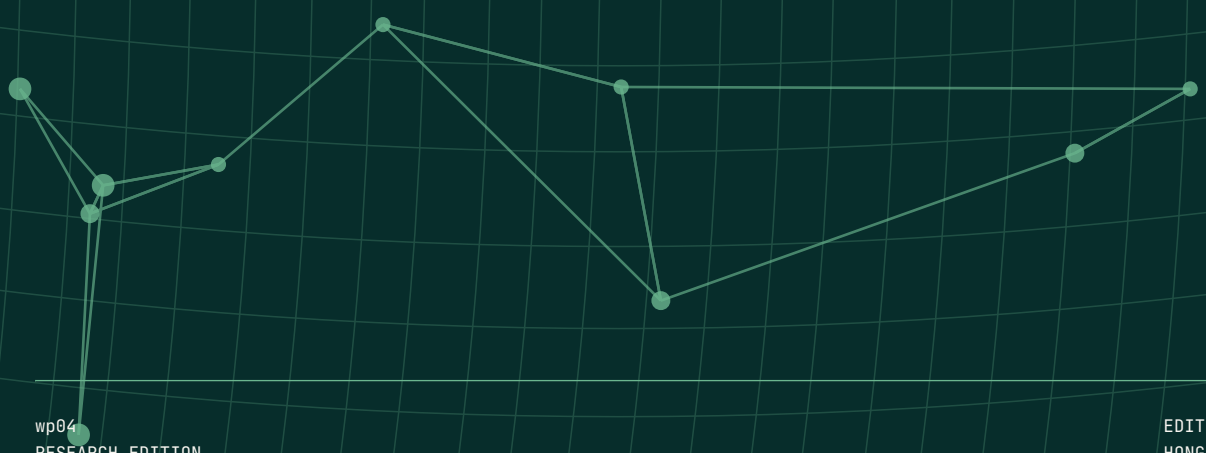
GREEN FUTURE TECH
LIMITED / HONG KONG SAR

GFT RESEARCH / WARP AND SIGNAL

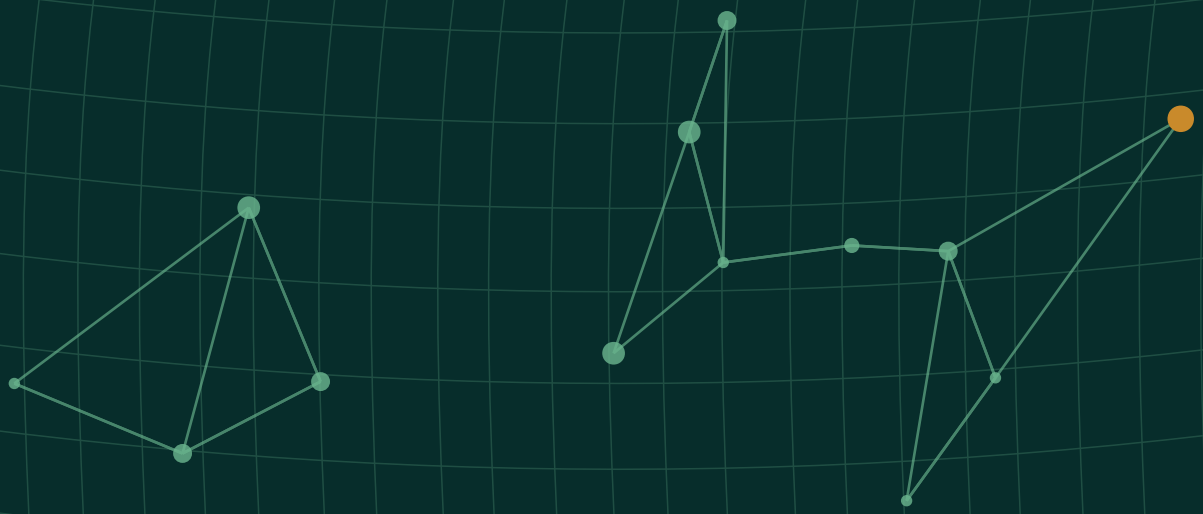
Forecasting From a Cold Start

Machine-learning demand forecasting for apparel makers with no sales history



wp04
RESEARCH EDITION

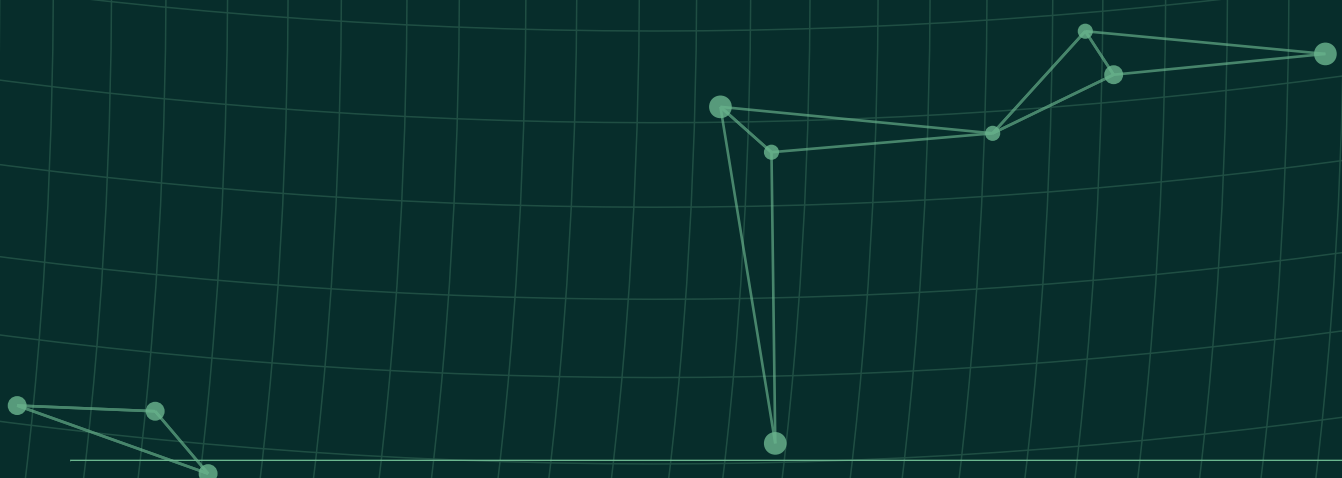
EDITION 2026-10-30
HONG KONG SAR



THE GFT THESIS

Value is made where it is measured.

GFT engineers sustainable factories that capture their own data from the first shift, so raw-material economies can keep the value they grow.



Read the evidence before acting on the recommendation.

ISSUER	Green Future Tech Limited, Hong Kong SAR
DOCUMENT	wp04 / Forecasting From a Cold Start
EDITION DATE	2026-10-30
AUDIENCE	Founders and operations leaders of greenfield apparel manufacturers, Demand planners, merchandisers and sales and operations planning teams, Industrial digital teams working with limited data and small technical staffs, Development finance and enterprise advisers assessing apparel launch plans, Researchers studying new-product and new-firm demand forecasting in African markets
STATUS	Research edition. Generated locally. Publication and operating-performance acceptance are separate decisions.
PROGRAMME	GFT's Cameroon textile programme: feasibility, performance-management system design and digital strategy. Prepared 31 January 2024; source-refresh annotations retained in 2026.

IMPORTANT NOTICE

This document is not investment, legal or tax advice. Proposed controls and illustrative calculations require project-specific validation. Targets and announced third-party projects are not evidence of delivered operating performance.

The confidential programme PDFs are not reproduced or published. No client identity or engagement code is included.

EVIDENCE REMAINS GRADED

A: official. B: institutional or major research. C: trade or local press and market-research vendors. D: unverified; excluded from use. A grade describes the supplied register and does not certify accuracy.

Inherited publication dates unrecorded: R01, R02, R03, R04, R05, R06, R07, R08, R09, R10, R11, R12, R13, R14, R15, R16, R18, R23, R24, R25, R26, R27, R29, R32. Temporal eligibility is unverified; these are inherited source-text checks, not independent source validation.

The operating case runs from fibre to feedback.

01 / No history is a forecasting condition, not a forecasting ban 15

1.1 / Greenfield apparel begins with transfer uncertainty rather than an empty evidence set 16

1.3 / GFT positions forecasting as a phased decision capability 18

1.5 / A new firm's cold start differs from a new SKU's cold start 19

1.7 / The forecast unit follows the commitment it must inform 21

1.9 / A useful first answer is a range with provenance 23

1.11 / Machine learning enters only after beating explicit baselines 25

1.13 / The argument moves from borrowed evidence to owned history 28

02 / The forecast target must be demand, not a convenient sales extract 31

2.1 / Unavailable stock can turn strong demand into weak sales 32

2.4 / Orders, shipments, sell-through and consumption answer different questions 34

2.7 / Returns and cancellations require event-based target definitions 35

2.9 / Product grain must separate style, colour, size and channel 38

2.11 / The horizon must match buying, production and replenishment clocks 39

2.13 / Sparse zeros need a meaning before they need a model 41

2.15 / A signed forecast contract makes accountability operational 44

03 / Analogue products provide the first outside view 47

3.1 / A reference class starts with comparable launches, not famous winners 48

3.2 / Attribute taxonomies make analogy auditable 50

3.3 / Visual similarity can transfer signals with explicit caveats 52

3.4 / Analogue markets require economic and channel comparability 54

3.5 / Partial pooling shares strength without forcing identical behaviour 56

3.6 / An analogue prior retains dispersion and disappointing cases 58

3.7 / Transfer weight should decay as target evidence accumulates 59

04 / External data can bound a market but cannot nominate a winner 62

4.1 / Reported trade flows describe supply, not consumer sell-through 63

4.4 / Mirror asymmetries need a reconciliation rule 66

4.7 / Values and quantities answer different diagnostic questions 68

4.9 / Population, income and urbanisation scale priors slowly 69

4.12 / Cotton statistics describe supply conditions rather than garment demand 72

4.14 / Search and weather must earn incremental value 73

4.17 / A covariate registry prevents a proxy from becoming a story 75

05 / Primary research turns intent into an explicit prior 80

5.1 / Respondent segments follow use occasions and buying roles 81

5.4 / Conjoint separates attribute preferences before launch 83

5.7 / Choice responses need calibration against purchase settings 85

5.9 / Price research produces scenarios, not a universal elasticity 89

5.12 / Delphi makes expert judgement anonymous and revisable 90

5.15 / Evidence synthesis must avoid counting one belief twice 92

5.18 / Pre-registration protects the prior from post hoc editing 94

06 / Small pilots convert hypotheses into purchase evidence 97

6.1 / A pilot tests a decision rather than decorating the launch plan 98

6.2 / Randomisation separates offer effects from self-selected demand 99

6.3 / Limited releases identify the ramp while containing exposure 102

6.4 / Availability must be designed into the pilot 103

6.5 / Channel and geography readouts remain separate 105

6.6 / Sequential updates preserve the prior and every revision 106

6.7 / Stop, adapt and scale rules are fixed before results arrive 108

07 / Clean launch records protect the history the model will learn 110

7.1 / Every sales event needs a stable identity and timestamp 111

7.2 / Availability and stockout timestamps are demand data 114

7.3 / Product attributes and images must be versioned with the SKU 116

7.4 / Price and promotion fields require an effective customer offer 117

7.5 / Returns, cancellations and substitutions stay visible 118

7.6 / External features need as-of timestamps to prevent leakage 119

7.7 / A small team needs validation, ownership and recovery before scale 121

08 / Simple models remain the cold-start benchmark 124

8.1 / Naive and seasonal-naive forecasts expose false sophistication 125

8.4 / Exponential smoothing becomes a durable baseline as history emerges 127

8.7 / Intermittent-demand methods address zero-heavy series with care 129

8.9 / Diffusion curves describe adoption only under strong assumptions 130

8.11 / Attribute regression links new items to analogue demand 131

8.14 / Equal-weight combinations are a serious candidate 132

8.16 / The model ladder promotes complexity only on evidence 134

09 / Global models borrow strength across products and attributes 139

9.1 / A global model learns one function across many series 140

9.5 / Static product attributes give unseen items a model entry point 142

9.9 / DeepAR offers a probabilistic cross-series candidate 144

9.12 / N-BEATS tests a generic deep architecture without covariate assumptions 145

9.15 / Temporal fusion adds input flexibility and operating burden 146

9.18 / Bayesian pooling balances shared and local parameters 149

9.21 / Similarity and global learning fail when the analogue mix shifts 151

10 / Pretrained forecasters must earn a place beside the baselines 154

10.1 / Zero-shot means an unseen dataset, not no evidence anywhere 155

10.4 / Chronos and Lag-Llama provide probabilistic candidates 157

10.7 / TimesFM tests decoder-only transfer across forecasting settings 158

10.10 / Univariate pretrained models do not use the full launch context by default 159

10.13 / Public benchmark rank is weak evidence under domain shift 161

10.16 / Compute, latency and vendor dependence belong in selection 162

10.18 / A challenger must beat the prior and the global baseline in launch replay 164

11 / Distributions and coherent hierarchies make forecasts usable 168

11.1 / Quantiles represent asymmetric demand risk 169

11.4 / Calibration must be checked rather than inferred from model class 171

11.7 / Proper scores compare distributions without collapsing them to one error 173

11.10 / Reconciliation keeps product, size, channel and market forecasts coherent 176

11.14 / Sparse count demand resists false continuous precision 179

11.17 / Scenario overlays remain distinct from statistical intervals 180

11.19 / Every distribution carries provenance, version and expiry 181

12 / Cold-start backtests must replay the information available at launch 185

12.1 / Rolling origins preserve chronology 186

12.2 / Held-out products test the unseen-item claim 188

12.3 / Held-out launches test the entire ramp 189

12.4 / As-of audits block feature and preprocessing leakage 190

12.5 / Scaled errors make heterogeneous series comparable 192

12.6 / Distribution scores sit beside point errors 193

12.7 / Decision-weighted readouts preserve operational meaning 194

12.8 / Release requires a baseline win, calibration and operating fitness 196

13 / Planners need a forecast service, not a model artefact 199

13.1 / One governed forecast version reaches the planning table 200

13.3 / The planner sees range, assumptions and freshness 202

13.5 / Production scenarios translate quantiles without automating commitment 203

13.6 / Overrides need reasons and subsequent scoring 205

13.8 / Monitoring watches inputs, errors, calibration and availability 206

13.10 / Rollback to a baseline is a normal control 208

13.12 / Value is measured in decisions rather than model loss alone 209

13.13 / Build-against-buy review protects data ownership and exit 211

14 / The eighteen-month roadmap replaces priors with owned evidence 213

14.1 / Recommendation one funds the evidence plan before the model platform 214

14.2 / Phase 0 defines the contract, analogues and pre-launch prior 215

14.3 / Phase 1 uses the first twelve weeks to validate evidence and capture 216

14.4 / Phase 2 uses months four to eight for controlled model comparison 218

14.5 / Phase 3 uses months nine to eighteen to scale only proven capability 220

14.6 / GFT would measure skill, calibration, contamination, adoption and planning outcome 221

14.7 / Decision gates keep uncertainty explicit for sponsors and planners 224

14.8 / The research agenda builds Cameroon apparel evidence rather than importing effect sizes 225

Executive summary 13

Evidence register 228

Glossary 251

Term index 254

Method and imprint 256

Every exhibit states a testable claim.

EXHIBIT 01 / flow
Evidence should change form as the launch matures 17

EXHIBIT 02 / matrix
New-item and new-firm cold starts require different evidence 20

EXHIBIT 03 / table
The forecast grain and horizon must match the commitment at risk 22

EXHIBIT 04 / keyfigures
Four questions discipline every cold-start forecast 27

EXHIBIT 05 / layered-architecture
Demand becomes learnable through four distinct event layers 33

EXHIBIT 06 / waterfall
An unadjusted invoice extract understates true launch demand 37

EXHIBIT 07 / matrix
Each planning clock needs its own forecast grain 40

EXHIBIT 08 / table
A target dictionary prevents silent metric substitution 42

EXHIBIT 09 / flow
Every zero record must pass an availability and listing audit before modelling 43

EXHIBIT 10 / table
A launch register keeps weak and failed cases in the evidence set 49

EXHIBIT 11 / matrix
An analogue scorecard exposes weak comparisons before modelling 52

EXHIBIT 12 / flow
Attributes become a prior through an inspectable transfer chain 54

EXHIBIT 13 / table
Each comparability layer has a limited transfer role 56

EXHIBIT 14 / range
Partial pooling preserves both shared and local uncertainty 57

EXHIBIT 15 / small-multiples
Disappointing launches remain part of the prior 59

EXHIBIT 16 / hbar
Reported apparel inflows need product and provenance detail 65

EXHIBIT 17 / waterfall
Mirror reconciliation bridges partner exports and reporter imports 67

EXHIBIT 18 / small-multiples
Slow market indicators belong above the SKU layer 71

EXHIBIT 19 / bar
Aggregate cotton output dwarfs individual mill intake 73

EXHIBIT 20 / layered-architecture
Each external signal stops at a defined evidence layer 75

EXHIBIT 21 / table
A covariate registry binds every external proxy to an as-of rule 77

EXHIBIT 22 / flow
Stated preference reaches a forecast through calibration 82

EXHIBIT 23 / range
Attribute utilities should remain distributions with sample provenance 84

EXHIBIT 24 / dumbbell
Calibration compresses raw stated choice into conservative purchase bounds 86

EXHIBIT 25 / table
A prior register makes dependence and uncertainty visible 93

EXHIBIT 26 / flow
Comparable test cells keep the launch readout interpretable 100

EXHIBIT 27 / table
A common denominator keeps the pilot comparison honest 101

EXHIBIT 28 / gantt
A twelve-week pilot can sequence learning without rushing scale 103

EXHIBIT 29 / range
Forecast revisions should show both learning and residual doubt 107

EXHIBIT 30 / layered-architecture
A learnable sales record connects six event and context layers 113

EXHIBIT 31 / flow
Stockout correction starts before model training 115

EXHIBIT 32 / table
A censored sales record supports scenarios rather than a unique recovery 116

EXHIBIT 33 / timeline
A forecast replay must use the feature value known at its origin 120

EXHIBIT 34 / heatmap
Completeness checks belong by field and sales week 122

EXHIBIT 35 / line
A baseline fan should widen where launch evidence is weak 127

EXHIBIT 36 / table
The intermittent baseline must expose its correction and inputs 130

EXHIBIT 37 / matrix
Evidence conditions determine which baseline is eligible 132

EXHIBIT 38 / bar
Relative error should be reported against each eligible baseline 134

EXHIBIT 39 / flow
The model ladder releases complexity only through a baseline gate 137

EXHIBIT 40 / layered-architecture
A global launch model combines shared learning with local evidence 141

EXHIBIT 41 / table
A feature becomes valid only when its temporal role is explicit 143

EXHIBIT 42 / matrix
Architecture choice follows evidence and input capacity 148

EXHIBIT 43 / range
Partial pooling should reveal both shrinkage and residual uncertainty 150

EXHIBIT 44 / flow
Shared learning remains conditional from corpus admission to rollback 152

EXHIBIT 45 / matrix
Zero-shot claims need precise evidence labels 156

EXHIBIT 46 / table
A pretrained candidate is approved with its operating evidence 163

EXHIBIT 47 / dumbbell
A pretrained challenger must clear two different benchmarks 165

EXHIBIT 48 / flow
Fallback belongs inside the pretrained-model design 166

EXHIBIT 49 / area
A launch fan exposes central and tail demand pathways across the pilot window 171

EXHIBIT 50 / small-multiples
Calibration can fail by product group or channel even when aggregate coverage looks sound 173

EXHIBIT 51 / table
Different distributional and coherence checks govern different launch failure modes 175

EXHIBIT 52 / sankey
Coherent reconciliation preserves volume conservation from style total to channels and sizes 177

EXHIBIT 53 / layered-architecture
A governed forecast distribution wraps raw quantiles in calibration, reconciliation and provenance layers 182

EXHIBIT 54 / timeline
Launch replay separates every training and decision cut-off 187

EXHIBIT 55 / flow
Every learnt transformation belongs inside the replay fold 191

EXHIBIT 56 / table
Paired launch errors expose skill and overforecasting together 192

EXHIBIT 57 / waterfall
Model value should survive each operational deduction 195

EXHIBIT 58 / heatmap
No single metric can carry the release decision 197

EXHIBIT 59 / layered-architecture
A forecast service joins evidence, models and planning controls 201

EXHIBIT 60 / table
A readable release carries decision context beside the forecast 203

EXHIBIT 61 / range
Production cases retain demand uncertainty across the planning window 204

EXHIBIT 62 / flow
The planning loop closes with measured forecast use 206

EXHIBIT 63 / keyfigures
The service scorecard balances skill, use and operating health 210

EXHIBIT 64 / flow
Every release requires evidence and a reversal path 219

EXHIBIT 65 / gantt
Preparation precedes the eighteen-month learning window 220

EXHIBIT 66 / line
Comparable exposure changes weight under an explicit rule 222

EXHIBIT 67 / table
The final measurement dictionary turns recommendations into management 223

EXHIBIT 68 / table
The final measurement dictionary turns recommendations into management, continued 224

Build the factory and its evidence as one system.

An apparel maker with no sales history does not face an empty evidence problem. It faces a transfer problem: which analogues, market statistics, expert judgements and early buying signals deserve to shape a launch forecast, and how quickly should their influence decay as real demand arrives? This paper argues that cold-start forecasting should be engineered as an evidence sequence, not bought as a single model. It begins with a forecast contract that separates latent demand from recorded sales, fixes product and channel grains, and makes stock availability a first-class field. It then builds explicit priors from comparable products, reported trade, population and income indicators, structured expert judgement and calibrated choice research. Small, availability-aware pilots convert those priors into owned evidence. Simple statistical and attribute models remain the benchmarks. Global neural models, Bayesian partial pooling and pretrained time-series models enter only when launch-replay backtests show gains in accuracy, calibration and planning utility. Probabilistic outputs, hierarchical reconciliation and rollback controls connect the model to production decisions without automating commitments. The closing roadmap shows how GFT would move a Cameroon apparel venture from pre-launch priors to a governed forecasting service over eighteen months. The result is not a promise of precise foresight. It is a measurable way to narrow uncertainty while protecting a new maker from false sophistication, stockout-contaminated learning and premature scale.

01 No sales history is not no evidence, but no model can create demand evidence from nothing. [S01] [S08] [S31]

02 A new firm's cold start is harder than a new SKU's cold start because attribute models still need other products on which to train. [S02] [S05]

03 Recorded sales must be corrected for availability before they can stand in for demand. [S06] [S07]

04 Analogue launch curves and partially pooled priors are more defensible starting points than an unverified complex model. [S28] [S30]

05 Naive forecasts, intermittent-demand methods and equal-weight combinations remain mandatory challengers throughout the model ladder. [S09] [S10] [S18] [S20]

06 Global and pretrained models are candidates for borrowing evidence, not proof that a Cameroon apparel launch will transfer from public benchmarks. [S21] [S22] [S25] [S26]

07 Usable launch forecasts carry calibrated quantiles and reconcile product, size, channel and market totals. [S14] [S15] [S16] [S17]

08 Cold-start validation must replay whole launches with as-of features, not randomise rows across time and products. [S02] [S05] [S13] [S37]

09 Calibrated choice research and controlled pilots provide the missing bridge between stated interest and owned purchase history. [S31] [S32]

10 Forecast value depends on governed planning use, drift response and measured decisions, not model loss alone. [S33] [S38] [S40]

GREEN FUTURE TECH

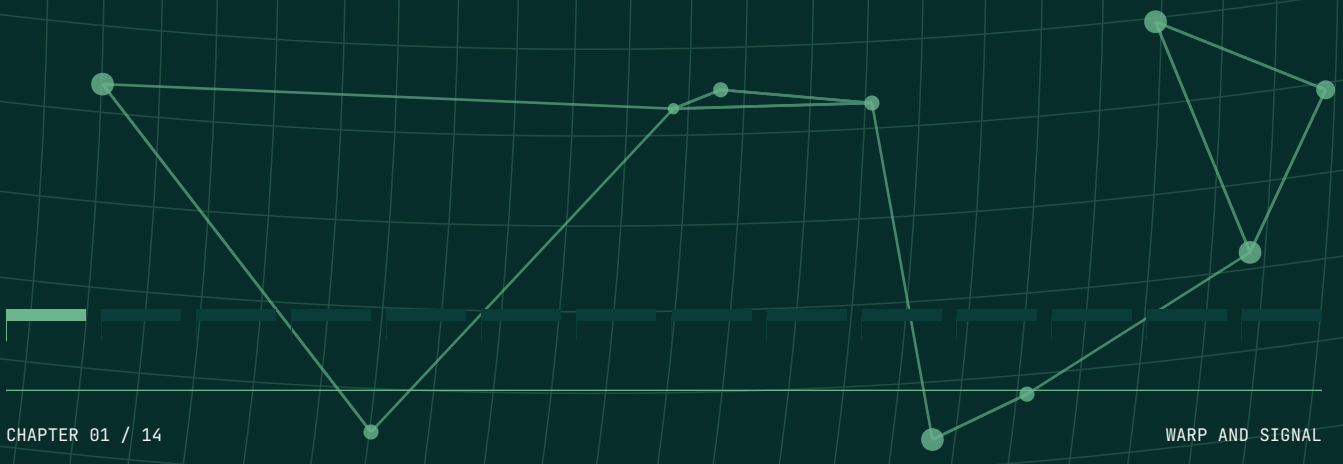
LIMITED / HONG KONG SAR

01

PART I: DEFINE THE DECISION AND THE PRIOR / THE COLD-START POSITION

No history is a forecasting condition, not a forecasting ban

A greenfield maker can forecast responsibly when it states what evidence transfers, what remains uncertain and what would change the plan.



CHAPTER 01
DECISION TEST

A greenfield maker can forecast responsibly when it states what evidence transfers, what remains uncertain and what would change the plan.

An apparel manufacturer opening its first production lines without a ledger of past orders does not face an obligation to guess blindly, nor does it possess a licence to defer demand planning until after launch. It faces a structured evidence-transfer problem: which external observations, product analogues, market bounds, buyer judgements and pilot purchase signals deserve weight before the first commercial run, and how rapidly should those borrowed priors yield to owned sales data once garments reach customers? Responsible cold-start forecasting treats the absence of target sales history as an explicit operating condition that governs model choice, uncertainty width and lot sizing. [S01] [S08] Rather than buying an opaque algorithm and hoping it can manufacture foresight from an empty table, a greenfield venture should engineer an auditable sequence of priors, availability-corrected launch pilots, simple statistical baselines and gated machine-learning challengers. [S08] [S32] [S33]

This position rejects two symmetrical errors that routinely damage new industrial ventures. The first is numerical fatalism: assuming that because a factory has zero historical invoices of its own, every pre-launch production commitment is pure intuition. The second is premature algorithmic theatre: deploying complex deep-learning or foundation-model pipelines on sparse, uncalibrated proxies and presenting a narrow point output as empirical fact. [S09] [S10] [S39] Between those failures lies an engineering middle path aligned with the Warp and Signal thesis of Green Future Tech Limited (GFT), Hong Kong SAR: production is the warp, measurement is the weft, and a factory that weaves measurement into its commercial and shop-floor processes from the first shift turns wide pre-launch priors into calibrated operational control. This paper provides a technical and operational framework for demand planning; nothing in it constitutes investment, legal or tax advice, or a guarantee of returns.

1.1

Greenfield apparel begins with transfer uncertainty rather than an empty evidence set

Missing historical data is not an anomaly in clothing manufacture; it is a defining structural condition of the sector. Academic reviews of apparel supply chains document that fashion demand combines taste volatility, seasonal cycles and acute lack of historical observations for new styles. [S01] Because clothing items have short commercial lifecycles, supply-chain scheduling and retail allocation decisions must be taken before a garment accumulates a multi-year sales series. [S0101] Even established apparel brands continuously launch silhouettes, fabrics and colourways that did not exist in prior seasons. [S01] [S02] What distinguishes a greenfield manufacturer is not novelty, but the depth of the transfer gap it must cross when constructing its initial expectation.

Forecasting theory addresses the zero-observation condition directly rather than treating it as a mathematical void. Comprehensive reviews of the discipline note that when no historical observations exist for the item of interest, structured judgemental methods, reference-class analogies and scenario design become the primary instruments of prediction. [S08] [S0102] When a planner estimates the completion time of a new facility or the initial uptake of an unlaunched product, the absence of an internal time series shifts the engineering task from parameter fitting to evidence curation.

[S08] The forecaster must still specify the target variable, bound the plausible range, document why chosen comparators are relevant and separate objective prediction from commercial target-setting. [S0102]

1.2 EVIDENCE DESIGN PRECEDES ALGORITHM SELECTION IN EVERY ZERO-HISTORY LAUNCH

Treating cold start as an evidence-design problem changes the first questions a technical team asks. Instead of asking which neural architecture handles short series best, the team asks where transferable signal resides and what distortions contaminate it. Prior information for a new garment venture arrives in four forms: historical launch trajectories of analogous products from external or partner catalogues, macro-level trade and demographic bounds, structured elicitation from domain buyers, and choice experiments conducted with prospective customers. [S02] [S29] [S31] None of these sources equals an owned, stockout-corrected transaction log, yet each carries partial information that narrows an otherwise arbitrary planning range.

EXHIBIT 01 / LAUNCH EVIDENCE STAGES

Evidence should change form as the launch matures



1. Screen attributes
2. Size small lots
3. Correct stockouts
4. Replay and gate

ILLUSTRATIVE [S28] [S31] [S32] Illustrative sequence of evidence stages for a cold-start apparel launch, drawing on hierarchical pooling, preference calibration and controlled pilot design. Node order shows information maturity, not a fixed calendar duration.

Exhibit 1, p. 17 illustrates the progression that organises this paper. Pre-launch, the venture relies on analogue distributions and expert priors to form an outside view. [S08] [S28] Before committing full factory capacity, it refines those priors through calibrated choice research and small, availability-aware pilot releases that generate the first owned purchase observations. [S31] [S32] Only after clean transaction and stockout records accumulate does the venture shift weight toward statistical smoothing, attribute regressions and cross-series models. [S02] [S06] [S08] Crucially, **Exhibit 1**, p. 17 is an operational maturity sequence rather than a guarantee that uncertainty

shrinks monotonically; if a pilot reveals that an analogue pool was poorly matched to local buyers, the right response is to widen the interval and revise the assortment before scaling.

Separating forecast construction from target setting is essential during this transition. Guidance on judgemental forecasting warns that when staff responsible for commercial quotas also author the demand forecast without independent rules, optimism bias and political anchoring corrupt the numbers. [S0102] A business plan states the volume required to amortise machinery, whereas a demand forecast estimates what customers will buy under stated price, channel and availability conditions. [S08] [S0102] Maintaining a hard wall between break-even targets and empirical demand priors is the first governance rule of a greenfield launch.

1.3

GFT positions forecasting as a phased decision capability

Within GFT's three-part Cameroon textile programme, prepared on 31 January 2024 with source refreshes flagged through 2026, demand forecasting is architected as a Phase 2 digital capability rather than a day-one software purchase. [R06] [R07] Across the programme's feasibility study, performance-management system design and digital strategy, Phase 1 focuses on foundational instrumentation: reliable enterprise resource planning records, standardised stock-keeping unit (SKU) identifiers, shop-floor quality logs and clean commercial transaction capture. A machine-learning model fitted before the factory can distinguish an unfulfilled order from a cancelled invoice, or a stockout from a zero-demand week, only automates the propagation of measurement noise. [S06] [S37]

Waiting for Phase 2 to deploy complex algorithms does not mean that pre-launch and Phase 1 decisions occur without quantitative support. Before a plant trains its first cross-series model, management must choose initial garment categories, book yarn and fabric capacities, negotiate institutional uniform tenders and size opening cut-and-sew batches. This white paper specifies the pre-Phase-2 and transition machinery that bridges that gap: writing the demand contract during project preparation, assembling an auditable prior from analogue and primary research, running small launch pilots that earn clean history, and governing graduation from simple baselines to global probabilistic models in Phase 2. [S02] [S21] [S28] [S32]

1.4 INDUSTRIAL POLICY TARGETS JUSTIFY LAUNCH DISCIPLINE BUT NEVER SUBSTITUTE FOR DEMAND EVIDENCE

The macroeconomic context in Cameroon explains why disciplined demand engineering carries high stakes for local textile industrialisation. According to the inherited sector register, Cameroon produced 330,000 tonnes of seed cotton in 2022, 314,455 tonnes in 2023 and about 340,000 tonnes in 2024, with 2025 projections exceeding 350,000 tonnes. [R01] [R02] Yet roughly 95 percent of Cameroonian cotton fibre is exported raw while only about 5 percent is processed locally, alongside a national SND30 target to reach 50 percent local processing by 2030. [R06] [R07] Large industrial initiatives reflect this policy push, including the announced SODECOTON

and CICAM merger targeted by 2030 and the Camtext SA integrated project announced on 12 November 2025 at US$300 million (180 billion XAF) to process 12,000 tonnes a year toward optimal capacity around 2033. [R08] [R09] [R10]

A greenfield apparel maker must treat national processing targets and incumbent metrics with strict analytical care. The gap between 5 percent local fibre processing and a 50 percent national target demonstrates raw-material supply potential, not guaranteed consumer or institutional absorption for any specific garment line. [R06] [R07] Historical distress at existing domestic facilities underscores the cost of mis-judging market fit and utilisation: public reports indicate that CICAM's domestic market share fell from about 85 percent three decades ago to about 5 percent, with July 2023 capacity utilisation reported at 33.7 percent at Garoua, 48.9 percent at CICAM 1 and 10.83 percent at CICAM 2, and underutilisation losses of about 2 billion XAF in 2020. [R11] [R12] Meanwhile, Statista estimates Cameroon clothing and footwear expenditure at about US$94.70 per person (around 8.0 pieces) in 2024 and apparel e-commerce revenue rising from US$20.32 million in 2023 to US$33.61 million by 2027, while McKinsey reports that AI at scale in established fashion firms has been associated with 20 to 50 percent forecast-error reductions and 35 to 40 percent inventory-cost reductions. [R26] [R27] None of those inherited benchmarks may be copied into a greenfield plan as an expected GFT factory result.

> **GFT NOTE**
>
> Never convert a national processing target, an aggregate import figure or an industry AI benchmark into a style-level sales forecast. In GFT's Cameroon textile programme, the 95 percent raw-export baseline and the 50 percent SND30 processing target establish why local value addition matters, whereas McKinsey's 20 to 50 percent error-reduction range reflects mature enterprises with existing data spines. [R06] [R07] [R27] A new apparel plant must earn its own demand parameters through a signed forecast contract, explicit analogue priors and availability-controlled launch pilots.

1.5

A new firm's cold start differs from a new SKU's cold start

Much of the academic and vendor literature on cold-start fashion forecasting solves a narrower problem than the one confronting a new manufacturer. In standard retail machine-learning studies, "cold start" denotes a new SKU introduced by an established multi-brand retailer that already possesses years of transaction logs across thousands of prior items, stores and customer segments. [S02] [S03] When such a retailer launches a new cotton poplin shirt, its data team can train a supervised regression, tree ensemble or neural network on historical sales of hundreds of previous shirts, using shared product attributes, store features and calendar seasonality to predict the unseen item's trajectory. [S02] [S03]

A new apparel firm entering production without an owned historical catalogue faces a harder regime. It lacks not only the target sales series for the new garment, but also the internal training table of historical products and outcomes from which an attribute-to-demand mapping is learned. [S02] If an engineer runs an off-the-shelf

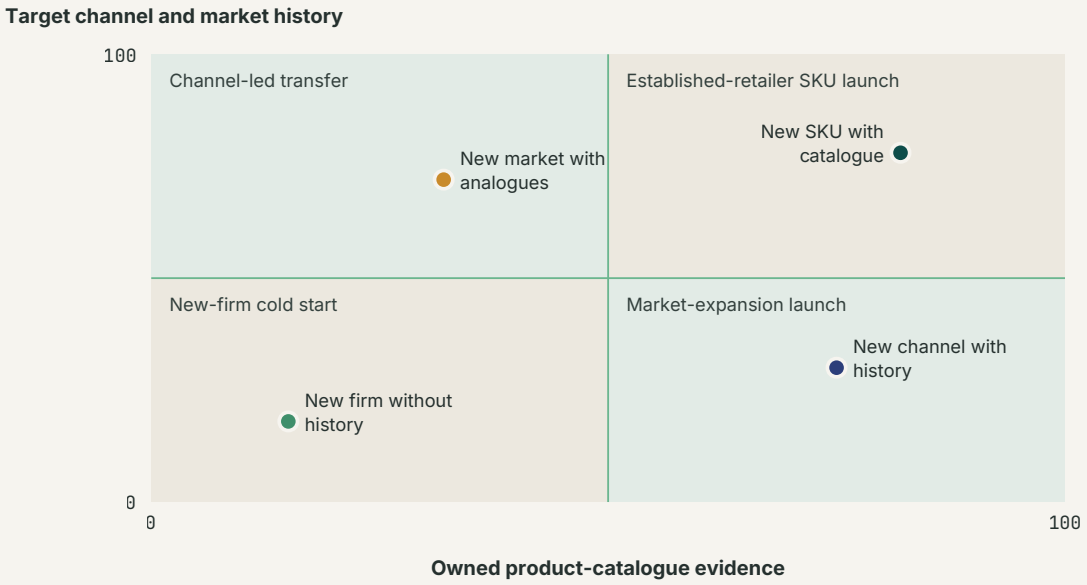
attribute-transfer model when the internal database holds zero completed product lifecycles, the model has no empirical support on which to estimate the effect of fabric weight, collar style, price band or sales channel.

1.6 TRANSFER MODELS REQUIRE A POPULATED REFERENCE CLUSTER BEFORE THEY CAN GENERALISE

Recent advances in multimodal deep learning make this distinction sharper. Modern cold-start architectures extract visual embeddings from garment photographs, group historical items into visual and behavioural clusters, and transfer stockout-corrected demand profiles from the nearest historical neighbours to the newly designed item. [S05] That transfer mechanism works only if the reference clusters already contain relevant, demand-corrected historical products sold under comparable price and channel conditions. [S05] A greenfield manufacturer in Cameroon launching school uniforms, corporate workwear or casual knitwear cannot populate those clusters from its own empty ledger; it must curate an external analogue corpus, elicit structured priors from experienced regional buyers, or generate its first clusters through staged pilot releases. [S05] [S28] [S32]

EXHIBIT 02 / CONCEPTUAL EVIDENCE COORDINATES FROM 0 TO 100

New-item and new-firm cold starts require different evidence



DERIVED [S01] [S02] [S05] Derived conceptual mapping of four cold-start regimes across owned catalogue depth and target market history, synthesised from fashion cold-start reviews and attribute-transfer studies. Coordinates are conceptual positions on an index from 0 to 100, not empirical accuracy measurements.

Exhibit 2, p. 20 separates four cold-start regimes along two conceptual axes: owned product-catalogue evidence on the horizontal axis and target channel and market history on the vertical axis, each expressed on an illustrative conceptual scale from 0 to 100 derived from the problem definitions in the fashion forecasting literature. [S01] [S02] [S05] In the upper-right quadrant sits the established-retailer SKU launch (conceptual coordinates 82 and 78), where rich attribute and store histories support supervised attribute models immediately. [S02] [S03] In the lower-right quadrant sits a known product range entering a new sales channel (coordinates 75 and 30), where product preferences are partially known but channel fulfilment and seasonality require

fresh calibration. [S04] In the upper-left quadrant sits an expansion into a familiar channel with a new product category supported by external analogues (coordinates 32 and 72). Finally, in the lower-left quadrant sits the new-firm cold start (coordinates 15 and 18), where neither an owned product catalogue nor an owned channel transaction log exists at launch. [S01] [S08]

What **Exhibit 2**, p. 20 establishes is that borrowing a method designed for the upper-right quadrant and deploying it unmodified in the lower-left quadrant is a category error. What the matrix cannot establish is the exact numerical accuracy penalty of moving between quadrants, because empirical error depends on the homogeneity of the target category; an institutional workwear contract with fixed headcount specifications carries less style uncertainty than a consumer fashion drop even when both start in the lower-left quadrant. [S01] [S08]

1.7

The forecast unit follows the commitment it must inform

A demand forecast has no operational meaning until its grain, horizon and target event are bound to a specific physical or financial commitment. In apparel manufacturing, candidate dimensions include style, fabric construction, colourway, size, sales channel, geography and week. [S02] [S03] [S04] Publishing a forecast at the finest combination of all dimensions months before launch creates an illusion of precision while guaranteeing extreme noise, whereas publishing only an annual factory tonnage figure leaves planners unable to order dye lots or cut size curves. [S08] [S11] The published forecast grain must match the finest level at which a named decision-maker can alter an operational commitment. [S08] [S0102]

Different clocks govern an apparel venture's commitments, and each clock demands a different forecast structure. Long-range capacity and raw-material booking (reserving greige cotton fabric weights or yarn counts 16 to 26 weeks ahead of delivery) depends on aggregate fabric-family demand across styles. [S01] [S0101] Medium-range assortment and initial cut-and-sew lot sizing (6 to 12 weeks before launch) requires style-by-colour volume ranges and a prior size-distribution profile so the cutting room can lay markers efficiently. [S02] [S03] Short-cycle channel allocation and replenishment (1 to 4 weeks ahead) operates at the SKU-by-channel-by-week grain, where observed sell-through and stockout flags update the initial prior. [S03] [S06]

EXHIBIT 03 / DECISION ALIGNMENT SCHEDULE

The forecast grain and horizon must match the commitment at risk

Commitment	Published grain	Primary horizon	Decision focus
Fabric and yarn booking	Fabric family by season	16 to 26 weeks	Upper quantiles for raw-material buffer
Initial cut-and-sew lot	Style, colour and size curve	6 to 12 weeks	Central range and downside overstock cap
Channel allocation	Style, colour and channel	2 to 6 weeks	Availability-corrected sell-through balance
Short-cycle replenishment	SKU by location and week	1 to 4 weeks	Upper tail service and stockout prevention

DERIVED [S03] [S06] [S08] [S16] Derived operational mapping of apparel launch commitments to published forecast grain, horizon and primary quantile focus, grounded in retail forecasting practice and multi-quantile evaluation literature.

Exhibit 3, p. 22 summarises how commitment clocks map to forecast grains, illustrative planning horizons in weeks, and the portion of the predictive distribution that drives the decision. [S03] [S06] [S08] [S16] When booking undyed or standard-dyed fabric 16 to 26 weeks ahead, pooling across styles that share the same base cloth reduces variance, because fabric can still be allocated across silhouettes later in the calendar. When releasing the initial cut-and-sew lot 6 to 12 weeks before launch, the planner commits to specific styles, colours and size ratios, but caps the initial batch using the downside and central quantiles so untested styles do not lock up working capital in finished goods. [S16] [S32] Once the launch is live, 1-to-4-week replenishment shifts focus to SKU availability and upper-tail quantiles to prevent stockouts on fast sellers. [S06] [S16] While **Exhibit 3**, p. 22 provides an operational template, each factory must substitute its actual spinning, dyeing, cutting and shipping lead times into the contract.

1.8 RECORDED SALES ARE A CENSORED VIEW OF LATENT DEMAND

Aligning the forecast unit with the decision also requires strict separation between recorded sales and latent demand. Retail forecasting research establishes that estimating demand solely from sales transactions introduces a systematic negative bias whenever products experience stockouts or partial availability. [S06] Stockout-annotated retail datasets confirm that the sales-equals-demand assumption collapses when inventory runs dry, because observed sales are right-censored at the available stock ceiling. [S07] In a cold-start apparel launch, where initial pilot lots are kept small to contain financial exposure, stockouts on popular sizes or colours are frequent. [S06] [S32] Feeding raw, uncorrected pilot sales into a replenishment model causes the model to mistake a stockout week for a demand plateau and starve the winning SKU. [S06] [S07]

Channel structure further complicates the target definition. An order placed by a wholesale distributor or an institutional buyer is a lumpy commitment at the factory gate, whereas consumer sell-through in direct retail or e-commerce reflects end-buyer uptake day by day. [S04] [S06] Treating wholesale sell-in and retail sell-through

as interchangeable rows in one time series corrupts both level and timing. Chapter 2 therefore formalises a binding forecast contract that separates order, dispatch, sell-through, return and stockout events before any model runs. [S06] [S07] [S36]

1.9

A useful first answer is a range with provenance

Because a greenfield forecast rests on transferred evidence and explicit assumptions, publishing a single deterministic number is scientifically indefensible and operationally dangerous. A point forecast of 4,000 shirts hides whether the analytical team believes demand will fall between 3,600 and 4,400 illustrative units or could span 800 to 9,500 illustrative units. Statistical theory on probabilistic prediction demonstrates that predictive distributions and quantile intervals are required to represent uncertainty honestly and align forecasts with asymmetric decision costs. [S14] When overproducing an untested style leaves dead stock below cost while underproducing a uniform contract triggers penalties, the optimal production quantity rarely equals the mean or median forecast. [S06] [S14]

Modern retail forecasting competitions have codified multi-quantile representation as the standard for operational uncertainty. In the M5 Uncertainty competition, forecasters predicted nine quantiles (0.005, 0.025, 0.165, 0.250, 0.500, 0.750, 0.835, 0.975 and 0.995) across hierarchical retail series so that central tendency and tail risks could be evaluated rigorously. [S16] Adopting a multi-quantile contract from the first cold-start prior allows a greenfield apparel maker to speak a consistent probabilistic language from pre-launch elicitation through to Phase 2 machine learning. [S08] [S14] [S16]

1.10 PROVENANCE METADATA MAKES A WIDE PRIOR INSPECTABLE AND REVISABLE

A probability interval is useful only if planners and sponsors can inspect where its bounds came from and how they were calibrated. Even deep-learning regression algorithms produce miscalibrated uncertainty intervals unless checked against held-out observations, and post-hoc calibration guarantees require sufficient sample support. [S15] Before target observations exist for empirical recalibration, the honesty of a predictive distribution depends on recording its provenance: which analogue launches were pooled, what weights were assigned to buyer surveys versus expert panels, and which channel assumptions condition the range. [S08] [S14] [S28]

METHOD

To publish an inspectable cold-start forecast distribution at a given forecast origin and horizon, record four mandatory elements alongside the numeric quantiles: 1. **Quantile vector**: at minimum the 10th, 25th, 50th, 75th and 90th percentiles, supplemented by outer tail quantiles (2.5th and 97.5th percentiles) when capacity or safety-stock commitments are evaluated. [S14] [S16] 2. **Evidence lineage and weights**: the explicit non-negative mixture weights summing to 1.0 assigned to each independent evidence source (analogue pool prior, calibrated choice survey and anonymous Delphi panel), together with sensitivity to alternative weightings. [S20] [S28] [S29] 3. **Hierarchical coherence rule**: how variant-level distributions (style, colour and size) reconcile with style, channel and factory totals so planners do not act on contradictory numbers across aggregation levels. [S11] [S17] 4. **Update and expiry trigger**: the observation count, pilot milestone or stockout threshold at which the prior updates into a posterior distribution or retires in favour of an empirical baseline. [S28] [S40]

Recording provenance in this form also supports hierarchical coherence. Size-level and colour-level forecasts must reconcile with style-level fabric bookings and channel budgets. [S11] [S17] Research on optimal forecast reconciliation demonstrates that base forecasts across a hierarchy can be reconciled via trace minimisation (MinT) to minimise mean squared error across the coherent collection under unbiasedness assumptions. [S17] Tracking both the quantile range and the aggregation key from day one keeps planning tables consistent as products graduate from priors to observed series. [S16] [S17]

WORKED EXAMPLE: SIZING A TWO-STAGE LAUNCH COMMITMENT FROM AN ILLUSTRATIVE COLD-START PRIOR

Consider an illustrative greenfield workwear maker planning a 12-week launch of a cotton twill utility shirt. With no historical sales, it constructs an illustrative pre-launch prior for 12-week style demand by combining an analogue pool and a calibrated buyer choice exercise. [S28] [S31] Suppose the illustrative predictive quantiles for the 12-week horizon are 1,200 units at the 10th percentile, 1,800 units at the 25th percentile, 2,600 units at the 50th percentile, 3,500 units at the 75th percentile and 4,600 units at the 90th percentile.

Assume an illustrative unit variable cost of 6,000 XAF, an illustrative wholesale price of 10,000 XAF (an illustrative contribution margin of 4,000 XAF per unit sold), and an illustrative salvage value of 3,500 XAF for unsold shirts at week 12 (an illustrative overstock loss of 6,000 minus 3,500 = 2,500 XAF per excess unit). Under a single blind batch commitment before launch, the classical critical fractile is 4,000 ÷ (4,000 + 2,500) = 0.6154 (about the 62nd percentile), sitting between the 50th percentile (2,600 units) and 75th percentile (3,500 units) at roughly 3,015 illustrative units. Committing 3,015 units upfront exposes the venture to 3,015 minus 1,200 = 1,815 unsold shirts (an illustrative overstock loss of 1,815 × 2,500 = 4,537,500 XAF) if demand lands at the 10th percentile of 1,200 units.

Instead, GFT's staged evidence protocol splits the commitment across a 3-week availability-monitored pilot. [S06] [S32] Before launch, the factory books greige fabric to the 75th percentile equivalent (3,500 shirt equivalents, as undyed twill can be redirected to standard trousers without salvage loss), but releases an initial cut-and-sew lot at the 25th percentile of 1,800 shirts for weeks 1 to 4. During weeks 1 to 3, the team records transactions and stockout timestamps across sizes. [S06] [S07] If availability-corrected demand tracks the upper quartile, the cutting room releases a second batch of 1,700 shirts in week 4 to reach 3,500 total units; if demand tracks the 10th percentile of 1,200 units, sewing stops at 1,800 units, capping unsold finished goods at 1,800 minus 1,200 = 600 shirts (an illustrative overstock exposure of 600 × 2,500 = 1,500,000 XAF, cutting downside loss by 3,037,500 XAF). All figures in this box are illustrative calculations; they are not observed GFT factory results.

1.11

Machine learning enters only after beating explicit baselines

Model complexity must be earned through empirical cold-start backtesting rather than assumed from architectural novelty. Large-scale empirical evaluations have repeatedly shown that complex machine-learning models do not automatically outperform simple statistical baselines or transparent forecast combinations. [S09] [S10] In a systematic comparison of statistical and machine-learning methods on monthly series, researchers found that a seasonal random walk benchmark (Naive 2) achieved greater accuracy than more than half of the evaluated machine-learning methods. [S09] Similarly, the M4 competition across 100,000 series reported that all six pure machine-learning submissions performed poorly relative to statistical combinations, with none beating the combination benchmark and only one outperforming Naive 2. [S10]

Those benchmark lessons apply directly to a greenfield apparel maker with small technical staffs and short histories. When target series are short, noisy or intermittent, high-capacity neural networks easily overfit launch spikes, promotional artefacts or unrecorded stockouts. [S06] [S08] [S18] Meanwhile, an equal-weight combination of simple, distinct baselines provides a strong, low-variance challenger; a fifty-year review of forecast combinations confirms that equally weighted averages frequently match or beat elaborate weighting schemes because they avoid weight-estimation error. [S20]

1.12 STATISTICAL LOSS MUST BE PAIRED WITH OPERATIONAL AND SYSTEM-FITNESS GATES

Even when a candidate machine-learning model (such as DeepAR, an attribute tree ensemble, or a pretrained foundation model like Chronos or TimesFM) improves upon naive, intermittent and equal-weight baselines in a statistical error metric, it has not yet earned deployment. [S21] [S25] [S26] Machine-learning engineering rules emphasise that a change improving mathematical loss while degrading system behaviour or decision utility should be rejected. [S33] Seminal work on technical debt in machine-learning systems warns that solutions delivering a tiny accuracy gain at the cost of massive increases in system complexity are rarely wise practice. [S39]

EXHIBIT 04 / MANDATORY PRE-LAUNCH CONTRACT GATES

Four questions discipline every cold-start forecast

Commitment

Gate 1: Decision

Which lot, purchase or allocation changes if the range shifts?

Latent demand

Gate 2: Target

How are stockouts, returns and channel timing separated?

Prior weights

Gate 3: Provenance

Which analogues, surveys and judgements form the range?

Stop rule

Gate 4: Reversal

What pilot or replay outcome forces a fallback to baseline?

ILLUSTRATIVE [S08] [S14] [S33] [S36] Illustrative governance tiles summarising the four questions required before a cold-start forecast may inform a production commitment. Tiles state governance gates rather than empirical performance metrics.

Exhibit 4, p. 27 sets out the four governance gates that every cold-start forecast must clear before reaching a planning table. [S08] [S14] [S33] [S36] Gate 1 requires a named commitment so the team never runs models for decorative reporting. [S08] Gate 2 requires a clean target definition accounting for stockouts, returns and channel timing so the model does not learn supply limits as demand ceilings. [S06] [S07] Gate 3 requires explicit prior weights and provenance so reviewers can inspect how analogues, statistics and expert judgements were combined. [S14] [S28] Gate 4 requires a pre-agreed reversal rule so that if a complex model drifts, leaks information or fails calibration during launch replay, the planning service reverts automatically to an approved baseline. [S33] [S36] [S40] Clearing all four gates in **Exhibit 4**, p. 27 guarantees procedural auditability rather than commercial profitability; market risk must still be managed through small pilot lots.

Evaluating challengers against these gates requires strict chronological hygiene. Random cross-validation across rows leaks future demand information into training folds and overstates how well an algorithm predicts an unseen launch. [S02] [S13] [S37] Valid cold-start evaluation holds out entire product launches and steps forward across expanding chronological origins using only as-of feature snapshots available

at each decision cut-off. [S02] [S13] [S37] Where mean absolute scaled error (MASE) compares series across scales, its naive denominator must use only observations prior to the forecast origin, and be reported as undefined when no target history exists yet. [S12]

> **WARNING**
>
> Unchecked data leakage is the most common reason a cold-start machine-learning prototype appears impressive in a notebook and fails in the factory. [S37] Computing category-mean encodings across the full season, normalising prices using post-launch averages, selecting analogue clusters using realised sales, or backfilling macro and search covariates without as-of publication timestamps all inject future truth into pre-launch features. [S05] [S37] Any model whose validation pipeline violates strict launch-origin cut-offs must be disqualified regardless of its reported error score.

1.13

The argument moves from borrowed evidence to owned history

The fourteen chapters of this paper are organised as four sequential parts that take a greenfield apparel manufacturer from its initial pre-launch prior to a governed, eighteen-month forecasting service. Every chapter leaves a small industrial team with an inspectable engineering artefact, an empirical validation test or a human-governed release gate, ensuring that the methodology can be executed without a large data-science department. [S36] [S39] [S40]

Part I (Chapters 1 to 4) defines the decision boundary and builds the outside-view prior. Following this opening chapter, Chapter 2 writes the demand contract, separating gross orders, shipments, sell-through, returns and stockout-censored intervals across planning horizons. [S06] [S07] [S18] Chapter 3 constructs the analogue prior engine, scoring comparable launches, including slow sellers and withdrawn lines, across attributes, prices, channels and visual clusters, and using hierarchical Bayesian partial pooling to share strength without erasing dispersion. [S01] [S02] [S05] [S28] Chapter 4 bounds the market using external covariates, including UN Comtrade inflows, mirror asymmetries, World Bank indicators, FAOSTAT fibre data, search queries and weather. [R06]

Part II (Chapters 5 to 7) earns target-specific evidence before scaling production. Chapter 5 turns primary research into a calibrated prior, combining conjoint designs with realisation-probability calibration and anonymous Delphi elicitation while preventing double counting. [S29] [S31] Chapter 6 designs small, availability-aware launch pilots over twelve weeks to convert hypotheses into purchase and repeat-buying observations while capping inventory exposure. [S06] [S32] Chapter 7 specifies the pilot data spine, capturing order timestamps, stockout intervals, versioned attributes, effective offers and as-of feature snapshots. [S02] [S06] [S37] [S40]

Part III (Chapters 8 to 11) builds the modelling ladder under scarcity and uncertainty. Chapter 8 establishes the floor of simple baselines: naive and seasonal-naive walks, exponential smoothing, Croston and Syntetos-Boylan intermittent estimators, Bass

diffusion scenarios, attribute regressions and equal-weight combinations. [S02] [S08] [S09] [S18] [S20] [S30] Chapter 9 evaluates global cross-series models (DeepAR, N-BEATS, Temporal Fusion Transformers and hierarchical Bayesian pooling) that borrow strength across a growing catalogue. [S21] [S22] [S28] Chapter 10 subjects pretrained foundation models (Chronos, TimesFM and Lag-Llama) to the same cold-start contract, separating public benchmarks from domain-shifted launches. [S25] [S26] Chapter 11 ensures that winning models produce calibrated multi-quantile distributions and hierarchically coherent totals across sizes, colours, styles and channels. [S14] [S15] [S16] [S17]

Part IV (Chapters 12 to 14) proves operational value and governs graduation. Chapter 12 specifies the whole-launch replay backtest harness, enforcing expanding windows, held-out launches, as-of leakage audits, undefined-denominator rules for MASE, proper scoring rules and decision-weighted gates. [S02] [S12] [S13] [S14] [S33] [S37] Chapter 13 embeds winning forecasts into a monthly sales and operations planning service with versioned releases, scenario translators, override tracking, drift monitors and rollback controls. [S29] [S36] [S38] [S40] Finally, Chapter 14 synthesises the eighteen-month roadmap for GFT's Cameroon textile programme from Phase 0 contracts and priors through Phase 1 pilots, Phase 2 model competition and Phase 3 governed scale without importing external vendor effect sizes as promised returns. [R06] [R07] [S32] [S39]

Across all four parts, the governing principle remains constant. A new manufacturer does not need to apologise for lacking a decade of sales invoices, nor gamble its working capital on black-box algorithms that hide their assumptions. By treating no history as a structured forecasting condition, weaving measurement into every pilot lot, and promoting models only when they beat transparent baselines in launch replay, a greenfield apparel maker can narrow uncertainty shift by shift and keep the value it engineers.

A cold-start forecast is not a substitute for history; it is the disciplined contract by which a factory earns its first history without betting the mill.

CHAPTER SOURCE NOTES

S01 InTech, International Journal of Engineering Business Management. 2013.

S08 International Journal of Forecasting, Elsevier; full text through arXiv. 2022.

S32 Springer Nature, Data Mining and Knowledge Discovery. 2008-07-30.

S33 Google for Developers. 2025-08-25.

S09 PLOS ONE, Public Library of Science. 2018-03-27.

S10 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2018.

S39 Neural Information Processing Systems Foundation, NIPS 2015. 2015.

S0101 arXiv. 2022-11-09.

S02 arXiv. 2019-06-27.

S0102 OTexts. 2021-05-31.

S29 RAND Corporation. 2023-12-29.

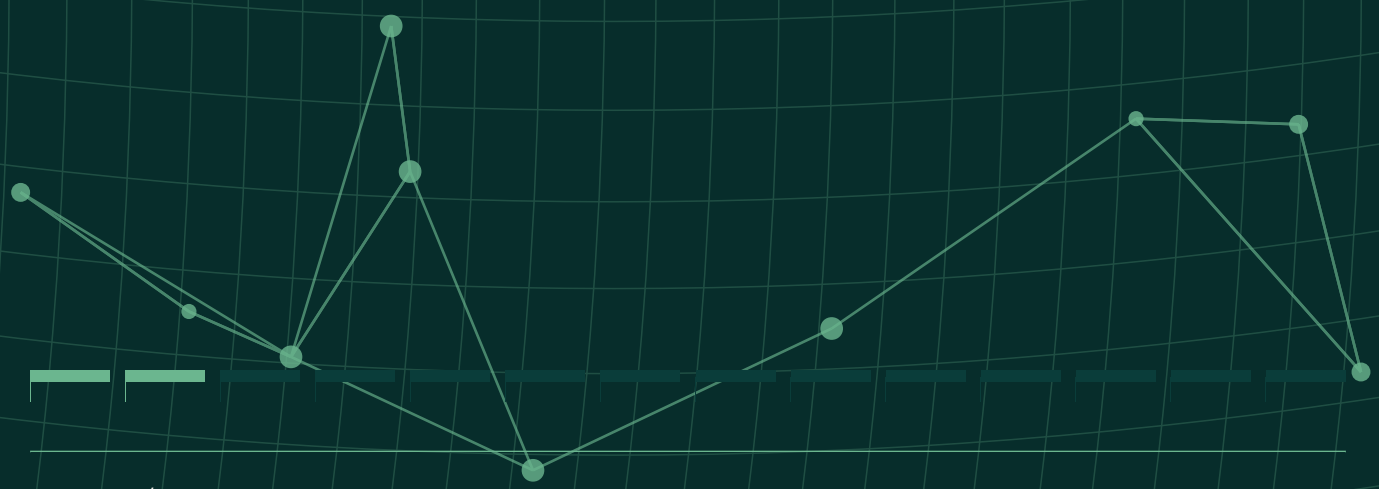
S31 INFORMS, Marketing Science; open manuscript from MIT. 2020.
S28 Journal of Industrial Ecology, Wiley. 2022-10-17.
S06 Lancaster University Management School, Munich Personal RePEc Archive. 2018-11-19.
R06 Cameroon Trade Hub. Publication date unrecorded; inherited source-text check.
R07 Ecofin Agency. Publication date unrecorded; inherited source-text check.
S37 Cell Press, Patterns; open full text from PubMed Central. 2023-08-04.
S21 arXiv; journal version in International Journal of Forecasting, Elsevier. 2019-02-22.
R01 Business in Cameroon. Publication date unrecorded; inherited source-text check.
R02 Fibre2Fashion; Business in Cameroon. Publication date unrecorded; inherited source-text check.
R08 Business in Cameroon. Publication date unrecorded; inherited source-text check.
R09 Ecofin Agency. Publication date unrecorded; inherited source-text check.
R10 Ecofin Agency. Publication date unrecorded; inherited source-text check.
R11 Business in Cameroon; MMI News. Publication date unrecorded; inherited source-text check.
R12 Business in Cameroon. Publication date unrecorded; inherited source-text check.
R26 Statista. Publication date unrecorded; inherited source-text check.
R27 McKinsey, The State of Fashion 2025. Publication date unrecorded; inherited source-text check.
S03 arXiv. 2020-07-10.
S05 Springer Nature, International Journal of Data Science and Analytics. 2026-01-08.
S04 MDPI, Processes. 2021-09-03.
S11 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2022.
S16 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2022.
S07 arXiv. 2026.
S36 National Institute of Standards and Technology, U.S. Department of Commerce. 2023.
S14 Journal of the American Statistical Association, American Statistical Association. 2007.
S15 Proceedings of the 35th International Conference on Machine Learning, PMLR. 2018-07-03.
S20 arXiv; journal version in International Journal of Forecasting, Elsevier. 2022-09-23.
S17 Journal of the American Statistical Association, Taylor & Francis; abstract through RePEc IDEAS. 2019.
S40 Google Cloud Documentation. 2024-08-28.
S18 International Journal of Forecasting, Elsevier; abstract through Lancaster EPrints. 2005.
S25 arXiv. 2024-11-04.
S26 arXiv. 2024-04-17.
S13 scikit-learn. 2026-10-30.
S12 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2006.
S30 Management Science, INFORMS; abstract through EconPapers. 1969.
S22 arXiv. 2021-03-26.
S38 ACM Computing Surveys; open record from Bournemouth University. 2014-04-30.

02

PART I: DEFINE THE DECISION AND THE PRIOR / FORECAST CONTRACT

The forecast target must be demand, not a convenient sales extract

A model cannot repair a target that confuses orders, sales, availability and consumption.



CHAPTER 02
DECISION TEST

A model cannot repair a target that confuses orders, sales, availability and consumption.

A greenfield apparel manufacturer should never train a forecasting model on raw accounting invoices or unadjusted point-of-sale extracts, because recorded sales measure the intersection of customer desire and physical supply rather than demand itself. [S06] [S07] Before a technical team selects an analogue prior, designs a launch pilot or writes machine-learning code, commercial and production leaders must sign an explicit forecast contract. [S08] [S36] [S0202] That contract defines the event being predicted, separates customer orders from wholesale shipments and retail sell-through, specifies how stockouts and returns are treated, fixes the product, channel and temporal grain, locks the information cut-off for each planning clock, classifies zero-observation periods, and names the human owner whose decision changes when the forecast shifts. [S06] [S07] [S18] [S36] [S37] Without this specification, even an advanced neural architecture learns to replicate the factory's own stockouts, allocation errors and invoicing delays as if they were laws of consumer preference. [S06] [S07] [S37]

In the Warp and Signal operating philosophy of Green Future Tech Limited (GFT), Hong Kong SAR, production is the warp and measurement is the weft. Within GFT's three-part Cameroon textile programme, prepared on 31 January 2024 with source refreshes flagged through 2026, this chapter extends the programme's performance-management system design into commercial demand events, establishing the measurement boundary that precedes statistical modelling. [R06] [R07] Just as a spinning or weaving mill cannot manage mechanical yield without defining where fibre waste is weighed, an apparel enterprise cannot forecast demand until it defines the boundary between an unfulfilled request, a shipped carton and a retained garment. [S06] [S0202] This framework is an operational engineering specification; it does not constitute investment, legal or tax advice, or a guarantee of commercial outcomes.

2.1

Unavailable stock can turn strong demand into weak sales

The foundational error in retail and manufacturing demand planning is equating historical sales records with customer demand. [S06] [S07] In an unconstrained setting where every style, colourway and size is perpetually on the shelf, completed transactions equal realised demand. In real garment operations, and especially during the capital-constrained launch runs of a new manufacturer, inventory is finite and unevenly distributed across sizes and locations. [S01] [S06] Whenever a garment variant stocks out, the transaction ledger stops recording purchases even while prospective buyers continue arriving at the counter or browsing the catalogue. [S06] [S07] Treating those truncated sales rows as the training target introduces a systematic downward distortion: the forecasting system mistakes an empty shelf for an uninterested market. [S06] [S07]

Methodological reviews of retail forecasting confirm that estimating demand solely from sales data produces a negative bias for the focal product. [S06] Dataset engineering for censored retail settings states the mechanism directly: the sales-equals-demand assumption collapses during stockouts because observed volumes underrepresent latent demand. [S07] In apparel, size curves and substitution compound this censoring. [S02] [S06] If an outlet in Douala runs out of medium and large cotton shirts

in week two while extra-small shirts remain on the rack, raw sales totals make the style appear to stall after week two and make extreme sizes appear artificially popular. [S03] [S06] Training on unadjusted sales cuts popular styles and shifts size ratios toward slow movers, creating lost revenue and dead stock. [S06] [S07]

2.2 CENSORED OBSERVATIONS CORRUPT BOTH ANALOGUE PRIORS AND TARGET UPDATES

Stockout censoring harms a cold-start manufacturer at two points. First, before launch, uncorrected stockouts in external reference series pull the outside-view prior downward, which is why cold-start architectures pair visual similarity clustering with demand correction before transferring profiles. [S05] [S06] Second, during initial pilots, failing to log stockout timestamps contaminates the likelihood update that turns the prior into a posterior distribution. [S06] [S07]

EXHIBIT 05 / SUPPLY-CHAIN AND DEMAND MEASUREMENT LAYERS

Demand becomes learnable through four distinct event layers



DERIVED [S06] [S07] [S0201] Derived conceptual architecture separating availability and latent demand, customer order capture, wholesale fulfilment dispatch and post-sale returns or cancellations, grounded in retail forecasting, censored-demand and supply-chain distortion literature.

Exhibit 5, p. 33 organises the commercial measurement boundary into four layers derived from retail demand and supply-chain research. [S06] [S07] [S0201] At the top sits the latent demand and availability layer, tracking whether an item is listed, whether physical units are available by size, what substitute garments are on the fixture, and whether the interval is censored. [S06] [S07] Beneath it sits the customer order and sell-through layer, capturing when a buyer commits to a purchase at a recorded effective price within a named channel. [S03] [S06] The third layer records wholesale shipment and factory fulfilment events, which follow warehouse batching and logistics schedules rather than real-time consumer arrival. [S06] [S0201] The bottom layer captures pre-dispatch cancellations, post-delivery returns, defect-versus-fit reason codes, and the rule attributing reversals to a calendar period. [S06]

What **Exhibit 5**, p. 33 establishes is that a learnable forecast target cannot be extracted from the third or fourth layer without joining it to the first and second layers. [S06] [S07] What the diagram cannot establish on its own is the numerical magnitude of unrecorded demand during a stockout; recovering latent demand requires observed sub-daily velocity prior to stock exhaustion, waitlist capture, or pilot inventory buffers designed to prevent censoring during the test window. [S06] [S07]

2.3 AVAILABILITY BELONGS INSIDE THE TARGET EQUATION RATHER THAN IN AN ANALYST FOOTNOTE

To make availability operational, the forecast contract defines an availability exposure variable a between 0 and 1 for every forecasted product variant, channel and time bucket. [S06] [S07] In a digital portal, a measures the fraction of active selling hours in which the variant was listed as in-stock. [S07] In a physical store or market stall, a is approximated from opening stock, replenishment timestamps and closing counts, flagging any period where closing inventory reaches zero or falls below a display minimum. [S06] [S07] When $a = 1$, observed gross sales provide an uncensored observation of latent demand; when a lies between 0 and 1, the observation is right-censored; when $a = 0$, the period carries zero information about customer preference and must be masked during model fitting. [S06] [S07] [S37]

2.4

Orders, shipments, sell-through and consumption answer different questions

Even when stock is available, departments inside an apparel firm routinely use the word "demand" to describe four incompatible events: upstream factory orders placed by institutional buyers or wholesalers, factory gate shipments invoiced to distributors, downstream retail sell-through to final purchasers, and garment usage by wearers. [S06] [S0201] A forecast contract that leaves the measured event ambiguous invites commercial, manufacturing and finance teams to feed different columns into the same pipeline, producing incoherent forecasts. [S08] [S0202]

Supply-chain theory explains why upstream orders must never be equated with downstream consumer sell-through. [S0201] Lee, Padmanabhan and Whang demonstrate that information transferred upstream as orders is systematically distorted by four mechanisms: demand signal processing, rationing games during perceived shortages, order batching to meet minimum transport or production quantities, and forward buying driven by price variations. [S0201] Because of this bullwhip effect, the variance of wholesale orders arriving at a factory gate exceeds the variance of retail garment sales, and its timing reflects distributor inventory rules rather than consumer buying rhythms. [S06] [S0201]

2.5 INDUSTRIAL CONTEXT IN CAMEROON HIGHLIGHTS THE COST OF CONFUSING CAPACITY, SHIPMENTS AND SELL-THROUGH

The distinction between production capacity, wholesale shipment and end-market absorption is sharp in Cameroon's textile sector. Inherited sector evidence shows that Cameroon harvests 314,455 to 340,000 tonnes of seed cotton annually across 2022 to 2024 (with 2025 projections above 350,000 tonnes) across more than 220,000 producers supported by SODECOTON, yet exports roughly 95 percent of its cotton fibre raw and processes about 5 percent locally, against an SND30 target of 50

percent by 2030. [R01] [R02] [R03] [R06] [R07] While major projects such as Camtext SA (announced at US$300 million, or 180 billion XAF, to process 12,000 tonnes a year) expand transformation capacity, domestic history warns that capacity does not guarantee sell-through. [R09] [R10] [R11] Public reporting notes that CICAM's market share fell from about 85 percent three decades ago to about 5 percent, with July 2023 utilisation at 33.7 percent in Garoua, 48.9 percent at CICAM 1 and 10.83 percent at CICAM 2, and underutilisation costs of about 2 billion XAF in 2020. [R11] [R12]

Meanwhile, Statista estimates Cameroon's clothing and footwear spend at about US$94.70 (around 8.0 pieces) per person in 2024, with apparel e-commerce revenue projected to rise from US$20.32 million in 2023 to US$33.61 million by 2027 at 13.41 percent a year as penetration grows from 18.0 percent to 22.7 percent. [R26] A Cameroonian maker selling across institutional tenders, wholesale distributors and direct digital channels cannot merge these streams into one series. [S04] [S06] Adding a 2,000-unit quarterly wholesale order in Yaounde to 45 direct online consumer sales in week one fabricates a 2,045-unit spike followed by eleven weeks of apparent collapse. [S04] [S06] [S0201]

2.6 CHANNEL RECONCILIATION RULES PREVENT SILENT AGGREGATION ERRORS

To prevent bullwhip distortion and channel mixing, the contract establishes three separation rules: [S04] [S06] [S0201]

1. **Direct-to-consumer and consignment channels** use customer order timestamps (for gross demand) and delivery handover (for fulfilled sell-through), paired with availability logs. [S04] [S06] [S07]
2. **Wholesale and distributor channels** separate distributor sell-in (factory order and dispatch) from downstream sell-through. [S06] [S0201] Where downstream sell-through is unavailable, wholesale orders are modelled as an intermittent batch-replenishment process, never pooled with consumer transactions. [S18] [S0201]
3. **Institutional contract channels** (uniforms, linens, workwear or public tenders) follow tender awards, headcount schedules and call-off windows in an order-book pipeline rather than retail SKU series. [S01] [S08] [S11]

2.7

Returns and cancellations require event-based target definitions

A transaction is a sequence of state transitions that can reverse weeks after the initial order. [S06] [S37] In apparel commerce, two reversal events separate initial order capture from retained revenue: pre-dispatch cancellations (from payment failure, buyer remorse, credit rejection or factory stock unavailability) and post-delivery returns or exchanges (from size misfits, shade or stitching defects, or late delivery). [S06] If a data pipeline overwrites historical sales rows in place whenever a cancellation or return arrives, every retrospective backtest becomes irreproducible and leaks future outcome information into earlier forecast origins. [S37] [S40]

Instead of mutating a single sales column, the forecast contract defines an immutable event log in which every unit transition carries an event timestamp and a parent transaction identifier. [S37] [S40] From this log, three demand measures are derived for specific operational decisions: [S06] [S08]

- **Gross demand:** Uncensored customer purchase intent expressed at the order timestamp, adjusted for stockout availability, excluding duplicate payment retries, before deducting supply-driven cancellations or customer returns. [S06] [S07] Merchandising teams use gross demand to evaluate whether the garment concept and price offer attracted buyers. [S02] [S03]
- **Fulfilled gross shipments:** Units picked, packed and dispatched from the factory or warehouse at the dispatch timestamp, used by logistics schedulers to plan handling capacity. [S06] [S08] [S0202]
- **Net retained demand:** Gross demand minus customer-initiated cancellations and non-defective customer returns, attributed according to a fixed temporal rule and used by finance and replenishment planners to determine net manufacturing requirements. [S06] [S08]

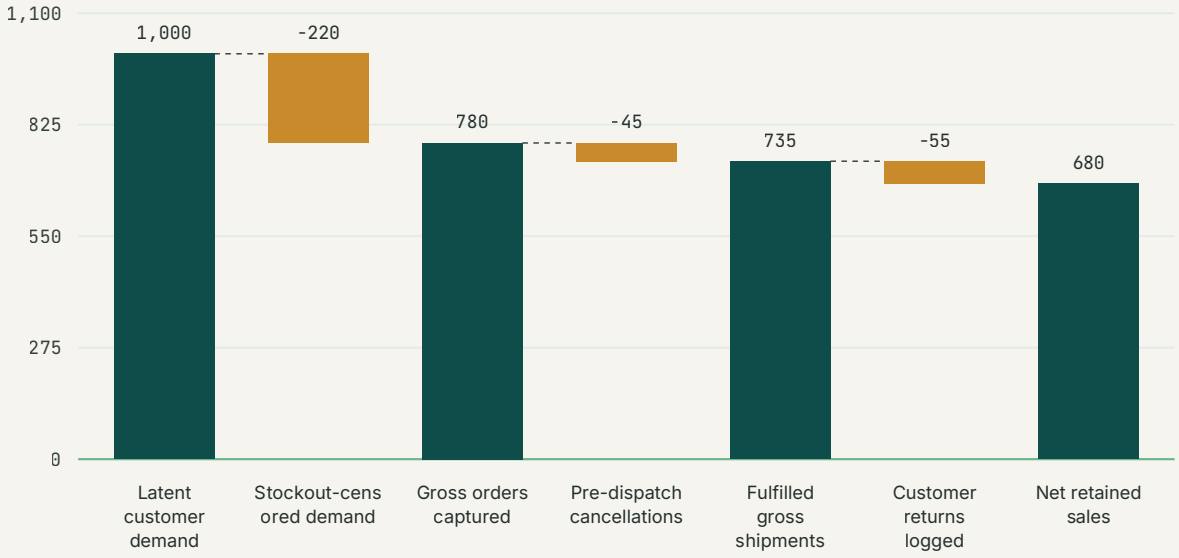
2.8 TEMPORAL ATTRIBUTION RULES MUST PREVENT HINDSIGHT LEAKAGE

When a customer buys a garment in week `t` and returns it in week `t + 3`, the contract must state whether that return reduces the demand target of week `t` (origin-period attribution) or enters as a returned-inventory inflow in week `t + 3` (arrival-period attribution). [S06] [S37] For evaluating style attractiveness and net demand, origin-period attribution links the return to the purchase cohort of week `t`, stored as a bitemporal table indexed by both target week and observation cut-off so a model replayed at week `t` sees only returns recorded by that date, paired with a prior return-rate distribution for immature cohorts. [S13] [S14] [S37] For physical warehouse replenishment, arrival-period attribution records returned garments inspected and passed as first-quality stock at week `t + 3` as a positive inventory inflow, keeping the stock balance exact. [S06]

Similarly, the contract separates return and cancellation reasons by root cause. [S06] If an order is cancelled because the factory failed to deliver on time or stocked out of the requested size, deducting that cancellation from demand blames customer preference for a manufacturing failure. [S06] [S07] Supply-caused cancellations and manufacturing-defect returns remain inside gross and latent demand, while triggering a quality or fulfilment exception under GFT's performance-management system. [R29] [S06]

EXHIBIT 06 / ILLUSTRATIVE GARMENT UNITS ACROSS A 4-WEEK PILOT

An unadjusted invoice extract understates true launch demand



ILLUSTRATIVE [S06] [S07] Illustrative arithmetic reconciliation of a 1,000-unit latent demand pilot across stockout censoring, pre-dispatch cancellations and customer returns. Values are constructed to demonstrate the contract accounting method, not an observed GFT factory result.

Exhibit 6, p. 37 shows how a four-week pilot launch can produce four different headline figures depending on which event boundary an analyst extracts. [S06] [S07] In this illustrative pilot, true underlying customer demand is 1,000 garment units. Because medium and large sizes stock out midway through the test, 220 units of latent demand go unrecorded, leaving 780 gross orders captured. [S06] [S07] Pre-dispatch cancellations remove 45 units, resulting in 735 fulfilled shipments. Post-delivery returns remove 55 units, leaving 680 net retained sales on the accounting ledger. Mistaking the 680-unit net sales extract for unconstrained demand sizes the next run 32 percent below latent customer interest; mistaking 780 gross orders for net cash collection overstates revenue by 100 units. **Exhibit 6**, p. 37 illustrates the accounting rule only; in live operations, the unobserved 220-unit stockout wedge must be estimated with uncertainty bounds rather than treated as a deterministic count. [S07] [S14]

WORKED EXAMPLE: RECONCILING CENSORED SALES, CANCELLATIONS AND RETURNS IN A LAUNCH LOT

Consider an illustrative four-week pilot launch for a cotton workshirt priced at 8,500 XAF across direct digital and retail counter channels. The factory releases an 800-shirt lot across four sizes. Order logs capture 780 gross orders (`780 × 8,500 = 6,630,000 XAF`). Availability telemetry shows Size L and Size XL stocked out after 78 percent of weighted selling hours (`a = 0.78`), while Sizes S and M remained in stock (`a = 1.00`) with 20 unsold units. [S06] [S07]

Before dispatch, 45 ordered units are cancelled: 25 from mobile-money payment timeouts (buyer-side) and 20 from a four-day sewing delay (supply-side). The warehouse dispatches `780 - 45 = 735` fulfilled shirts (`6,247,500 XAF`). Within the 14-day maturity window, customers return 55 shirts: 40 for fit preferences (35 pass inspection and re-enter stock) and 15 for a seam defect. [R29] [S06] Net retained sales equal `735 - 55 = 680` shirts (`5,780,000 XAF`).

Under the GFT contract, the planner does not feed 680 or 735 into the model as unconstrained demand. [S06] [S07] Exposure-rate reconstruction on stocked-out sizes (`780 ÷ 0.78 = 1,000` illustrative gross units, with a 10th to 90th percentile interval of 910 to 1,120 units) estimates gross latent demand at 1,000 units (`1,000 - 780 = 220` censored units). [S07] [S14] Deducting only buyer-side cancellations (`25 ÷ 780 = 3.21 percent`) and fit returns (`40 ÷ 735 = 5.44 percent`) while routing supply cancellations and defect returns to factory corrective action gives a retention factor of `(1 - 0.0321) × (1 - 0.0544) = 0.9153`, yielding 915 net latent shirts. [R29] [S06] With 35 restocked fit returns on hand, the factory cuts `915 - 35 = 880` new shirts. Using the 680-unit accounting extract would under-produce by 235 shirts (`1,997,500 XAF` illustrative gross revenue). All figures are illustrative calculations demonstrating contract rules, not observed GFT factory outcomes.

2.9

Product grain must separate style, colour, size and channel

Defining the target event solves what is measured; defining the forecast grain solves the structural resolution at which that event is indexed and aggregated. [S0202] In apparel manufacturing, a product is a nested hierarchy of technical and commercial decisions: product family (woven tops), style or silhouette (short-sleeve workshirt), fabric construction (100 percent cotton twill at 180 grams per square metre), colour-way or dye lot (indigo or khaki), size (S, M, L, XL), and sales channel (direct e-commerce, Douala retail counter, regional wholesale distributor or institutional contract). [S02] [S03] [S04]

Machine-learning research on unseen fashion items shows why this attribute and grain hierarchy must be locked in the contract before data capture begins. [S02] [S03] Attribute-based demand models infer the sales trajectory of a new item by learning how clothing attributes (category, fabric, colour, cut) and merchandising factors (selling price, product age, store context, promotion) drove demand across prior items. [S02] [S03] If an enterprise table concatenates style, colour and size into an unparsed text string, or reuses SKU codes across seasons when fabric weight or price changes, the model loses the structural bridges needed to share statistical strength across variants. [S02] [S17] [S22]

2.10 SURROGATE KEYS AND IMMUTABLE ATTRIBUTE VERSIONS PROTECT TRANSFER LEARNING

To preserve cross-product learning and hierarchical coherence, the forecast contract enforces four schema rules across the product and channel hierarchy: [S02] [S03] [S11] [S17]

1. **Separated composite grain keys:** Every demand record carries explicit foreign keys for family, style, fabric, colour, size, channel and region, alongside the atomic SKU identifier (`style × fabric × colour × size`). [S02] [S03] [S11]
2. **Attribute immutability and versioning:** If a garment's fabric blend, weight, construction specification or visual appearance changes, it receives a new style or SKU version rather than overwriting the existing attribute record. [S02] [S05] [S37]
3. **Channel and store separation:** Because multichannel retail studies show that online, wholesale and physical store channels respond differently to calendar events, promotions and regional conditions, the channel identifier is a mandatory primary key in the demand table. [S04] [S11]
4. **Coherent aggregation mapping:** Size-level forecasts aggregate into style-by-colour dye lots, style-level cutting plans, fabric procurement requirements (via bill-of-materials yardage per garment), and channel revenue budgets, preparing the hierarchy for trace minimisation (MinT) reconciliation across levels. [S11] [S17]

> METHOD
>
> **Contract rule for size-curve decoupling in cold-start apparel launches**
>
> At the pre-launch and early pilot stage, forecasting every `style × colour × size × channel × week` cell independently produces count sparsity and high variance. [S18] [S0203] The GFT forecast contract decouples the target into two linked components recombined through hierarchical reconciliation: [S11] [S17] 1. **Style-colour-channel volume:** Modelled as the primary demand series in garment units per week at the `style × colour × channel` grain, where analogue priors, price offers and channel covariates act directly. [S02] [S03] [S04] 2. **Conditional size-share vector:** Modelled via a Dirichlet-multinomial prior informed by anthropometric data, institutional head-count rosters or category-level size sales, updated exclusively using uncensored periods where all sizes of the style were simultaneously in stock (`a = 1` across all sizes). [S06] [S07] [S28]
>
> Updating the size curve exclusively on all-sizes-available windows prevents a stock-out in Size L from inflating the learned share of Size S. [S06] [S07]

2.11

The horizon must match buying, production and replenishment clocks

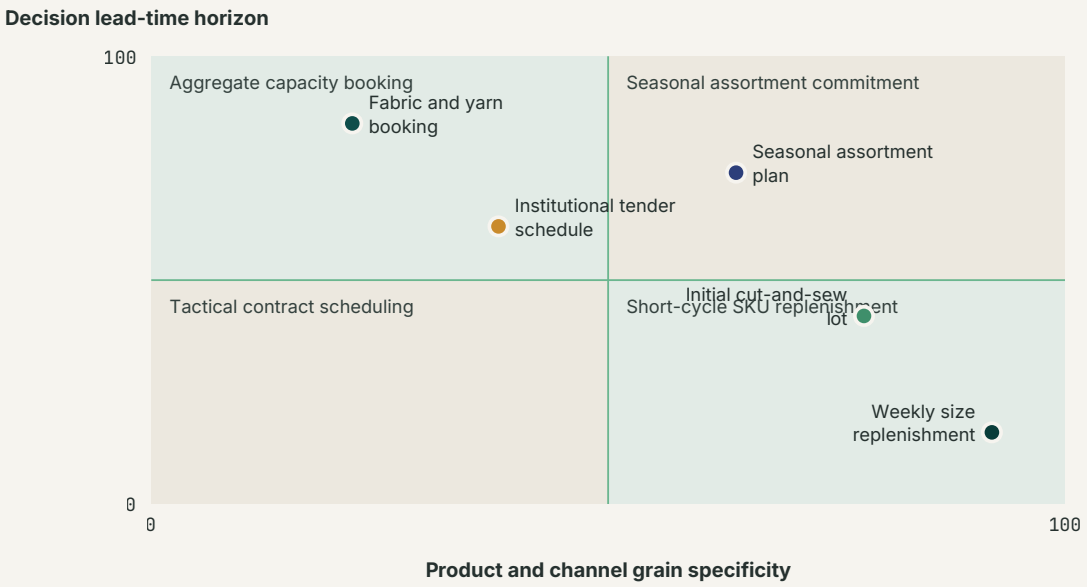
A forecast is actionable only if it arrives before the physical lead time of the decision it guides. [S08] [S0202] Forecasting theory and practice emphasise that method selection, temporal aggregation and evaluation must be tied directly to the planning horizon and decision context. [S08] [S0202] Short-term forecasts support personnel scheduling, machine loading and transport dispatch; medium-term forecasts determine raw-material purchasing and batch sizing; long-term forecasts guide capacity and market

strategy. [S0202] Evaluating a sixteen-week fabric-procurement forecast using one-week-ahead point accuracy is a methodological error: a model that achieves low one-week error by echoing last week's sales provides zero utility when yarn must be ordered four months in advance. [S08] [S12] [S33]

In an apparel manufacturing venture, four operational clocks run in parallel, and the forecast contract assigns each clock its own target grain, forecast horizon, update cadence, information cut-off and decision owner. [S01] [S08] [S0202]

EXHIBIT 07 / ILLUSTRATIVE SPECIFICITY AND HORIZON INDEX FROM 0 TO 100

Each planning clock needs its own forecast grain



ILLUSTRATIVE [S02] [S03] [S06] [S08] [S0202] Illustrative mapping of five apparel planning decisions across product-and-channel grain specificity (horizontal axis, from aggregate fabric to atomic SKU-by-location) and required decision lead-time horizon (vertical axis, from immediate weekly replenishment to multi-month raw-material booking). Positions are conceptual planning coordinates from 0 to 100, not empirical accuracy scores.

Exhibit 7, p. 40 maps five apparel commitments across two conceptual dimensions from 0 to 100: product-and-channel grain specificity on the horizontal axis and decision lead-time horizon on the vertical axis. [S02] [S03] [S06] [S08] [S0202] In the upper-left quadrant sits fabric and yarn booking (illustrative coordinates 22 and 85), reserving cotton yarn counts and greige fabric weights 16 to 26 weeks ahead at a fabric-family grain. [S01] [S0202] In the upper-right quadrant sits the seasonal assortment plan (coordinates 64 and 74), committing style counts and colourways 12 to 20 weeks before launch. [S02] [S08] Near the centre sits institutional tender scheduling (coordinates 38 and 62), reserving sewing capacity 8 to 16 weeks ahead. [S08] [S0202] In the lower-right quadrant sit initial cut-and-sew lot sizing (coordinates 78 and 42, 4 to 8 weeks ahead at style-by-colour-by-size grain) and weekly size replenishment (coordinates 92 and 16, 1 to 3 weeks ahead at SKU-by-channel grain). [S03] [S06]

What **Exhibit 7**, p. 40 establishes is that one horizon cannot serve all five clocks; a forecast service must publish multi-horizon distributions reconciled across the hierarchy. [S16] [S17] [S24] What the matrix cannot establish is exact factory lead times

without a supply-chain audit, as lead times depend on whether the maker buys finished fabric, contracts local dyeing, or integrates spinning and finishing. [R09] [R10] [S01]

2.12 INFORMATION CUT-OFFS AND FREEZE WINDOWS PREVENT CHRONOLOGICAL LEAKAGE

Every horizon in the contract requires an explicit forecast origin, an information cut-off timestamp, and a production freeze window. [S13] [S37] Machine-learning reproducibility research defines data leakage as a spurious relationship between independent variables and the target arising from collection, sampling or preprocessing strategies that expose information unavailable at prediction time. [S37] In multi-horizon apparel forecasting, leakage occurs whenever an analyst evaluates a 6-week-ahead cutting decision using inventory, price, weather or return records from the intermediate five weeks. [S13] [S37]

Physical factories also impose freeze windows: if cutting and sewing require two weeks from fabric issue to garment packaging, a replenishment forecast issued today cannot alter finished goods arriving in weeks one or two. [S06] [S08] Its actionable target is cumulative demand after the two-week freeze window across the review cycle, evaluated across the nine-quantile structure of the M5 Uncertainty competition (or a contracted subset such as the 0.10, 0.25, 0.50, 0.75 and 0.90 quantiles). [S14] [S16]

2.13

Sparse zeros need a meaning before they need a model

Once an apparel maker moves from style totals to size-by-colour-by-channel weekly records, its tables fill with zeros. [S18] [S0203] In intermittent demand research, demand appears sporadically, with many periods showing no demand at all. [S18] When counts are small (0, 1, 2 or 3 garments a week at a regional counter), continuous Gaussian assumptions fail and count or intermittent-demand estimators are required. [S18] [S19] [S0203]

John Croston's method addresses sporadic demand by constructing two series from the original series: nonzero demand size and inter-arrival time (periods between successive nonzero demands), applying separate exponential smoothing updates to each and forecasting demand per period as their ratio. [S18] [S0203] Because the original Croston ratio is biased upward, Syntetos and Boylan introduced a modified estimator that multiplies the ratio by a damping factor to give approximately unbiased demand-per-period estimates. [S19] [S0203]

2.14 UNCLASSIFIED ZEROS CORRUPT INTER-ARRIVAL TIMES IN CROSTON AND MACHINE-LEARNING MODELS

A severe distortion occurs when engineers apply Croston's method, the Syntetos-Boylan correction, or a global autoregressive model to raw database tables in which every empty cell is filled with zero. [S06] [S18] [S37] [S0203] In Croston-type estimators, every zero period increments the inter-arrival counter by one period. [S0203] If an item records four consecutive zeros because it was unlaunched, paused online, stocked out in Size M, or disconnected from the point-of-sale network, feeding those zeros

into the estimator tells the algorithm that customers rejected an available garment for four weeks, inflating the inter-arrival estimate and collapsing the forecast. [S06] [S07] [S18] [S19] [S0203]

EXHIBIT 08 / CONTRACT TARGET AND ZERO-STATE DEFINITIONS

A target dictionary prevents silent metric substitution

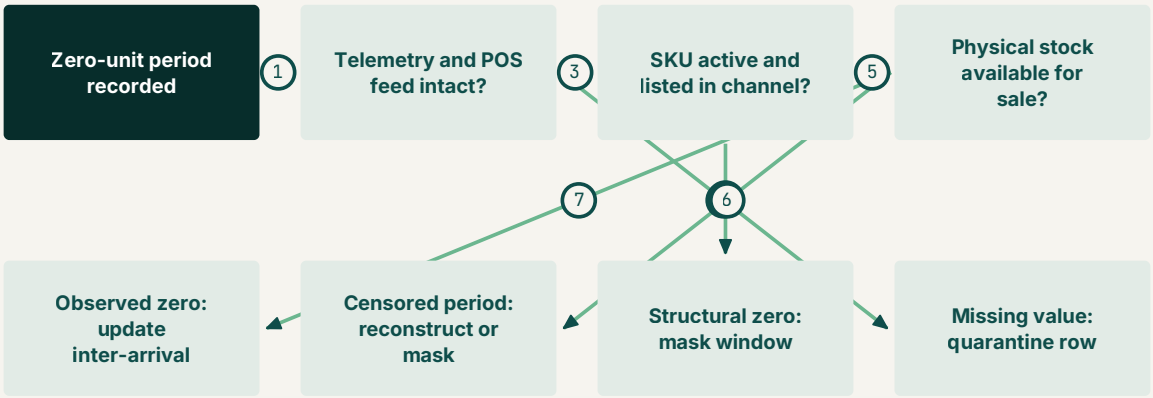
Contract term	Event boundary and timestamp	Permitted planning use
Gross orders	Customer order submission timestamp	Gross commercial interest and payment funnel audit
Fulfilled shipments	Warehouse dispatch and invoice timestamp	Logistics capacity and working-capital billing
Net sell-through	Final purchase minus origin-dated fit returns	Commercial revenue and style retention review
Latent demand	Available sell-through plus stockout recovery	Production lot sizing and replenishment target
Observed zero	Zero units sold while listed and in stock	Valid input to intermittent demand estimators
Structural zero	Item unlisted, pre-launch or channel closed	Excluded from active selling window and denominator
Censored zero	Zero units sold during verified stockout	Flagged for availability correction or interval masking
Missing value	POS, telemetry or sync feed failure	Quarantined from training until pipeline recovery

ILLUSTRATIVE [S06] [S07] [S18] [S19] Illustrative GFT target dictionary template grounded in retail forecasting, stockout-censored demand and intermittent-demand literature. Rows define mandatory schema states rather than observed GFT factory volumes.

Exhibit 8, p. 42 sets out the target and zero-state dictionary that every GFT forecasting pipeline enforces before feature engineering or baseline fitting begins. [S06] [S07] [S18] [S19] The top four rows lock positive volume boundaries (gross orders, fulfilled shipments, net sell-through and latent demand), while the bottom four rows partition every zero or empty period into four mutually exclusive states: observed zero, structural zero, censored zero and missing value. [S06] [S07] [S18] While **Exhibit 8**, p. 42 prevents silent metric substitution across teams, the dictionary alone cannot guarantee that store staff scan stockouts accurately; physical audit checks during Phase 1 pilots remain essential to verify that censored zeros are not mislabelled as observed zeros. [S06] [S07] [S40]

EXHIBIT 09 / ZERO-RECORD CLASSIFICATION SEQUENCE

Every zero record must pass an availability and listing audit before modelling



1. Inspect log
2. Feed failed
3. Feed valid
4. Unlisted
5. Listed
6. Stocked out
7. In stock

DERIVED [S06] [S07] [S18] [S19] [S0203] Derived ingestion decision sequence for classifying a zero-unit period before fitting Croston, Syntetos-Boylan or global demand models, synthesising intermittent-demand and stockout-censoring rules.

Exhibit 9, p. 43 translates the dictionary into an automated ingestion gate executing on every zero-unit row before any model is called. [S06] [S07] [S18] [S0203] First, the gate checks whether the channel telemetry feed reported heartbeat events; if a shop in Garoua or Douala lost connectivity, the cell is quarantined as a missing value. [S37] [S40] Second, if the feed was intact, the gate checks the listing calendar: pre-launch, delisted or closed-holiday periods become structural zeros and are masked from the selling window. [S03] [S04] Third, for active listings, if opening stock was zero (`a = 0`), the row becomes a censored zero routed to latent-demand reconstruction or masked from loss calculation. [S06] [S07] Only when the feed is valid, the offer is listed and stock is confirmed on hand does a zero-sale period graduate to an observed zero eligible to increment the inter-arrival time in Croston or Syntetos-Boylan estimators. [S18] [S19] [S0203] Partial-period censoring (`0 < a < 1`) is handled by scaling exposure time or flagging periods with `a < 0.80` as partially censored. [S07]

> **WARNING**
>
> Never execute an unconditional zero-fill across a full Cartesian grid of all SKUs, channels and calendar weeks before model training. [S18] [S37] In a cold-start apparel catalogue where styles launch in staggered drops across the year, padding pre-launch and post-withdrawal weeks with unconditional zeros injects hundreds of fabricated zero-demand observations into the training set. [S03] [S37] This single query mistake biases naive seasonal denominators downward, ruins MASE scaling, and teaches global neural networks that every garment experiences months of zero demand before suddenly spiking. [S12] [S22] [S37]

2.15

A signed forecast contract makes accountability operational

A data dictionary and a target equation protect the codebase only if they are backed by human governance. [S36] The National Institute of Standards and Technology (NIST) Artificial Intelligence Risk Management Framework emphasises that human judgement must be employed when deciding on the specific metrics related to system trustworthiness and the precise threshold values that trigger intervention. [S36] Similarly, machine-learning engineering guidance shows that optimising a mathematical loss function detached from system and decision outcomes wastes engineering effort, while unmanaged pipeline dependencies create hidden technical debt that overwhelms small industrial teams. [S33] [S39]

In GFT's methodology, no forecasting work begins until commercial, production and digital leads sign a one-page Forecast Contract Charter for each active planning clock. [S08] [S36] [S0202] The charter is both a managerial agreement and a machine-readable configuration schema consumed by the validation and release pipeline. [S40] Every signed contract records eight mandatory fields: [S12] [S14] [S36] [S37] [S40]

1. **Decision and owner:** The named role and the physical or financial commitment controlled by the forecast (such as releasing a 500-shirt cut-and-sew order or booking 3,000 metres of twill). [S08] [S36] [S0202]
2. **Target event and availability rule:** The target series from **Exhibit 8**, p. 42 (net latent demand for production lots, gross latent demand for assortment review) and the minimum availability threshold flagging stockouts. [S06] [S07]
3. **Grain and hierarchy keys:** The prediction grain (`style × colour × channel × week` or `sku × channel × week`) and parent aggregation tree for hierarchical reconciliation. [S11] [S17]
4. **Horizon, origin and freeze window:** The forecast origin, information cut-off timestamp, manufacturing freeze window and coverage horizon. [S08] [S37]
5. **Required output quantiles:** The published probability levels (such as the 0.10, 0.25, 0.50, 0.75 and 0.90 quantiles) weighing overstock against stockout risk. [S14] [S16]

6. **Evaluation metrics and undefined-denominator rules:** Mean Absolute Scaled Error (MASE) where the as-of naive denominator is nonzero, paired with MAE, relative error against the frozen analogue prior and bias for zero-history series, plus quantile loss and coverage. [S12] [S14] [S15] [S16]
7. **Escalation and rollback thresholds:** Maximum tolerated stockout censoring, missing-feed rate, calibration drift or relative error above which the pipeline reverts to the approved prior or baseline. [S33] [S36] [S40]
8. **Pre-launch decision test:** The forecast shift required to alter the commitment; if an opening lot is fixed at 1,000 units regardless of the forecast, no modelling resource is spent pretending otherwise. [S08] [S33] [S39]

> **GFT NOTE**
>
> In GFT's Cameroon textile programme, a small industrial team cannot afford to maintain five competing spreadsheets defining "sales" differently for finance, merchandising, spinning, sewing and logistics. [R06] [S39] GFT recommends implementing the Forecast Contract as a single version-controlled schema table before Phase 1 pilot production begins, linking demand event definitions directly to the factory's performance-management KPI dictionary. [S36] [S40] When commercial and production managers sign the contract before launch, post-mortem reviews shift from debating whose spreadsheet is right to inspecting measured demand against the pre-agreed prior. [S08] [S36]

2.16 THE CONTRACT CONNECTS CHAPTER 1'S COLD-START POSITION TO CHAPTER 3'S ANALOGUE PRIOR

With the forecast contract signed and the event dictionary locked, the manufacturer has an auditable target definition. [S06] [S07] [S36] Models predict availability-corrected latent demand rather than stockout-capped invoices, wholesale batches stay separate from retail sell-through, returns carry bitemporal timestamps, and structural or stockout zeros are separated from observed zero demand. [S06] [S07] [S18] [S37] [S0201] Chapter 3 applies these same availability, attribute and channel rules to external reference classes to build the initial analogue prior. [S01] [S02] [S05] [S28]

A forecast is only as honest as the event it measures: when a factory confuses what it managed to ship with what customers wanted to buy, its models learn to repeat its own constraints.

CHAPTER SOURCE NOTES

S06 Lancaster University Management School, Munich Personal RePEc Archive. 2018-11-19.

S07 arXiv. 2026.

S08 International Journal of Forecasting, Elsevier; full text through arXiv. 2022.

S36 National Institute of Standards and Technology, U.S. Department of Commerce. 2023.

S0202 OTexts. 2021-05-31.

S18 International Journal of Forecasting, Elsevier; abstract through Lancaster EPrints. 2005.

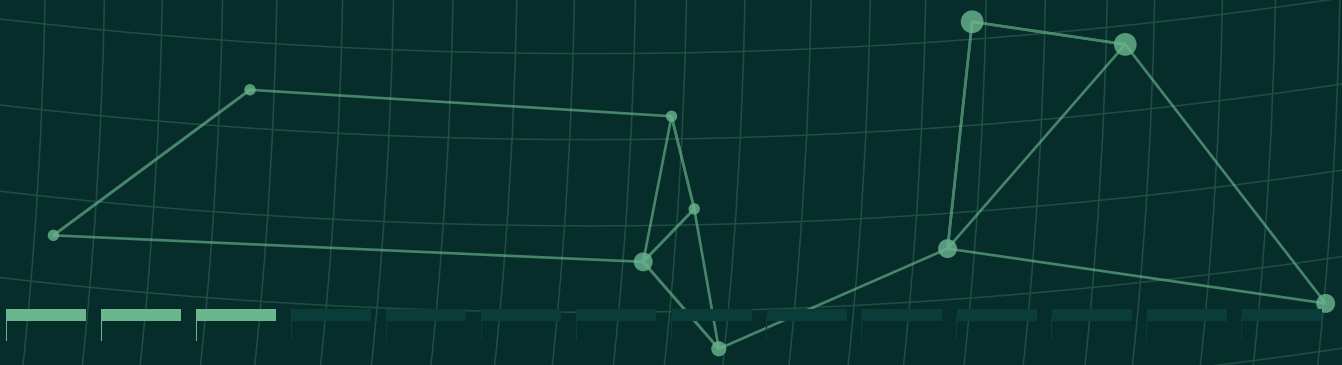
S37 Cell Press, Patterns; open full text from PubMed Central. 2023-08-04.
R06 Cameroon Trade Hub. Publication date unrecorded; inherited source-text check.
R07 Ecofin Agency. Publication date unrecorded; inherited source-text check.
S01 InTech, International Journal of Engineering Business Management. 2013.
S02 arXiv. 2019-06-27.
S03 arXiv. 2020-07-10.
S05 Springer Nature, International Journal of Data Science and Analytics. 2026-01-08.
S0201 Management Science, INFORMS; abstract through RePEc IDEAS. 1997-04.
R01 Business in Cameroon. Publication date unrecorded; inherited source-text check.
R02 Fibre2Fashion; Business in Cameroon. Publication date unrecorded; inherited source-text check.
R03 SODECOTON. Publication date unrecorded; inherited source-text check.
R09 Ecofin Agency. Publication date unrecorded; inherited source-text check.
R10 Ecofin Agency. Publication date unrecorded; inherited source-text check.
R11 Business in Cameroon; MMI News. Publication date unrecorded; inherited source-text check.
R12 Business in Cameroon. Publication date unrecorded; inherited source-text check.
R26 Statista. Publication date unrecorded; inherited source-text check.
S04 MDPI, Processes. 2021-09-03.
S11 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2022.
S40 Google Cloud Documentation. 2024-08-28.
S13 scikit-learn. 2026-10-30.
S14 Journal of the American Statistical Association, American Statistical Association. 2007.
R29 MDPI; industry inspection literature. Publication date unrecorded; inherited source-text check.
S17 Journal of the American Statistical Association, Taylor & Francis; abstract through RePEc IDEAS. 2019.
S22 arXiv. 2021-03-26.
S0203 OTexts. 2021-05-31.
S28 Journal of Industrial Ecology, Wiley. 2022-10-17.
S12 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2006.
S33 Google for Developers. 2025-08-25.
S16 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2022.
S24 arXiv. 2020-09-27.
S19 International Journal of Production Economics, Elsevier; abstract through Lancaster EPrints. 2001.
S39 Neural Information Processing Systems Foundation, NIPS 2015. 2015.
S15 Proceedings of the 35th International Conference on Machine Learning, PMLR. 2018-07-03.

03

PART I: DEFINE THE DECISION AND THE PRIOR / ANALOGUE EVIDENCE

Analogue products provide the first outside view

A transparent set of comparable launches is more useful than a precise forecast whose provenance cannot be inspected.



CHAPTER 03
DECISION TEST

A transparent set of comparable launches is more useful than a precise forecast whose provenance cannot be inspected.

A new apparel maker should construct its first forecast from a documented reference class of launches, not from a memorable success story or an unexplained model score. The useful output is a distribution with provenance: which products and markets supplied evidence, why they were comparable, how availability affected their recorded sales, and how much weight each case received. Attribute matching, visual similarity and hierarchical pooling can help, but only inside that visible chain of reasoning. The resulting prior is a starting position for a reversible production decision, not a promise about realised demand.

This distinction matters because the absence of target sales does not remove the need for a forecast, while the presence of other products' sales does not establish transfer. Judgemental methods are legitimate when the item of interest has no observations, and probabilistic assessment remains preferable to an untraceable point estimate. [S08] [S14] The engineering task is therefore to make each analogy inspectable before giving it influence.

3.1

A reference class starts with comparable launches, not famous winners

The reference class is the evidence set from which the outside view is built. It should contain ordinary sellers, slow sellers, withdrawn products and launches that missed their plan alongside strong performers. Fashion demand already combines uncertainty, short histories and seasonal effects. Selecting only products that remained visible long enough to become case studies adds a further bias: the evidence set describes survival rather than the distribution of launch outcomes. [S01]

The unit of selection is a launch in context, not an isolated garment. Two offers can share a product code and still represent different launches when their prices, channels, dates or selling regions differ. Conversely, several colour variants can belong to one commercial launch and provide less independent evidence than their row count suggests. The corpus therefore records a launch identifier, the product variants it covers, the selling window and the conditions under which customers could buy. This prevents a large export of related rows from masquerading as a large number of independent analogues.

Eligibility rules should be written before outcomes are compared. A defensible rule might require a defined category, an observable launch date, a complete selling window, and sufficient availability records to distinguish weak buying from interrupted supply. A launch that fails the rule is logged with its exclusion reason rather than silently deleted. A launch that passes the rule is not excluded merely because its result is disappointing. Missing information remains missing; it is never converted into zero demand or an assumed average.

The evaluation corpus register gives every admitted case an owner, access right and extraction date. The owner is accountable for definitions and corrections. The access right states whether the team may use the record for analysis, retain derived features and show it to reviewers. The extraction date protects reproducibility when a source system later revises its history. The register also carries the complete launch window, availability correction, outcome status and any exclusion reason. Visual-transfer

research explicitly combines corrected historical demand with similarity matching, which makes the correction part of the evidence lineage rather than an optional cleaning step. [S05]

Availability requires particular care at both ends of the launch. Initial sell-in can exceed customer purchases, while a stockout can cap later sales. Returns, cancellations and delisting can also change what the recorded series means. The analogue curator should not invent a latent-demand value when the source lacks the fields needed to estimate one. Instead, the record identifies censored intervals, states the correction method where one is defensible, and carries the uncorrected and corrected versions together. A wide uncertainty contribution is more honest than a precise correction with no observable basis.

The corpus also needs dependence labels. Launches sharing a campaign, distribution footprint or institutional order may have common causes that a product-level table hides. Near-duplicates should be grouped before any count of supporting cases is reported. Where independence is doubtful, the prior should retain more between-launch uncertainty and less confidence in the apparent sample size. The report can show both the raw launch count and the effective independent groups without pretending that the latter is an exact statistical quantity.

Exhibit 10, p. 49 specifies the minimum register. Its purpose is to make survivorship, censoring and access limitations visible before modelling begins. It does not prove that any admitted launch is sufficiently similar to the target; that test comes next.

EXHIBIT 10 / REQUIRED REGISTER ENTRIES

A launch register keeps weak and failed cases in the evidence set

Register field	Required entry	Decision protected
Launch identity	Product, variants, market and selling window	Prevent duplicate evidence
Outcome coverage	Strong, ordinary, slow and withdrawn launches	Limit survivorship bias
Availability	Stock, listing and censored intervals	Separate demand from supply limits
Correction	Method, inputs, version and unresolved cases	Keep demand estimates auditable
Provenance	Owner, access right and extraction date	Permit lawful reproduction
Selection	Eligibility result and exclusion reason	Block post hoc case selection
Dependence	Shared campaign, channel and order groups	Avoid overstating support

ILLUSTRATIVE [S01] [S05] Illustrative GFT register derived as a practical template from fashion cold-start and corrected-demand research. Rows are required fields, not observed analogue records or performance results.

Adoption curves can inform the shape of an admitted launch, but they cannot rescue a weak reference class. Bass developed and tested his model on consumer durables, with initial-purchase timing related to previous buyers. [S30] That scope is not weekly apparel sell-through. A diffusion shape may be considered as one candidate for timing, aligned to the actual launch date and checked against disappointing cases. It should not set market potential, curve steepness or total sales by analogy to durable goods.

3.2

Attribute taxonomies make analogy auditable

Comparable launches become usable only when the reason for their similarity is explicit. Category, silhouette, fabric, colour, price band, occasion, channel and launch timing should be fields with definitions, not impressions attached to a product name. Research on unseen fashion items uses clothing, footwear and merchandising attributes to infer demand from other products, which supports attributes as a transfer route while still requiring a relevant historical catalogue. [S02]

The taxonomy should begin with the decisions the venture can actually make. A silhouette field is useful if designers and planners can apply its labels consistently. A price-band field is useful only if its boundaries and currency basis are fixed. An occasion field should distinguish work, school, ceremony, sport and other use cases only where the proposed assortment requires those distinctions. Excess detail creates missing values and reviewer disagreement without necessarily improving transfer.

Each field needs a controlled vocabulary, a definition and a missingness rule. Unknown, not applicable and not recorded are different states. A fabric description may be absent because the source never captured it, or because the garment genuinely combines materials that the vocabulary cannot represent. Collapsing those cases into a single other category would make unlike products appear similar. Versioning the taxonomy also matters: a later reclassification must not silently rewrite the features used for an earlier forecast.

Product age and store context belong beside physical attributes. A fashion retail study models sales as a function of age, attributes, selling price, temporal features and store features. [S03] Age should be measured from the launch that customers actually experienced in each channel, not from a design approval or factory completion date. Launch timing then serves two purposes: it aligns curves by elapsed selling time and preserves calendar conditions that may make two otherwise similar launches incomparable.

Price requires an effective-offer definition. List price, transaction price, promotional discount, delivery charge and institutional contract terms can describe different economics. The analogue record should use the customer-facing price basis that matches the target decision, with taxes and channel fees handled consistently. A product that sits in the same nominal band as the target may still be a poor comparator if one offer includes delivery or credit and the other does not.

Channel is an attribute of the launch, not a column to be standardised away. The relevant fields include who owns inventory, who chooses the assortment, how the customer discovers the product, whether returns are accepted and how quickly the offer can be replenished. These fields explain why the same garment can produce different launch curves. They also stop an online visual match from automatically becoming evidence about an institutional uniform contract.

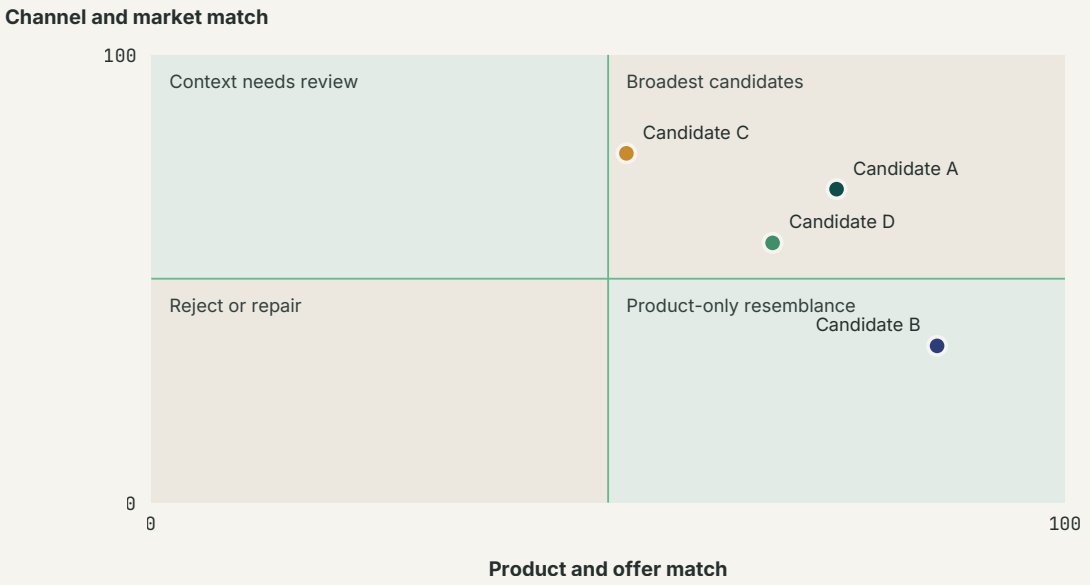
A similarity score can summarise this taxonomy, but it should remain a screening device until launch-replay evidence supports a more formal model. Weights reflect the transfer hypothesis. Category and occasion may matter more for one decision, while channel and price dominate another. The planner should approve the weights before seeing which analogue produces the most attractive forecast. Sensitivity runs then show whether a small weight change alters the selected reference class or the resulting range.

> **WORKED EXAMPLE: A WEIGHTED ANALOGY IS A SCREEN, NOT A PROBABILITY**
>
> An illustrative target is compared with one candidate using weights of 25 for category, 15 for occasion, 15 for fabric, 15 for price, 15 for channel, 10 for geography and 5 for launch timing. The candidate's respective match scores are 100, 80, 60, 80, 40, 60 and 100, all on an illustrative scale from 0 to 100. The weighted score is (25 × 100 + 15 × 80 + 15 × 60 + 15 × 80 + 15 × 40 + 10 × 60 + 5 × 100) ÷ 100 = 75. This value is only a comparability score. It is not a probability of purchase, an expected sales level or evidence that the candidate explains 75 percent of target demand. A reviewer should inspect the weak channel match before admitting the candidate, and should repeat the calculation under alternative approved weights.

Exhibit 11, p. 52 places several illustrative candidates on separate attribute and context axes. The separation is deliberate. Strong physical similarity should not compensate invisibly for weak channel or market comparability, and a single total score should never conceal the fields that drove admission.

EXHIBIT 11 / ILLUSTRATIVE MATCH SCORE FROM 0 TO 100

An analogue scorecard exposes weak comparisons before modelling



ILLUSTRATIVE [S02] [S03] Illustrative candidates and scores demonstrate screening only. Horizontal positions use the worked weights for category, occasion, fabric and price. Vertical positions combine channel, geography and timing with reviewer-approved illustrative weights. Scores are not probabilities or observed demand.

Missingness should reduce confidence in the comparison rather than receive a free middle score. If a candidate lacks price and channel evidence, its position should carry a coverage flag and its eligibility should remain provisional. Reviewers can accept it for a wide scenario pool, request more evidence, or exclude it from the central reference class. Those choices belong in the register. A score that hides missing fields creates precision without improving the analogy.

3.3

Visual similarity can transfer signals with explicit caveats

Images add information that a structured taxonomy may miss, particularly texture, styling and the overall visual impression of a garment. Recent cold-start research matches new products to clusters of similar historical items and transfers corrected demand signals from those items. [S05] This makes visual similarity a plausible candidate for finding analogues, not a licence to import their sales without inspection.

The practical sequence starts with the structured screen. Category, use occasion, price basis and channel eligibility narrow the pool before image retrieval begins. Visual neighbours then help rank candidates within that admissible set. This order reduces the chance that an attractive photograph connects products whose buying occasions or distribution mechanics are fundamentally different. It also leaves a human-readable explanation when an image model returns a surprising neighbour.

Every image needs an offer version. The picture used for matching should be the picture associated with the product as customers saw it, together with its date, channel and product identifier. A later campaign image can change colour, styling emphasis or apparent quality. If that later image enters a historical comparison, the model may use information that was absent at the original decision. The same rule applies to cropped marketplace images, technical drawings and digital samples: their roles should be labelled rather than treated as interchangeable.

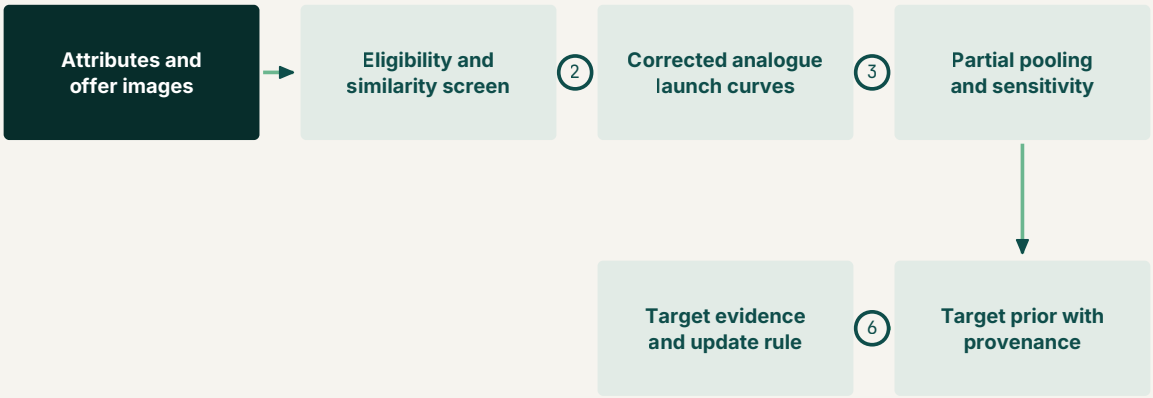
Visual resemblance does not establish identical willingness to pay, size fit or channel response. The product-age study explicitly cautions that similar visual attributes alone do not imply similar sales, and instead combines attributes with price, age, temporal and store features. [S03] Fit requires measurements, grading rules and return reasons where available. Price response requires the effective offer. Channel response requires the selling context. None should be inferred from proximity in an image embedding.

The transferred target also needs its availability history. A visual cluster built from stockout-capped sales can reproduce supply constraints as if they were customer preferences. The visual-similarity study addresses this by correcting censored sales before transferring demand, but that architecture still depends on a relevant historical catalogue and defensible correction inputs. [S05] A new firm without such a catalogue cannot claim the same evidential basis merely by adopting the architecture.

Exhibit 12, p. 54 shows the proposed transfer chain and its review points. The chain joins structured attributes and images to corrected launch curves, partial pooling and an updating rule. It is a method design, not evidence that these stages have produced a forecast for GFT or for a Cameroon apparel venture.

EXHIBIT 12 / EVIDENCE AND DECISION SEQUENCE

Attributes become a prior through an inspectable transfer chain



2. admit with reasons
3. retain dispersion
6. version or restore

DERIVED [S02] [S05] [S28] Derived GFT method synthesis: structured and visual attributes identify eligible analogues; availability-corrected launch curves supply outcomes; partial pooling forms a target prior; versioned updating adds target evidence. No implemented performance is claimed.

Review should test cluster composition as well as forecast error. A reviewer examines representative items, extreme outcomes and rejected neighbours, then records whether the grouping reflects the intended occasion and offer. Cluster stability can be checked under reasonable changes to imagery and candidate pools. If a small processing change replaces most neighbours, the analogue set is not a dependable foundation for a narrow prior. The correct response is broader uncertainty and more target evidence, not another unexplained model layer.

3.4

Analogue markets require economic and channel comparability

A similar garment sold in another retail system is a hypothesis about transfer, not an established comparator. Multichannel fashion research evaluates physical and non-store channels separately and recommends testing regional and channel-specific covariates rather than assuming a universal relationship. [S04] The analogue market screen should therefore begin with the mechanics through which a customer encounters, evaluates and pays for the offer.

Channel comparability covers assortment control, inventory ownership, customer discovery, payment, delivery, returns and replenishment. A distributor's initial order is not equivalent to repeated consumer sell-through, even if both are recorded as sales in their respective systems. An online launch with immediate delivery is not directly comparable with a made-to-order offer that has a long waiting period. These differ-

ences can alter both the level and timing of recorded demand. Where they cannot be reconciled, the launch should move to a scenario pool or leave the central reference class.

Geography adds more than a location label. Climate, calendar, sizing, language, retail concentration and the availability of substitutes can all affect the buying condition. The analogue record should identify which of these are observed and which remain assumptions. It should not convert a broad national similarity statement into an exact adjustment multiplier. A documented mismatch can justify exclusion or a wider prior; it does not justify an arbitrary percentage correction.

Macroeconomic indicators are useful for screening scale and context at this stage, not for predicting a style's weekly sales. The World Bank defines total population using a de facto count of residents and reports midyear estimates. [S44] That provides a clear denominator for a market-scale discussion. It does not identify the subset of residents who share the target product's occasion, price access or channel reach.

Income requires the same restraint. The Bank's GNI per capita series divides an economy-wide income measure by population and expresses it in constant prices using a stated reference year. [S45] An average income measure can help rule out obviously different market contexts and support slow-moving scenarios. It cannot establish clothing expenditure, willingness to pay for a particular garment, or the income distribution among reachable customers.

Urbanisation can describe locational context, but the underlying definitions are national. The World Bank explicitly advises caution in cross-country comparisons because countries classify urban and rural populations differently. [S46] The market register should preserve those definitions and the observation period. A similar urban share does not prove similar retail access, and a national urban measure cannot substitute for evidence about the actual launch cities.

Exhibit 13, p. 56 separates what each layer can contribute from what it cannot establish. The table is a control on scope. It keeps macro context above the product forecast and prevents channel resemblance from being inferred from a product photograph.

EXHIBIT 13 / PERMITTED AND PROHIBITED INFERENCE

Each comparability layer has a limited transfer role

Evidence layer	Can inform	Cannot establish alone
Product and occasion	Candidate eligibility	Identical willingness to pay
Effective offer	Price and terms comparability	Customer response at another date
Channel mechanics	Launch-curve transferability	Consumer purchases from shipments
Population context	Broad market scale	Reachable apparel buyers
Average income	Slow market scenarios	Style-level purchasing power
Urban definition	Locational context	Equivalent retail access

DERIVED [S04] [S44] [S45] [S46] Derived qualitative synthesis of opened channel research and indicator definitions. No cross-country numerical similarity or demand elasticity is estimated.

Market and product matching should remain separate until both pass review. A strong product match in a weak market may belong in a downside or upside scenario, depending on the documented difference, but not automatically in the central prior. A strong market match with a weak product match is similarly limited. This two-stage discipline makes it possible to explain whether forecast uncertainty comes from the garment, the selling system or the wider market.

3.5

Partial pooling shares strength without forcing identical behaviour

Once several eligible launch groups exist, the analyst faces two unattractive extremes. Full pooling combines every case and suppresses real differences among products, channels and markets. No pooling treats each launch as unrelated and can produce unstable local estimates from very little evidence. Hierarchical modelling offers a middle route: local parameters are informed by a shared distribution while remaining allowed to differ.

The supporting demand study makes this distinction explicit. It estimates local and global parameters together and shares information across heterogeneous sites, contrasting hierarchical estimates with fully pooled and unpooled alternatives. [S28] Its empirical application is copper demand across regions and sectors, not apparel launches. The transferable idea is the statistical structure, not an elasticity, uncertainty reduction or forecast result that can be quoted for garments.

For an analogue prior, the shared distribution might describe launch curves within a product family, channel or market group. Each analogue retains a local departure from that shared pattern. A new target without sales then draws its initial expectation from the relevant shared distribution and the variation observed among comparable launches. It should not simply inherit the arithmetic mean of the most convenient neighbours.

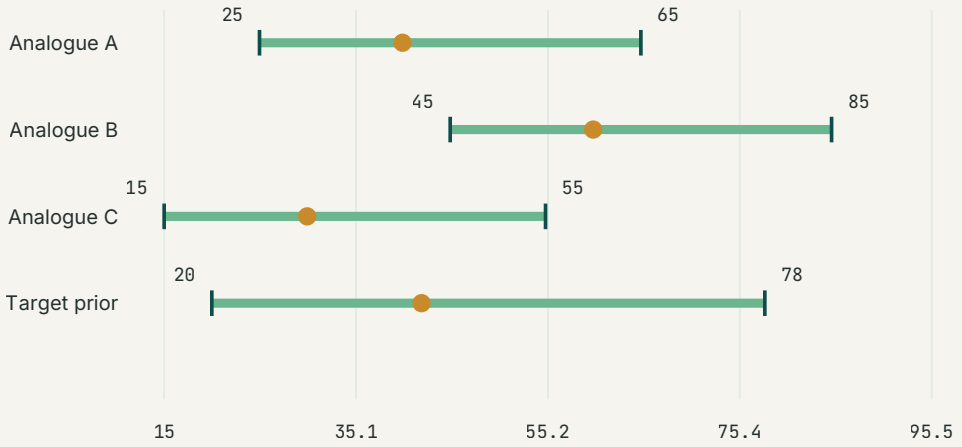
Partial pooling also changes what must be reported. The model version should name the grouping variables, the parameters allowed to vary, the prior choices and the data included at each level. Reviewers need to see which groups contain many independent launches and which rely on one or two records. They also need sensitivity results for reasonable alternative groupings and priors. A pooled answer that moves materially under those alternatives is not ready to support a narrow production commitment.

Shrinkage should be an outcome of the evidence and model assumptions, not a fixed rule chosen to make forecasts look stable. Where local launches are genuinely similar and well observed, stronger common structure may be defensible. Where they are heterogeneous, weakly matched or censored, local and between-group uncertainty should remain visible. The amount of pooling is therefore part of model validation rather than a hidden consequence of concatenating every row.

Exhibit 14, p. 57 illustrates why the target range should retain both shared and local variation. The analogue groups have different centres and spreads, while the target prior remains wider than a simple average of their medians would suggest. All values are indexed illustrations. They are not garment sales, fitted posteriors or calibrated intervals.

EXHIBIT 14 / ILLUSTRATIVE INDEXED FIRST-WINDOW DEMAND

Partial pooling preserves both shared and local uncertainty



ILLUSTRATIVE [S28] Illustrative medians and bounds assumed to demonstrate partial pooling. Analogue A low, mid and high are 25, 40 and 65; B are 45, 60 and 85; C are 15, 30 and 55. The target values 20, 42 and 78 are constructed to retain between-group and transfer uncertainty. No statistical coverage is claimed.

The target product must stay outside the fitting process used to evaluate its cold start. Analogue groups, normalisation and pooling choices are fitted on reference launches only. Only then can a held-out launch test whether the method transferred to an unseen case. This is a necessary check on the analogy, not an optional technical refinement. Detailed whole-launch replay belongs to the later validation chapter; the requirement here is to preserve the data and versions that make such a replay possible.

GFT NOTE

GFT would approve a partially pooled prior only with a companion sensitivity pack. The pack shows the eligible launches, independent support by group, pooled and unpooled comparisons, alternative grouping choices, and the effect of removing each influential analogue. If the reference class is too small or weakly comparable for a stable readout, the approved result is a wide prior and a plan to collect better launch evidence. Pooling must never become a way to manufacture confidence from a thin corpus.

3.6

An analogue prior retains dispersion and disappointing cases

The prior should preserve the launch distribution, not collapse it into a central estimate with an arbitrary band attached. Each admitted curve contributes its level, timing, censoring and outcome status. Slow sellers and withdrawals remain in the readout even when a planner considers them unlikely for the target. Their presence is what allows the outside view to correct the natural attraction of a persuasive launch story.

Three uncertainty sources should remain distinct. Observation uncertainty concerns errors, missing events and demand censoring within an analogue. Between-launch dispersion concerns the variation among eligible outcomes. Transfer uncertainty concerns the possibility that the target will not behave like the reference class even after matching. Combining them into one unexplained interval hides the lever that later evidence can realistically reduce. Early target purchases may reduce the first two under suitable conditions, while a changed channel or market can increase the third.

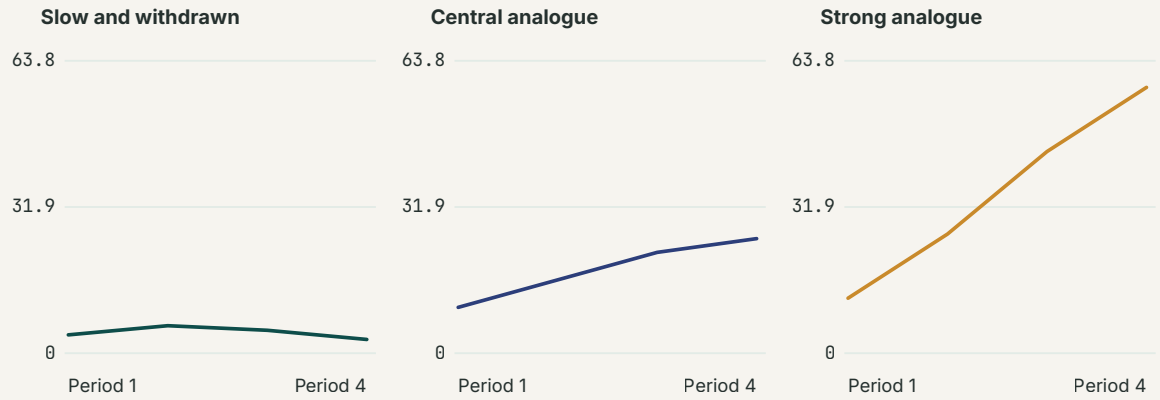
Launch curves should be aligned by elapsed selling time before their distribution is summarised. The alignment uses the date on which each offer became available in the relevant channel and keeps calendar conditions as separate fields. Availability-censored intervals remain marked, and a withdrawn product is not given an imaginary continuation. Where a correction is not defensible, the report can show the observed curve and a sensitivity case rather than select a single reconstructed history.

Weights should change influence without quietly deleting tails. A weakly comparable strong performer may receive little weight, but its outcome can still inform the upper scenario boundary. A close but disappointing analogue should not be removed because it pulls the median down. The prior register therefore stores both the weighted empirical readout and the unweighted sensitivity readout, together with the independent launch support behind each.

Exhibit 15, p. 59 illustrates three launch paths that should remain visible together. The central path does not replace the slow or strong case. These indexed curves are deliberately illustrative and do not represent an adoption law, an observed apparel series or a recommended production ramp.

EXHIBIT 15 / ILLUSTRATIVE INDEXED DEMAND PER LAUNCH PERIOD

Disappointing launches remain part of the prior



ILLUSTRATIVE [S01] [S30] Illustrative launch paths constructed to show retained dispersion. Values are assumed indices, not observed apparel demand. The consumer-durable adoption source supports only the general need to distinguish curve shape from product-specific evidence.

A range formed this way is an evidence summary whose calibration still needs testing. Proper scoring rules provide a principled way to evaluate probabilistic forecasts and encourage honest assessments; they do not certify an analogue distribution before held-out outcomes exist. [S14] The launch review should eventually compare stated quantiles with realised outcomes across comparable cold starts. Until that evidence exists, the report calls the range a prior, states its construction and avoids presenting its bounds as a measured coverage guarantee.

Scenarios sit outside that statistical statement. Forecasting practice distinguishes probabilistic forecasts from conditional views formed under specified assumptions. [S08] A price change, a new distributor or a supply interruption can be explored as a named scenario with an owner and rationale. It should not be smuggled into the prior by widening every quantile until the preferred production plan looks safe.

3.7

Transfer weight should decay as target evidence accumulates

The analogue prior is the beginning of an evidence sequence. As the target earns its own launch history, the forecast should rely less on borrowed outcomes and more on observed purchases, availability and repeat behaviour. Global models supply another source of shared evidence by fitting one function across series rather than an independent function for each product. [S22] The update record should therefore distinguish analogue weight, shared-model weight and target-history weight instead of describing the whole system as machine learning.

Decay should follow effective evidence, not elapsed time or row count. A week with well-recorded availability, stable offer conditions and clean customer identities may deserve more influence than a longer period with stockouts and changing promotions. Multiple transactions from one institutional order may also carry less independent

information than the transaction total implies. The updating rule needs fields for censoring, offer changes, channel mix and dependence before it can assign target weight defensibly.

Hierarchical updating provides a coherent candidate for this transition because shared and local parameters can be estimated together. [S28] Yet the operational rule should remain visible to planners. Each forecast version records the prior used, the new evidence admitted, the resulting range and the contribution still attributable to analogues. A reviewer can then reconstruct why the forecast moved and challenge a transfer assumption without inspecting model code.

Confidence need not increase monotonically. Early sales can reveal that the target occupies a different channel, price or size context from the analogues. A stockout can make the new series less informative than expected. A sudden departure from the reference distribution can indicate that the prior was wrong, that the target definition changed, or that the observation process failed. In each case, widening the range or restoring an earlier prior may be more defensible than forcing rapid adaptation.

The update rule should also specify what happens when target and shared evidence disagree. One option is to retain both forecasts side by side until availability and definitions pass review. Another is to publish a combined range whose additional transfer uncertainty reflects the disagreement. What the team should not do is select whichever output best supports the existing production commitment. Combination and reversion rules are fixed before the launch readout and changed only through a documented review.

A small team can operate this discipline through a compact version register. It contains the forecast identifier, information cut-off, analogue corpus version, taxonomy version, availability correction, pooling specification, target evidence summary and approval status. It also records the reason for every material revision. These artefacts make the prior reproducible and allow later backtests to deny the model information that was unavailable at the original decision.

The final control is a stopping rule for weak evidence. If independent analogue launches are too few, their attributes are poorly observed, or their channels cannot be reconciled with the target, the prior remains wide. The next expenditure goes to better launch records, calibrated research or a limited pilot rather than a more elaborate transfer model. This is a successful application of the method: it identifies what the current evidence cannot support.

Before the first production commitment, the planning owner should receive the reference-class register, attribute scorecards, visual-review exceptions, market comparison, pooled and unpooled sensitivity cases, the full prior distribution and the updating rule. Commercial and production reviewers approve different parts of that

pack, but neither can silently rewrite the other's evidence. The launch then has an outside view that can be inspected, challenged and gradually replaced by owned history.

Borrow the distribution, not the success story; retire the analogy only when the launch has earned its own evidence.

CHAPTER SOURCE NOTES

S08 International Journal of Forecasting, Elsevier; full text through arXiv. 2022.

S14 Journal of the American Statistical Association, American Statistical Association. 2007-03.

S01 InTech, International Journal of Engineering Business Management. 2013.

S05 Springer Nature, International Journal of Data Science and Analytics. 2026-01-08.

S30 Management Science, INFORMS; abstract through EconPapers. 1969.

S02 arXiv. 2019-06-27.

S03 arXiv. 2020-07-10.

S28 Journal of Industrial Ecology, Wiley; abstract through RePEc IDEAS. 2023-02.

S04 MDPI, Processes. 2021-09-03.

S44 World Bank, World Development Indicators. 2026-10-30.

S45 World Bank, World Development Indicators. 2026-10-30.

S46 World Bank, World Development Indicators. 2026-10-30.

S22 arXiv. 2021-03-26.

04

PART I: DEFINE THE DECISION AND THE PRIOR / MARKET COVARIATES

External data can bound a market but cannot nominate a winner

Trade, demographic and digital signals are most credible when each is assigned a limited and testable forecasting role.

CHAPTER 04
DECISION TEST

Trade, demographic and digital signals are most credible when each is assigned a limited and testable forecasting role.

External statistics cannot tell a greenfield apparel manufacturer which shirt, dress or uniform variant will clear its first production lot. Customs tables, national accounts, agricultural balances, search indices and weather records measure aggregates at boundaries far removed from a customer choosing a specific garment in a specific channel at a specific price. A disciplined cold-start forecasting system therefore assigns every external series a bounded structural task before any model is fitted. Reported trade flows help bound formal category inflows and unit-value distributions, but they do not measure consumer sell-through. [S41] [S43] Demographic and income series scale multi-year market scenarios and screen candidate analogue markets, but their annual cadence cannot drive weekly stock keeping unit (SKU) replenishment. [S44] [S45] [S46] Agricultural and industrial cotton statistics describe upstream raw-material availability and policy ambition, not downstream garment demand. [R06] [S47] Higher-frequency digital and meteorological series enter only as challenger covariates that must beat simple baselines in strictly chronological, ex-ante launch replay. [S09] [S34] [S35] [S0401]

This separation of duties protects a new venture from a frequent cold-start error outside the factory floor: mistaking a large macro or trade number for product-level commercial proof. Fashion demand combines demand uncertainty, short product life cycles and seasonal shifts even when a firm holds years of historical transactions. [S01] [S0402] When a firm holds no sales history of its own, judgemental and prior-building methods are required, and every external proxy must carry an explicit boundary definition, an as-of timestamp and a falsifiable test of predictive value. [S08] [S37] In Green Future Tech Limited (GFT) programme work for Cameroon's textile sector, prepared on 31 January 2024 with source refreshes flagged in 2026, external market indicators establish why local industrial capacity is strategically relevant while leaving SKU-level demand commitments to calibrated priors, controlled pilots and audited demand capture. [R06] [R07]

4.1

Reported trade flows describe supply, not consumer sell-through

A customs import table records border crossings of classified merchandise, not retail demand. When a planning team extracts apparel import series from the United Nations Comtrade database, it observes data sourced directly from national authorities and submitted to the United Nations Statistics Division under international merchandise trade conventions. [S41] Those records capture formal consignments declared at customs ports. They do not record garments produced and sold by domestic tailors, stock held in wholesale warehouses between seasons, unrecorded cross-border trade, or the date and price at which a final consumer buys an item. Treating an annual or monthly customs inflow as a direct proxy for retail sell-through confuses wholesale supply entering a territory with consumer demand clearing at the point of sale. [S06] [S41]

4.2 CUSTOMS CLASSIFICATIONS AGGREGATE HETEROGENEOUS GARMENTS AND CHANNELS

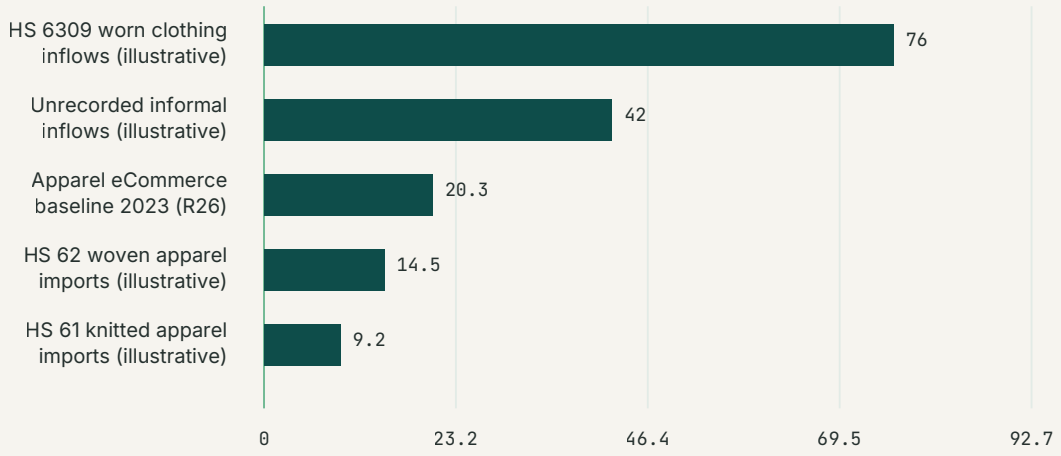
International trade platforms organise apparel flows primarily under Harmonized System (HS) Chapter 61 (articles of apparel and clothing accessories, knitted or crocheted) and Chapter 62 (articles of apparel and clothing accessories, not knitted or crocheted), alongside Chapter 63 for other made-up textile articles and worn clothing. The International Trade Centre (ITC) Trade Map system provides both values and quantities for the trade flows of goods across these tariff headings. [S43] Even at the six-digit subheading level, a single customs line groups garments that differ in silhouette, fabric weight, stitch density, branding, sizing, retail channel and target occasion. New-item forecasting research demonstrates that clothing and footwear demand depends on fine-grained merchandising attributes, selling price, product age and store context. [S02] [S03] Because a customs declaration strips away those merchandising attributes, a six-digit HS flow can bound the volume of a broad category entering formal distribution, but it cannot identify which design attributes or price points will win inside that category. [S02] [S43]

Market context in Cameroon illustrates why channel and product boundaries must remain explicit before any trade series touches a forecast prior. Inherited sector evidence records that Cameroon's domestic textile player CICAM saw its domestic market share fall from about 85 percent three decades ago to about 5 percent in the baseline period, while informal employment accounts for about 90 percent of total employment across the economy. [R11] [R16] At the consumer level, inherited market estimates put Cameroon clothing and footwear expenditure at about US$94.70 a person in 2024, corresponding to about 8.0 pieces a person, while domestic apparel eCommerce revenue was estimated at US$20.32 million in 2023 and projected to reach US$33.61 million by 2027 at an annual growth rate of 13.41 percent. [R26] Each figure describes a different boundary: total personal expenditure, physical piece intake, formal online channel revenue, or domestic mill share. None is identical to customs-reported imports under HS Chapters 61 and 62, and none measures the sell-through velocity of a newly launched local garment line. [R11] [R26] [S41]

Exhibit 16, p. 65 contrasts these distinct channel and customs buckets on a common monetary scale, joining the sourced 2023 Cameroon apparel eCommerce baseline with illustrative customs and informal inflow slices to show why an analyst must decompose category supply before setting a market bound.

EXHIBIT 16 / MILLION US$ BY REPORTED OR ILLUSTRATIVE CHANNEL

Reported apparel inflows need product and provenance detail



ILLUSTRATIVE [S41] [S43] [R26] Illustrative diagnostic comparison anchored to the sourced 2023 Cameroon apparel eCommerce revenue baseline of US$20.32 million (R26) and structured using UN Comtrade (S41) and ITC Trade Map (S43) tariff boundaries. The slices for HS 6309 worn clothing (76.0), unrecorded informal inflows (42.0), HS 62 woven apparel (14.5) and HS 61 knitted apparel (9.2) are illustrative values constructed to show boundary separation; they are not observed GFT customs extracts or consumer sell-through totals.

Exhibit 16, p. 65 establishes a foundational rule for Part I of the workflow: the largest volume entering a clothing market may sit in worn clothing or informal distribution rather than formal new-apparel customs lines, and even formal HS 61 and HS 62 lines represent landed wholesale shipments rather than consumer purchases. [R16] [S41] [S43] The exhibit cannot establish addressable demand for a specific Cameroonian factory offer; instead, it forces the commercial team to specify which customs headings, price tiers and distribution channels its proposed garment contests.

4.3 POLICY AND REGIONAL TRADE RULES ALTER SUPPLY CEILINGS WITHOUT GUARANTEEING SALES

Trade series also reflect tariff regimes and preferential access rules. In the inherited register, Cameroon's eligibility under the United States African Growth and Opportunity Act (AGOA) was terminated effective 1 January 2020 and remained listed as ineligible for calendar year 2025. [R18] Within Africa, textiles and clothing remained among unresolved tariff categories under the African Continental Free Trade Area (AfCFTA), where 92.3 percent of lines had been agreed excluding vehicles, textiles and clothing. [R23] Continental market-research sources in the register estimate the Africa textile market at about US$37.3 billion in 2025, projected to reach US$49.41 billion by 2030 at 4.73 percent a year in one series, and projected to reach US$72.7 billion by 2035 at 6.9 percent a year with fashion and apparel representing 48.1 percent of the 2024 market in a second vendor series. [R24] [R25]

A cold-start team treats these vendor projections and trade-policy rules as scenario boundaries rather than model inputs. [S08] [S0403] Vendor market-size reports carry grade-C status and use incompatible definitions, so GFT retains both series without averaging them. [R24] [R25] Tariff-line resolution or preferential access changes the competitive ceiling of an export channel, belonging in conditional scenarios rather than in the statistical prior for domestic weekly sell-through. [R18] [R23] [S08]

4.4

Mirror asymmetries need a reconciliation rule

Even when trade data are used strictly to bound formal inflows, a single customs query rarely yields an unambiguous number. In UN Comtrade, bilateral trade asymmetries occur when reported exports from country A to country B do not match reported imports to country B from country A. [S42] Where national customs reporting experiences publication lags, incomplete capture or reclassification, partner-reported exports (mirror data) diverge from reporter imports. [S41] [S42] Overwriting reporter imports with mirror exports whenever a new extraction runs silently alters historical backtests. [S37] [S42]

4.5 STRUCTURAL CAUSES OF BILATERAL DIVERGENCE REQUIRE EXPLICIT ADJUSTMENTS

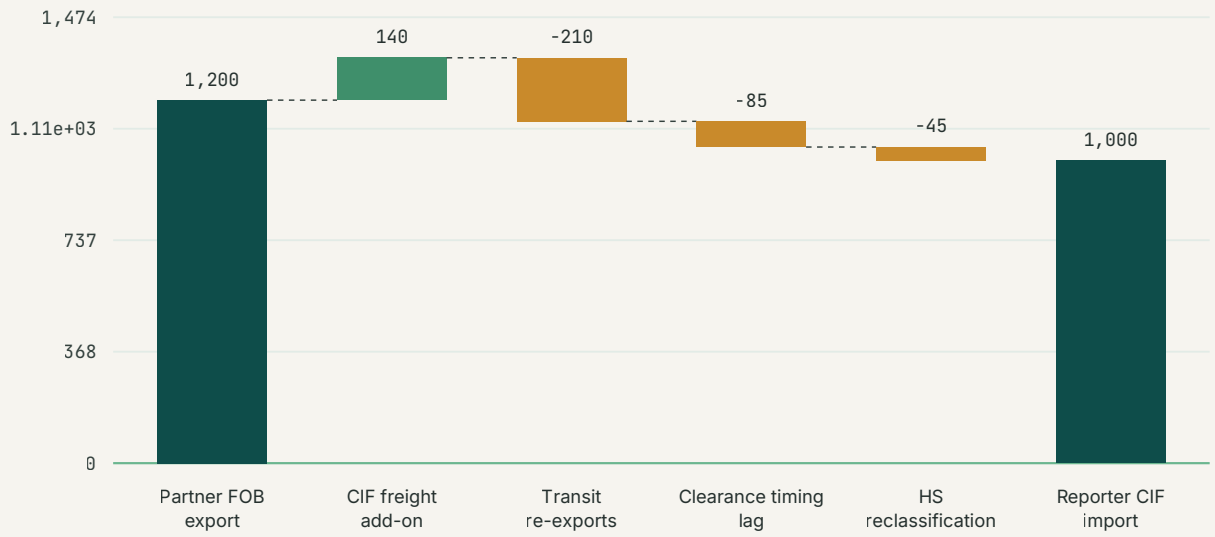
Before reconciling reporter and partner records, the engineer separates four mechanical causes of bilateral asymmetry: [S42] [S43]

- **Valuation basis (CIF versus FOB):** Importing countries record cost, insurance and freight (CIF) values, whereas exporting partners record free on board (FOB) values, creating a freight and insurance wedge. [S42]
- **Transit and re-export routing:** Consignments shipped through intermediate ports, or landed at Douala for transit to landlocked neighbours, may be attributed differently by origin and consignment rules. [S42]
- **Clearance timing lags:** Maritime shipments departing in November or December frequently clear destination customs in January or February, shifting flows across annual boundaries. [S42]
- **Tariff reclassification:** An exporter may declare a six-digit HS 61 or HS 62 subheading while the importing authority records mixed lots under another subheading or HS Chapter 63. [S42] [S43]

Exhibit 17, p. 67 illustrates how these four mechanisms bridge a partner's FOB export total and a reporter's CIF import total.

EXHIBIT 17 / ILLUSTRATIVE THOUSAND US$ IN BILATERAL HS 62 TRADE

Mirror reconciliation bridges partner exports and reporter imports



ILLUSTRATIVE [S41] [S42] [S43] Illustrative reconciliation walk showing how GFT decomposes bilateral trade asymmetries defined by UN Comtrade (S41, S42) and ITC Trade Map (S43). Starting from an illustrative partner FOB export record of 1,200 thousand US$, adding CIF freight and insurance (+140) and deducting third-country transit routing (-210), year-end customs clearance lag (-85) and tariff-line reclassification (-45) reconciles to the reporter CIF import record of 1,000 thousand US$. Values are illustrative accounting steps, not a measured Cameroon customs audit.

Exhibit 17, p. 67 shows why taking a simple average of reporter and partner numbers, or picking whichever number is larger, corrupts the market bound. Because CIF freight works in the opposite direction from transit re-exports and clearance lags, net bilateral discrepancy can be positive or negative depending on the corridor and quarter. [S42] What **Exhibit 17**, p. 67 cannot establish on its own is the exact freight ratio or transit share for a corridor without shipping-line and customs-manifest evidence; where those inputs are absent, the discrepancy must remain visible as an uncertainty interval around the inflow bound rather than disappear into a single point estimate. [S08] [S42]

4.6 EXTRACTION METADATA PREVENTS SILENT BACKTEST REWRITING

National authorities submit late declarations and revisions to UN Comtrade long after the calendar year closes. [S41] If an analyst pulls trade series in 2026 to backtest a 2024 launch origin without recording when each figure was published, the backtest commits temporal data leakage by using revised trade totals unavailable at the decision date. [S37] [S41] Every trade extraction in the GFT pipeline therefore locks seven mandatory provenance fields alongside the numeric series:

1. **Reporter jurisdiction and code:** The declaring customs territory and its numerical identifier. [S41]
2. **Partner jurisdiction and role:** Country of origin, country of consignment or partner export reporter, preserved as separate columns. [S42]
3. **Trade flow and regime:** Gross imports, net imports, domestic exports or re-exports. [S41] [S42]

4. **HS nomenclature revision:** The Harmonized System vintage (such as HS 2017 or HS 2022) used to define the six-digit code. [S41] [S43]
5. **Valuation and physical units:** Currency basis (CIF or FOB US dollars, or CFA francs at the euro peg of 655.957 XAF per euro) and supplementary quantity units (kilograms, pieces, pairs or dozens). [R32] [S43]
6. **Reference period versus publication date:** The calendar period of the customs event alongside the release date. [S37] [S41]
7. **Snapshot extraction timestamp:** The exact as-of date on which GFT froze the table for a forecast version. [S37] [S40]

4.7

Values and quantities answer different diagnostic questions

Trade databases following the ITC Trade Map convention report both monetary trade values and physical quantities for goods flows. [S43] In apparel diagnostics, value and quantity answer distinct operational questions and fail in different ways. Monetary import value (in US dollars or XAF) reflects total border expenditure at wholesale landed cost, which responds to inflation, exchange-rate movements, valuation rules and shifts between budget and premium garments. [R32] [S43] Physical quantity (in net kilograms or supplementary pieces) reflects material volume crossing the border, connecting to spinning, weaving, dyeing and sewing capacity measured in tonnes or pieces a year. [R10] [S43]

4.8 UNIT-VALUE RATIOS MIX PRICE, WEIGHT AND PRODUCT COMPOSITION

Dividing reported trade value by reported quantity yields an apparent unit value, such as US dollars per kilogram or per piece. [S43] Commercial teams frequently treat this ratio as a proxy for local selling price or quality tier. That inference is unsafe without independent validation. Customs quantity reporting is often incomplete across supplementary units, grouping multipacks or mixed bale weights under one line. [S43] Even when weight and piece counts are complete, a change in US dollars per kilogram between two periods can reflect a shift in fabric weight (such as lightweight cotton t-shirts versus heavy twill workwear trousers) rather than a change in garment price. Furthermore, a customs CIF unit value measures wholesale landed cost before importer margins, domestic transport, taxes and retail mark-ups. [S03] [S43]

WORKED EXAMPLE: DECOMPOSING A CUSTOMS UNIT-VALUE SHIFT BEFORE TOUCHING PRICE PRIORS

Consider an illustrative customs diagnostic for a six-digit woven trousers heading under HS Chapter 62. In Year 1, a reporter logs US$1,800,000 in CIF import value and 120,000 kilograms in net weight, giving an aggregate unit value of 1,800,000 ÷ 120,000 = US$15.00 per kilogram. Suppose the consignment mix in Year 1 consists of 90,000 kilograms of basic work trousers landed at US$12.00 per kilogram (US$1,080,000) and 30,000 kilograms of tailored cotton chinos landed at US$24.00 per kilogram (US$720,000). With an illustrative garment weight of 0.50 kilograms for work trousers and 0.40 kilograms for chinos, the physical flow equals (90,000 ÷ 0.50) + (30,000 ÷ 0.40) = 180,000 + 75,000 = 255,000 pieces, and the landed cost per piece is US$6.00 for work trousers and US$9.60 for chinos.

In Year 2, reported value rises to US$1,980,000 while net weight remains 120,000 kilograms, raising the apparent aggregate unit value by 10 percent to 1,980,000 ÷ 120,000 = US$16.50 per kilogram. A naive covariate rule would feed a 10 percent market price increase into the launch model. However, an illustrative sub-shipment audit shows that neither garment's landed price changed: basic work trousers fell to 75,000 kilograms at US$12.00 per kilogram (US$900,000) while tailored chinos rose to 45,000 kilograms at US$24.00 per kilogram (US$1,080,000). Total pieces imported rose from 255,000 to (75,000 ÷ 0.50) + (45,000 ÷ 0.40) = 150,000 + 112,500 = 262,500 pieces, while every SKU price remained flat. The 10 percent rise in aggregate customs unit value was entirely a composition and weight-per-piece effect.

This arithmetic demonstrates why GFT rejects single unit-value rankings from trade tables as direct proof of consumer willingness to pay or retail price elasticity. [S02] [S03] [S43] When an analyst uses Trade Map values and quantities to bound a market, weight in tonnes connects to upstream mill capacity (such as the 12,000 tonnes a year announced for Camtext SA), piece counts connect to per-capita intake benchmarks (such as the 8.0 pieces a person reported for Cameroon in 2024), and monetary values connect to expenditure ceilings (such as the US$94.70 per-person clothing and footwear spend in 2024), each with an explicit conversion range. [R10] [R26] [S43]

4.9

Population, income and urbanisation scale priors slowly

Macroeconomic and demographic indicators sit at the top of the external evidence hierarchy, above customs categories and far above individual SKUs. In Chapter 3, population, income and urbanisation helped screen whether a foreign city or country was an eligible analogue market. [S44] [S45] [S46] Here, their role as time-series covariates is defined. Because national demographic and income accounts update annually, incorporate census revisions and describe whole-population averages, they can scale multi-year market priors and capacity scenarios, but they cannot explain week-to-week demand variation across styles, colours and sizes. [S01] [S44] [S45] [S46] [S0402]

4.10 INDICATOR DEFINITIONS RESTRICT WHAT ANNUAL MACRO SERIES CAN REPRESENT

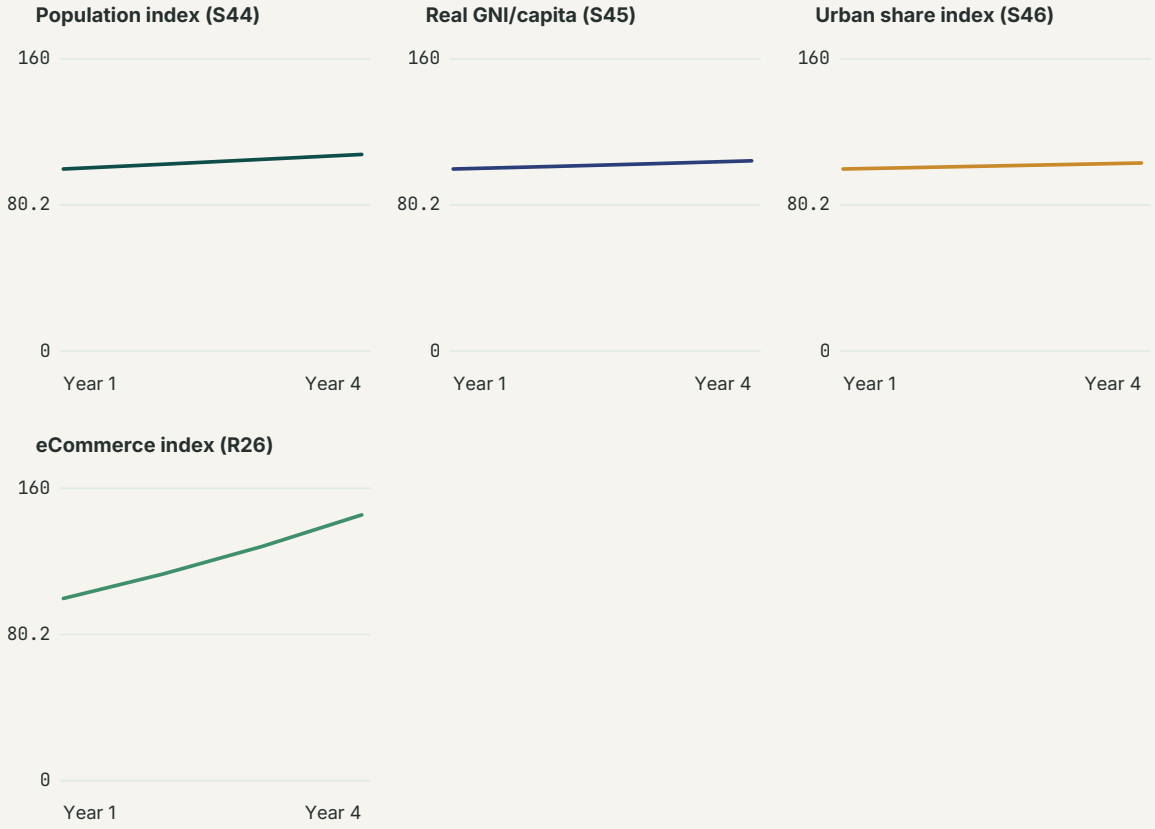
Each of the three World Development Indicators (WDI) carries a statistical definition that limits its forecasting role:

- **Total population (SP.POP.TOTL):** The World Bank defines total population on the de facto definition, counting all residents regardless of legal status or citizenship as midyear estimates. [S44] While SP.POP.TOTL supplies a resident-count denominator for converting per-person intake estimates (such as 8.0 pieces or US$94.70 per person in R26) into a national volume envelope, it treats infants, rural subsistence households and urban salaried buyers as equal units. [R26] [S44]
- **GNI per capita in constant 2015 US dollars (NY.GNP.PCAP.KD):** Gross national income measures total income earned by residents within an economic territory during an accounting period, divided by midyear population and adjusted to constant 2015 prices. [S45] It does not measure household disposable income within a retail catchment or supply a clothing-specific income elasticity. [S45] In Cameroon, inherited evidence shows wide income dispersion and informality: the statutory minimum wage (SMIG) rose by 5 percent from XAF 41,875 a month in 2023 to XAF 43,969 a month (about US$76) from 23 February 2024, whereas average monthly salary across sectors was reported at about US$359 a month and informal employment stood near 90 percent. [R14] [R15] [R16] A national mean like NY.GNP.PCAP.KD cannot tell a planner whether a garment priced for formal earners will clear in a specific store. [R14] [R15] [S45]
- **Urban population share (SP.URB.TOTL.IN.ZS):** This series reports the percentage of total population living in urban areas as defined by national statistical offices. [S46] Because administrative criteria differ across countries, cross-country urban comparisons require caution, and a national urban percentage does not measure store footfall in Douala or Yaounde. [S46]

Exhibit 18, p. 71 illustrates the frequency and magnitude mismatch between these slow-moving annual macro series and faster-moving channel adoption over a four-year window.

EXHIBIT 18 / ILLUSTRATIVE INDEX (YEAR 1 = 100)

Slow market indicators belong above the SKU layer



ILLUSTRATIVE [S44] [S45] [S46] [R26] Illustrative indexed four-year comparison (Year 1 = 100) showing why annual demographic and income series defined by World Bank WDI (SP.POP.TOTL in S44, NY.GNP.PCAP.KD in S45, SP.URB.TOTL.IN.ZS in S46) change too slowly to explain short-cycle apparel variation, whereas channel-specific adoption (calibrated to the sourced 13.41 percent annual growth rate of Cameroon apparel eCommerce from 2023 to 2027 in R26: 100.0, 113.4, 128.6, 145.9) moves far faster. Macro panel trajectories (100.0 to 108.0, 104.5 and 103.3) are illustrative annual paths.

As **Exhibit 18**, p. 71 makes clear, annual resident population, real GNI per capita and urban share move by only a few index points over four years, while a specific sales channel such as Cameroon apparel eCommerce (R26, growing at 13.41 percent a year from US$20.32 million in 2023 to a projected US$33.61 million in 2027, with user penetration rising from 18.0 percent to 22.7 percent) compounds by nearly 46 index points over the same span. [R26] [S44] [S45] [S46] At the weekly SKU level, variance is larger still and is dominated by product attributes, selling price, item age, store allocation and stock availability. [S02] [S03] [S06]

4.11 INTERPOLATING ANNUAL MACRO SERIES INTO WEEKLY MODELS CREATES SPURIOUS FIT

A common machine-learning error is to linearly interpolate annual population or GNI per capita into fifty-two weekly values and feed them into a tree or neural model alongside weekly SKU sales. Because an interpolated annual series is a smooth monotonic trend within each year, a flexible learner uses it as a surrogate time index, memorising in-sample launch ramps and failing when a new style launches under different conditions. [S09] [S37] In GFT's architecture, SP.POP.TOTL, NY.GNP.PCAP.KD

and SP.URB.TOTL.IN.ZS are restricted to scaling the annual category envelope in multi-year capacity planning and serving as annual market context in hierarchical Bayesian partial pooling. [S28] [S44] [S45] [S46]

4.12

Cotton statistics describe supply conditions rather than garment demand

For an integrated African textile venture, upstream agricultural statistics are prominent in feasibility studies. The Food and Agriculture Organization of the United Nations (FAO) publishes agricultural production statistics through FAOSTAT covering crop commodities from production volumes to harvested areas up to 2024. [S47] In Cameroon, the inherited GFT evidence register records substantial upstream cotton scale: national seed-cotton production reached 330,000 tonnes in 2022, 314,455 tonnes in 2023 and about 340,000 tonnes in 2024, with a 2025 projection above 350,000 tonnes. [R01] [R02] Across the CFA-zone cotton belt, output was reported at 2.3 million tonnes in 2024/25, down 11.5 percent. [R05] Within northern Cameroon, SODECOTON supports more than 220,000 cotton producers, operates nine ginning plants and two oil mills, and injects roughly US$100 million a year into rural communities. [R03] [R04]

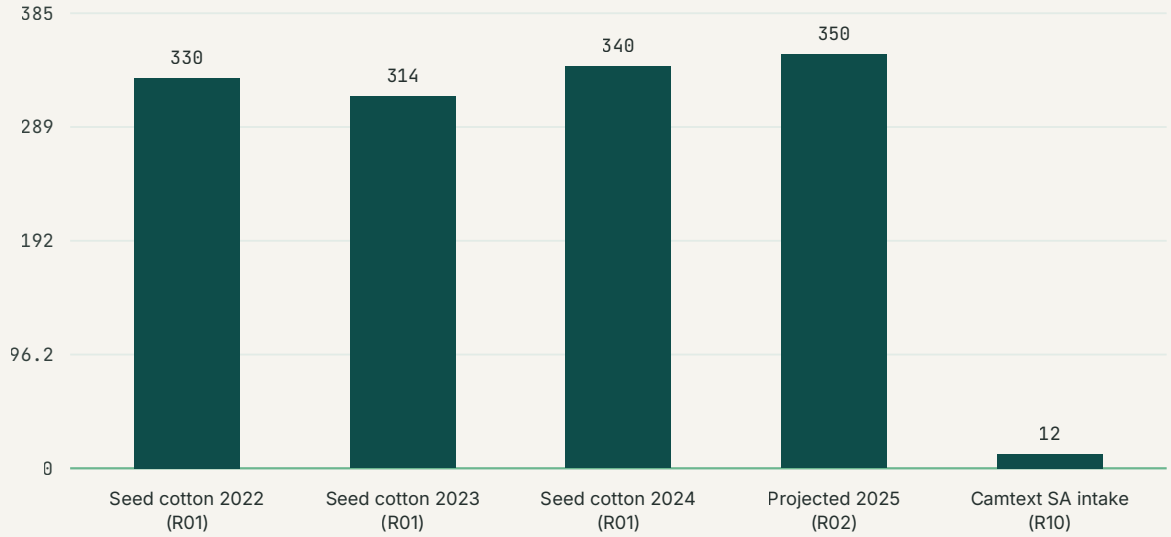
4.13 INDUSTRIAL PROCESSING TARGETS JUSTIFY A FACTORY, NOT A SALES FORECAST

Upstream abundance coexists with a severe downstream processing gap. About 95 percent of Cameroon's cotton fibre is exported raw against about 5 percent processed locally, while national industrialisation policy under SND30 sets a target of 50 percent local processing by 2030. [R06] [R07] Major industrial announcements respond to that gap, including a proposed SODECOTON and CICAM merger targeted by 2030 and the announced Camtext SA integrated project valued at US$300 million (180 billion XAF) on 12 November 2025, designed to process 12,000 tonnes a year, create about 12,000 jobs at Dibamba and 3,000 at Garoua, and reach optimal capacity around 2033 with an estimated return on investment near 11 years. [R08] [R09] [R10] Meanwhile, historic domestic mill operations faced severe constraints: a July 2023 report put CICAM capacity utilisation at 33.7 percent in Garoua, 48.9 percent at CICAM 1 and 10.83 percent at CICAM 2, with underutilisation costing about 2 billion XAF in 2020 and restructuring estimates ranging from 30.7 billion to 48.2 billion XAF (with one relaunch plan cited at 70.2 billion XAF). [R12] [R13]

Exhibit 19, p. 73 compares Cameroon's annual seed-cotton harvests against the planned fibre intake of a major integrated textile project on a common physical scale.

EXHIBIT 19 / THOUSAND TONNES A YEAR

Aggregate cotton output dwarfs individual mill intake



SOURCED [R01] [R02] [R10] [S47] Sourced comparison from the inherited GFT evidence register: Cameroon seed-cotton production of 330.0 thousand tonnes in 2022, 314.455 thousand tonnes in 2023 and about 340.0 thousand tonnes in 2024 (R01); 2025 seed-cotton projection above 350.0 thousand tonnes (R02); and announced Camtext SA annual fibre processing of 12.0 thousand tonnes a year (R10). FAOSTAT (S47) defines the agricultural production domain boundary. Original publication dates for R01 are unrecorded in the inherited register.

In a forecasting model, FAOSTAT crop series and national cotton balances inform fibre procurement and rural income calendars in northern cotton provinces (where SODECOTON injects roughly US$100 million a year), never SKU demand. [R01] [R03] [R04] [S47] Treating the 50 percent SND30 processing target or a factory's nameplate tonnage as a demand forecast violates the demand contract established in Chapter 2. [R06] [R07] [R10]

4.14

Search and weather must earn incremental value

Digital search queries and meteorological records arrive at weekly or daily frequency. Dynamic regression and multichannel fashion studies note that external predictor variables such as school and business holidays, commercial events (Easter, Christmas), economic conditions, search behaviour and weather may explain part of historical variation. [S04] [S34] [S35] [S0403] Yet high frequency does not guarantee predictive skill for a cold-start maker. Every high-frequency external signal must be formulated as an ex-ante feature and tested against naive, seasonal-naive and attribute baselines before influencing a production or replenishment commitment. [S09] [S10] [S33] [S0401]

4.15 WEB SEARCH VOLUME IS A NEAR-TERM ATTENTION PROXY, NOT A COLD-START GUARANTEE

Research on predicting consumer behaviour with web search shows that in the absence of other data sources, or where small improvements in predictive performance are material, search queries provide a useful guide to the near future. [S34] That finding carries four strict boundary conditions for a new apparel maker:

- **Horizon decay:** Search queries measure active user intent days or a few weeks before purchase; they do not forecast seasonal assortment demand six months ahead when fabrics are ordered. [S34] [S0402]
- **Brand versus category terms:** A greenfield brand generates negligible brand-specific search before launch, and generic category queries reflect all competitors and informal sellers rather than intent to buy the new maker's offer. [S34]
- **Penetration bias:** In Cameroon, where Statista estimates apparel eCommerce user penetration at 18.0 percent in 2023 rising toward 22.7 percent in 2027, search queries over-represent urban digital buyers and under-represent offline channels. [R26] [S04]
- **Index normalisation leakage:** Public search tools return indices scaled from 0 to 100 over the requested window. Pulling a multi-year index in one request scales early weeks by future peak volume after the forecast origin, introducing preprocessing leakage. [S37]

4.16 WEATHER ANOMALIES EXPLAIN MODEST TIMING SHIFTS RATHER THAN STRUCTURAL DEMAND

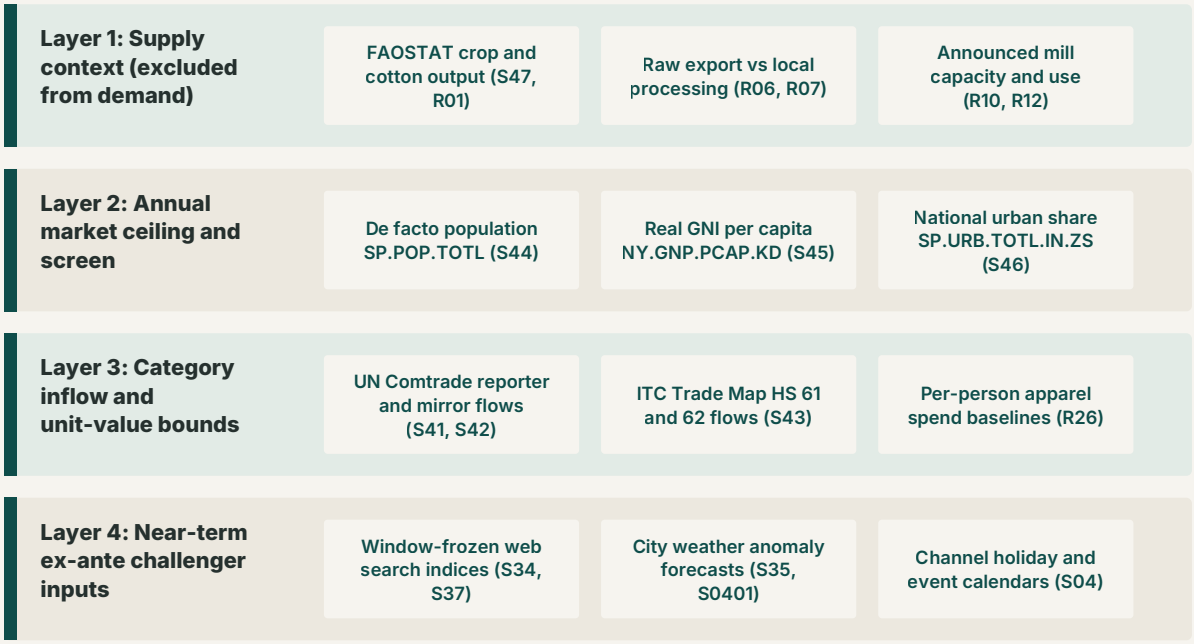
Using monthly retail sales and weather data from the National Weather Service, Federal Reserve research finds that unusual weather has a modest but significant role in explaining monthly sales fluctuations, while lagged effects frequently offset the initial impact across a quarter. [S35] Weather variables therefore obey four rules:

1. **Use anomalies against seasonal norms:** Raw temperature or rainfall mainly repeats the annual calendar, which a seasonal-naive benchmark captures at zero external-data cost. [S09] [S35]
2. **Match stations to selling cities:** National averages merge Garoua, Douala and Yaounde; weather series must match the city where the garment is sold. [S04] [S35]
3. **Separate ex-ante weather forecasts from ex-post station actuals:** Genuine ex-ante forecasts use only information available in advance and require forecasts of the predictors, whereas ex-post forecasts plug in realised predictors after the fact. [S0401] Models must be trained and tested on as-of weather forecasts. [S37] [S0401]
4. **Enforce complexity limits:** Weather features that add pipeline fragility without improving decisions over a simple baseline are rejected. [S33] [S35] [S39]

Exhibit 20, p. 75 synthesises the external data stack into four distinct evidence layers, defining where each external family stops.

EXHIBIT 20 / PERMITTED FORECASTING SCOPE

Each external signal stops at a defined evidence layer



DERIVED [S04] [S34] [S35] [S37] [S41] [S44] [S47] Derived GFT architecture mapping external data families to their maximum permitted forecasting role. Upstream agricultural data (S47) remain in supply context; World Bank macro indicators (S44) stop at annual market scaling and analogue screening; UN Comtrade trade series (S41) stop at category inflow bounds; and search, weather and calendar series (S04, S34, S35) enter near-term SKU models only as as-of audited challengers (S37).

Exhibit 20, p. 75 enforces a strict upward firewall: Layer 1 never crosses into demand estimation, Layers 2 and 3 shape only annual market bounds and hierarchical priors, and only Layer 4 signals may compete as short-horizon temporal features alongside product attributes and launch history. [S02] [S24] [S37] Even inside Layer 4, architectures with feature-selection gates such as Temporal Fusion Transformers (TFT) must prove in held-out launch replay that external inputs improve accuracy and calibration over models that suppress them. [S12] [S24] [S0403]

4.17

A covariate registry prevents a proxy from becoming a story

In a small industrial analytics team, external covariates create two compounding hazards: hidden technical debt and temporal data leakage. Machine-learning systems research warns that solutions providing a tiny accuracy benefit at the cost of massive increases in system complexity are rarely wise practice. [S39] Methodological audits show that data leakage is a spurious relationship between independent variables and the target variable arising as an artefact of data collection, sampling or preprocessing strategy. [S37] External covariates are prime carriers of leakage because macroeconomic tables are revised retroactively, trade statistics arrive months after the customs

month closes, search indices rescale across query windows, and weather archives record realised actuals rather than forecasts available at the decision origin. [S37] [S41] [S0401]

4.18 EVERY EXTERNAL FEATURE MUST PASS A BITEMPORAL REGISTRATION GATE

To prevent an external proxy from becoming a post-hoc story, GFT requires every external series to be logged in a governed Covariate Registry before feature engineering or training begins. [S36] [S37] [S40] The registry treats time as bitemporal: every record stores both the **valid time** (the calendar period the event describes) and the **as-of availability time** (the timestamp at which that value became observable to the pipeline). [S37] In expanding-window chronological cross-validation, where successive training sets are supersets of those that come before them, a feature value is admissible at the forecast origin only if its as-of availability timestamp is on or before that origin, and for future horizons only if it is a known calendar structure or an ex-ante predictor forecast issued by that origin. [S13] [S37] [S0401]

THE GFT FIVE-STEP ADMISSION TEST FOR EXTERNAL COVARIATES

1. **Boundary and layer assignment:** Map the candidate series to Layer 1, 2, 3 or 4 in **Exhibit 20**, p. 75 and reject any supply-side statistic proposed as a direct SKU demand predictor. [S47]
2. **Bitemporal as-of audit:** Verify that historical snapshots preserve publication lags, customs revisions and ex-ante predictor forecasts as they stood at each historical decision origin, with scaling parameters fitted strictly inside each expanding training window. [S13] [S37] [S0401]
3. **Baseline ablation in launch replay:** Run the cold-start backtest with and without the candidate covariate against naive, seasonal-naive and attribute-only baselines, scoring mean absolute scaled error (MASE, where defined), bias and quantile calibration. [S09] [S10] [S12]
4. **Ex-ante versus ex-post degradation check:** Compare backtest error using realised ex-post covariate values against backtest error using realistic ex-ante covariate forecasts to measure how much predictor noise erodes skill at the required horizon. [S0401]
5. **Complexity and drift sign-off:** Confirm with named commercial and technical owners that the decision improvement exceeds the maintenance and concept-drift burden; attach an automatic fallback to the covariate-free baseline if the external feed stalls. [S33] [S36] [S38] [S39] [S40]

Exhibit 21, p. 77 specifies the mandatory schema of the GFT Covariate Registry across the external data families examined in this chapter.

EXHIBIT 21 / MANDATORY GOVERNANCE FIELDS BY EXTERNAL SOURCE FAMILY

A covariate registry binds every external proxy to an as-of rule

Covariate family	Primary source	Cadence and lag	Leakage control (S37, S0401)	Permitted role
Merchandise trade (HS 61 to 63)	UN Comtrade and ITC Trade Map (S41, S43)	Annual or monthly; multi-month customs lag	Freeze snapshot date; retain reporter and mirror columns (S42)	Category inflow and unit-value bounds
Population and urbanisation	World Bank WDI SP.POP and SP.URB (S44, S46)	Annual midyear estimate; census revisions	Use vintage release date; forbid weekly linear interpolation	Market ceiling and analogue market screen
Real income per capita	World Bank WDI NY.GNP.PCAP.KD (S45)	Annual constant-price national accounts	Lock national accounts release date; never use ex-post GDP	Multi-year scenario scaling only
Cotton and crop output	FAOSTAT and national ginning reports (S47, R01)	Annual crop campaign	Firewall from demand target; log campaign publication date	Fibre supply context and rural calendar
Web search query volume	Search query index APIs (S34)	Weekly; near-real-time index rescaling	Scale strictly inside each training fold up to origin	Short-horizon challenger for active offers
City weather anomalies	Meteorological station feeds (S35)	Daily or weekly; forecast error grows with horizon	Train and test on as-of weather forecasts, not ex-post actuals	Short-horizon replenishment modifier

DERIVED [S34] [S35] [S37] [S41] [S44] [S47] [S0401] Derived GFT governance specification combining external source definitions (S34, S35, S41, S44, S47) with leakage controls (S37) and ex-ante predictor rules (S0401). Every external feature must populate these columns before entering an evaluation fold.

Exhibit 21, p. 77 turns methodological caution into an auditable engineering gate. When an external feed fails its bitemporal check, or when removing the feature leaves launch-replay error and planning decisions unchanged, the feature is retired from the model pipeline rather than retained for narrative appeal. [S33] [S37] [S39] Continuous monitoring tracks summary statistics of every retained covariate and triggers a notification or automatic rollback to the covariate-free baseline when an upstream agency alters a definition, delays a release or introduces a structural break. [S38] [S40]

GFT NOTE

In GFT's Cameroon textile programme, external statistics are separated strictly by decision horizon. Feasibility and capacity-sizing models use UN Comtrade and ITC Trade Map series, World Bank demographic and income indicators, Statista expenditure baselines and national cotton-processing benchmarks to bound the macro opportunity and stress-test multi-year scenarios. [R06] [R26] [S41] [S43] [S44] [S45] [S46] None of those series is permitted to nominate a winning garment style or set a first-lot SKU production commitment. Before launch, SKU priors come from audited analogue distributions (Chapter 3) and calibrated primary choice research (Chapter 5); after launch, short-term covariates such as search or weather enter the forecasting service only if they pass the bitemporal registry gate and beat simple baselines in held-out launch replay. [S09] [S34] [S35] [S37] This paper provides an engineering and forecasting methodology, not investment, legal or tax advice.

By closing Part I with a governed covariate registry, the cold-start maker establishes a complete outside view: a demand contract that distinguishes latent demand from sales (Chapter 2), an analogue prior that preserves launch dispersion (Chapter 3), and external market bounds that respect provenance and chronology (Chapter 4). Part II turns from borrowed context to owned pre-launch and pilot evidence, showing how structured primary research and controlled release experiments convert market bounds into style-level purchase signals.

External statistics draw the walls of the arena; only customer choices and clean launch records can name the garments that sell inside it.

CHAPTER SOURCE NOTES

S41 United Nations Statistics Division, UN Comtrade Help Center. 2026-10-30.

S43 International Trade Centre. 2014.

S44 World Bank, World Development Indicators. 2026-10-30.

S45 World Bank, World Development Indicators. 2026-10-30.

S46 World Bank, World Development Indicators. 2026-10-30.

R06 Cameroon Trade Hub. Publication date unrecorded; inherited source-text check.

S47 Food and Agriculture Organization of the United Nations. 2025.

S09 PLOS ONE, Public Library of Science. 2018-03-27.

S34 Proceedings of the National Academy of Sciences. 2010.

S35 Board of Governors of the Federal Reserve System, Finance and Economics Discussion Series. 2000.

S0401 OTexts. 2021-05-31.

S01 InTech, International Journal of Engineering Business Management. 2013.

S0402 arXiv. 2022-11-09.

S08 International Journal of Forecasting, Elsevier; full text through arXiv. 2022.

S37 Cell Press, Patterns; open full text from PubMed Central. 2023-08-04.

R07 Ecofin Agency. Publication date unrecorded; inherited source-text check.

S06 Lancaster University Management School, Munich Personal RePEc Archive. 2018-11-19.

S02 arXiv. 2019-06-27.

S03 arXiv. 2020-07-10.
R11 Business in Cameroon; MMI News. Publication date unrecorded; inherited source-text check.
R16 WageIndicator. Publication date unrecorded; inherited source-text check.
R26 Statista. Publication date unrecorded; inherited source-text check.
R18 Congressional Research Service IF10149; AGOA (Wikipedia). Publication date unrecorded; inherited source-text check.
R23 MacMap; AfCFTA. Publication date unrecorded; inherited source-text check.
R24 Mordor Intelligence. Publication date unrecorded; inherited source-text check.
R25 Market.us. Publication date unrecorded; inherited source-text check.
S0403 OTexts. 2021-05-31.
S42 United Nations Statistics Division, UN Comtrade Help Center. 2026-10-30.
R32 Wise; XE. Publication date unrecorded; inherited source-text check.
S40 Google Cloud Documentation. 2024-08-28.
R10 Ecofin Agency. Publication date unrecorded; inherited source-text check.
R14 WageIndicator. Publication date unrecorded; inherited source-text check.
R15 wage.is. Publication date unrecorded; inherited source-text check.
S28 Journal of Industrial Ecology, Wiley. 2022-10-17.
R01 Business in Cameroon. Publication date unrecorded; inherited source-text check.
R02 Fibre2Fashion; Business in Cameroon. Publication date unrecorded; inherited source-text check.
R05 Ecofin Agency. Publication date unrecorded; inherited source-text check.
R03 SODECOTON. Publication date unrecorded; inherited source-text check.
R04 Financial Fortune Media. Publication date unrecorded; inherited source-text check.
R08 Business in Cameroon. Publication date unrecorded; inherited source-text check.
R09 Ecofin Agency. Publication date unrecorded; inherited source-text check.
R12 Business in Cameroon. Publication date unrecorded; inherited source-text check.
R13 Business in Cameroon; MMI News. Publication date unrecorded; inherited source-text check.
S04 MDPI, Processes. 2021-09-03.
S10 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2018.
S33 Google for Developers. 2025-08-25.
S39 Neural Information Processing Systems Foundation, NIPS 2015. 2015.
S24 arXiv. 2020-09-27.
S12 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2006.
S36 National Institute of Standards and Technology, U.S. Department of Commerce. 2023.
S13 scikit-learn. 2026-10-30.
S38 ACM Computing Surveys; open record from Bournemouth University. 2014-04-30.

GREEN FUTURE TECH

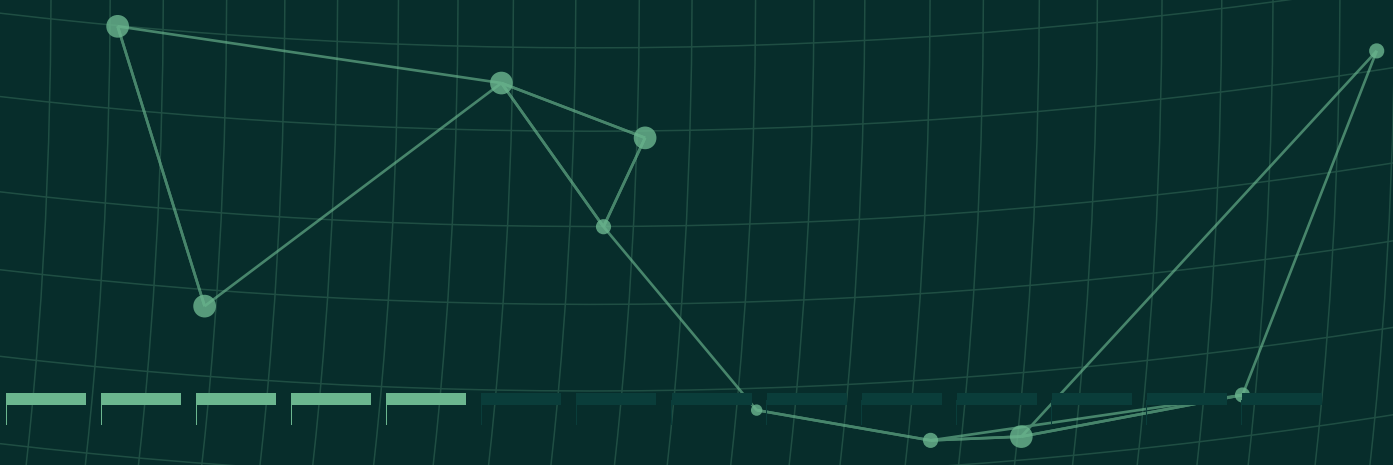
LIMITED / HONG KONG SAR

05

PART II: EARN EVIDENCE BEFORE SCALE / PRE-LAUNCH EVIDENCE

Primary research turns intent into an explicit prior

Stated interest becomes useful only when sampling, incentives and calibration connect it to a future purchase decision.



CHAPTER 05
DECISION TEST

Stated interest becomes useful only when sampling, incentives and calibration connect it to a future purchase decision.

Primary market research earns a place in a cold-start demand forecast only when it is engineered as a calibrated probability distribution rather than accepted at face value as a promise of future sales. [S31] A greenfield apparel manufacturer in Cameroon cannot train a supervised time-series model on owned transaction history before its first production lot reaches the market, nor can it infer style-level buyer preference from aggregate national clothing expenditure of about US$94.70 per person and about 8.0 pieces per person in 2024 or from reported raw-cotton export shares of roughly 95 percent against 5 percent local processing. [R06] [R26] When historical observations for the target item do not exist, forecasting theory requires structured judgemental and preference-elicitation methods that link stated human responses to a concrete operational commitment. [S08] [S0501] Unadjusted survey enthusiasm, informal distributor encouragement and open-room executive consensus systematically overstate willingness to pay because hypothetical choices carry no financial consequence for the respondent. [S0502] To convert pre-launch fieldwork into a defensible prior for initial lot sizing, assortment selection and channel allocation, an industrial planning team must stratify recruitment by actual buying roles, restrict choice attributes to specifications the factory can physically deliver, apply an explicit mathematical calibration from stated choice to realised purchase probability, isolate expert judgement in anonymous iterative panels, track information dependencies so a single belief is never counted twice, and lock the entire protocol in a pre-registered specification before the first interview begins. [S20] [S29] [S31] [S0503]

5.1

Respondent segments follow use occasions and buying roles

A pre-launch survey that pools every respondent into an undifferentiated consumer average conceals the operational boundaries at which garment buying decisions actually occur. [S01] [S31] In apparel markets, the person who wears a garment, the person who pays for it, the wholesale trader who finances inventory in regional market stalls, and the institutional officer who specifies uniform tenders operate under distinct budget constraints, order calendars and substitution rules. [S01] [S06] Treating these four roles as a single sample produces an uninterpretable preference score: high aesthetic approval from urban students who do not control household clothing budgets can mask price rejection by the parents or procurement officers who authorise payment. [S0501] [S0502]

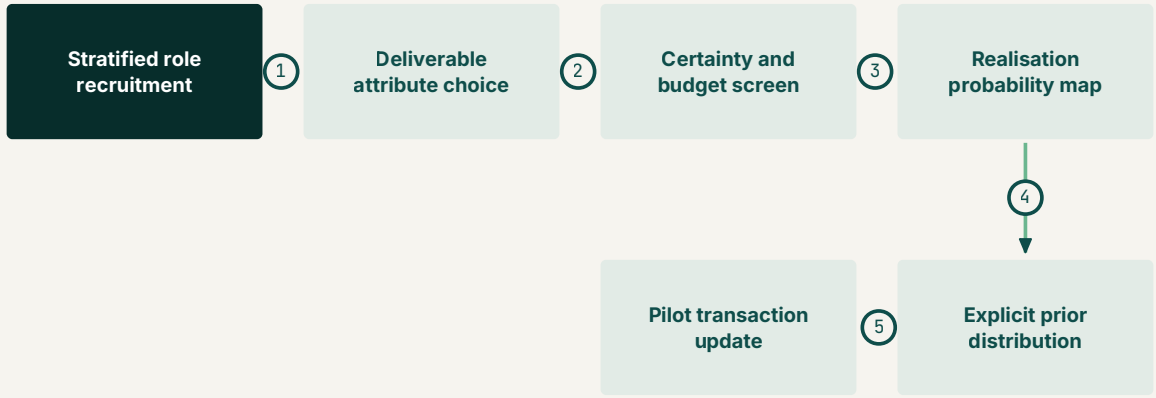
5.2 BUYING ROLES SEPARATE END-USER UTILITY FROM ORDER COMMITMENT

Research design for a new apparel maker must partition the target population along two orthogonal axes: the use occasion of the garment and the decision authority of the participant. [S01] [S31] In GFT's Cameroon textile programme, prepared in January 2024 with source refreshes in 2026, commercial scoping distinguishes everyday retail basics, school and academic uniforms, corporate and industrial workwear, and regional wholesale bale or dozen packs. Each category has its own buying unit and payment rhythm. In a market where informal employment accounts for about 90 percent of the workforce, the statutory minimum wage (SMIG) stands at XAF 43,969 a month from 23 February 2024 (about US$76, up 5 percent from XAF 41,875 in 2023), and average monthly earnings across formal sectors sit near US$359, house-

hold liquidity fluctuates sharply across the calendar month and the school-fee cycle. [R14] [R15] [R16] A retail wearer evaluating a cotton polo shirt considers fit, colour fastness and social occasion, whereas a wholesale distributor in Douala or Garoua evaluates pack margin, credit terms, size-run balance and the risk of slow turn against imported second-hand or low-cost Asian alternatives. [S02] [S06]

EXHIBIT 22 /

Stated preference reaches a forecast through calibration



1. Role quota
2. Raw utilities
3. Ex-post filter
4. Quantile bounds
5. Lot commitment

DERIVED [S31] [S48] [S0501] [S0502] Derived method sequence combining attribute conjoint design (S48), hypothetical-bias mitigation (S0502), preference-to-demand realisation calibration (S31) and customer-intention audit rules (S0501).

As shown in **Exhibit 22**, p. 82, primary research does not feed raw survey tallies directly into production scheduling. [S31] [S48] Instead, stratified role recruitment governs who enters the choice task, deliverable attribute constraints bound the trade-offs they evaluate, ex-post certainty and budget screens filter inconsequential selections, and a realisation probability map translates filtered utilities into an explicit prior distribution that awaits pilot transaction verification. [S31] [S0502] While **Exhibit 22**, p. 82 establishes the required information sequence from recruitment to pilot gate, it cannot by itself guarantee that a chosen sample represents the full target market if field enumerators recruit only convenient respondents.

5.3 RECRUITMENT METADATA PREVENTS UNREACHABLE SEGMENTS FROM INFLATING THE PRIOR

To prevent convenience sampling from masquerading as market demand, every interview record must capture verifiable screening covariates before presenting any product concept. [S31] [S37] Required fields include the respondent's geographic cell (such as Douala, Yaounde, Garoua or secondary urban centres), primary purchasing channel (open-air market stall, independent boutique, supermarket, institutional

tender or e-commerce platform), past-twelve-month purchase frequency in the specific garment category, actual unit price paid on the most recent purchase, and role in the buying decision (sole payer, joint household decider, wearer without budget authority, trade reseller or institutional buyer). [R26] [S03] [S31] In Cameroon, where apparel e-commerce revenue was estimated at US$20.32 million in 2023 and projected to reach US$33.61 million by 2027 with user penetration rising from 18.0 percent to 22.7 percent, an online-only survey automatically excludes the roughly four-fifths of residents who purchase garments exclusively through physical channels. [R26]

When post-stratification weights are applied to convert sample responses into a channel-level prior, cells with fewer than an illustrative threshold of 30 completed, quality-screened interviews are flagged as sparse rather than extrapolated with excessive weight into national volume estimates. [S28] [S36] If a recruitment cell for institutional uniform specifiers yields only 12 verified procurement officers, the forecasting team retains those responses as a narrow qualitative range for tender scenario planning and refuses to multiply them by a national corporate headcount. [S08] [S36] Recording channel, location, occasion and purchase responsibility at recruitment ensures that an attractive aggregate preference score in one sub-sample cannot hide zero reach or negative margin in the channel where the factory intends to sell. [S31] [S0501]

5.4

Conjoint separates attribute preferences before launch

Asking prospective buyers whether they like a single prototype garment invites polite agreement and provides no information on how demand will shift when price, fabric weight or delivery lead time changes. [S48] [S0502] Choice-based conjoint analysis (also termed a discrete choice experiment) resolves this limitation by presenting respondents with competing multi-attribute product profiles and requiring them to select one option or an explicit no-purchase alternative. [S48] [S0502] By observing trade-offs across orthogonal or near-orthogonal experimental profiles, the analyst estimates part-worth utility distributions for individual product features before a single commercial lot is cut. [S02] [S48]

5.5 FEASIBILITY BOUNDARIES GOVERN THE ATTRIBUTE MATRIX

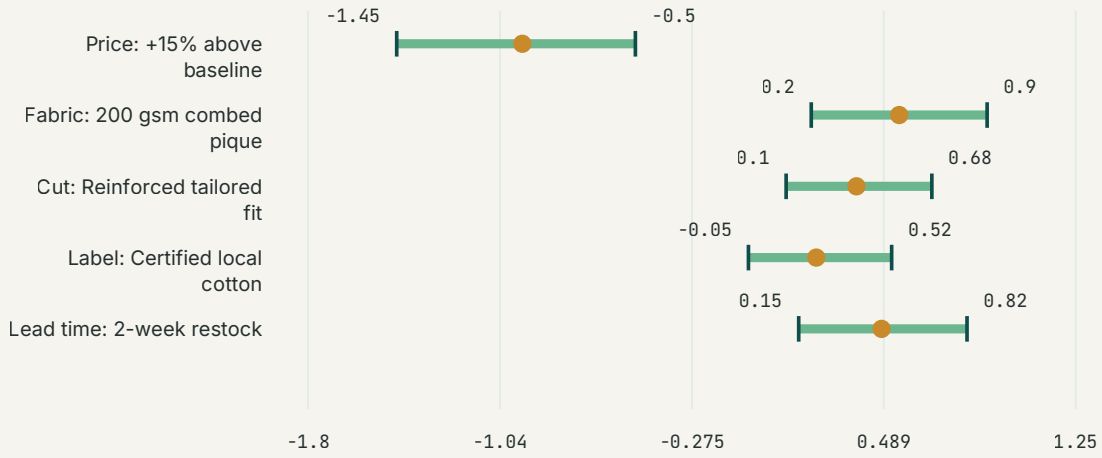
A recurring failure in pre-launch consumer research is testing attribute levels that the manufacturing line cannot physically or economically produce during its initial ramp. [S36] [S48] If a survey tests a combed ring-spun cotton shirt at a retail price equivalent to low-grade imported polyester and a seven-day restock lead time that the mill's spinning and dyeing schedule cannot meet, the resulting choice utilities describe a fictional product. [S02] [S0502] Before the conjoint instrument is finalised, production engineering and merchandising leads must sign off on a feasible attribute matrix in which every level corresponds to a verified bill of materials, machine capability and costed price band. [S02] [S36]

For a greenfield Cameroonian apparel line, the feasible attribute set typically spans five operational dimensions: 1. **Fabric construction and weight**: for example, 160 grams per square metre (gsm) carded single jersey versus 200 gsm combed piqué

versus a 65/35 cotton-polyester woven twill, each linked to known yarn counts and shrinkage tolerances in the factory's specification sheet (illustrative engineering levels). [S02] [S48] 2. **Garment silhouette and finish**: standard tubular cut versus side-seamed tailored fit, collar reinforcement, and stitch density, which directly govern sewing standard allowed minutes on the line. [S02] 3. **Origin and provenance marking**: unbranded basic label versus explicit Cameroon-grown and locally processed cotton certification, testing whether domestic buyers or corporate specifiers assign positive utility to local value addition beyond the 5 percent baseline processing share. [R06] [S48] 4. **Commercial pack and delivery terms**: single-piece retail hangtag, six-piece assorted size pack for market traders, or 100-piece institutional carton with a stated replenishment lead time of two, four or six weeks (illustrative commercial levels). [S06] 5. **Effective unit price in CFA francs (XAF)**: three to four discrete price tiers spanning the venture's break-even floor up to the prevailing retail price of imported branded equivalents, converted where relevant at the fixed euro peg of XAF 655.957 or reference US dollar rates. [R32] [S31]

EXHIBIT 23 / ILLUSTRATIVE LOG-ODDS UTILITY

Attribute utilities should remain distributions with sample provenance



ILLUSTRATIVE [S28] [S31] [S48] Illustrative hierarchical Bayes choice-model output across five deliverable garment attributes (10th, 50th and 90th percentile posterior draws). Values illustrate model structure based on S28, S31 and S48 and are not observed Cameroon market shares.

Exhibit 23, p. 84 illustrates how part-worth estimates from a hierarchical Bayesian multinomial logit model should be reported to commercial and production planners. [S28] [S48] Rather than publishing a single bar chart of average utilities, the exhibit retains the 10th, 50th and 90th percentile intervals across respondents in the sampled segment. [S14] [S16] In this illustrative readout, raising the unit price by 15 percent above the baseline tier imposes a negative utility shift with a median of -0.95 log-odds (spanning -1.45 to -0.50), whereas upgrading from light jersey to 200 gsm combed piqué adds a median of +0.55 log-odds (+0.20 to +0.90) and guaranteeing a two-week restock lead time for wholesale buyers adds a median of +0.48 log-odds (+0.15 to +0.82). [S31] [S48] By contrast, the illustrative interval for a certified local-cotton label (-0.05 to +0.52, median +0.22) crosses zero at its lower decile, warning planners that provenance messaging alone cannot compensate for a price penalty or a weak fabric handfeel. What **Exhibit 23**, p. 84 cannot establish is the absolute

volume of garments sold per week; relative utility differences must still be anchored to an outside no-purchase option and calibrated against real purchase frequencies. [S31] [S0502]

5.6 THE NO-PURCHASE OPTION ANCHORS RELATIVE UTILITY TO MARKET PARTICIPATION

A forced-choice conjoint task in which every screen compels the respondent to pick one of three garment cards measures only relative preference conditional on buying. [S31] [S48] If all three cards on a screen are overpriced relative to the buyer's real budget, a forced choice records a vote for the least-disliked shirt even though the customer would walk out of the shop empty-handed. [S0502] Every choice set in a cold-start demand study must therefore include a dual-response or direct outside option: "I would keep my money or buy my usual alternative instead of any of these options." [S31] [S0502] Including a realistic status-quo comparator (such as a specific grade of imported Asian polo shirt or second-hand market garment at its observed local street price in XAF) forces the new venture's attributes to compete against the actual outside alternatives available to Cameroonian buyers rather than against vacuum abstractions. [S06] [S31]

5.7

Choice responses need calibration against purchase settings

Even with a deliverable attribute matrix and an explicit no-purchase option, raw stated-choice probabilities systematically overestimate real-world purchase rates. [S0501] [S0502] Across empirical valuation and marketing studies, hypothetical bias arises because survey selections are non-binding, social desirability encourages respondents to endorse locally made or higher-specification goods, and cognitive attention during a fifteen-minute interview differs from cash allocation at a market stall. [S0501] [S0502] Preference-learning research demonstrates that inexpensive pre-launch choice data can still produce accurate demand forecasts in real purchase settings, provided the model explicitly estimates and applies realization probabilities that map stated utility to actual transaction execution. [S31]

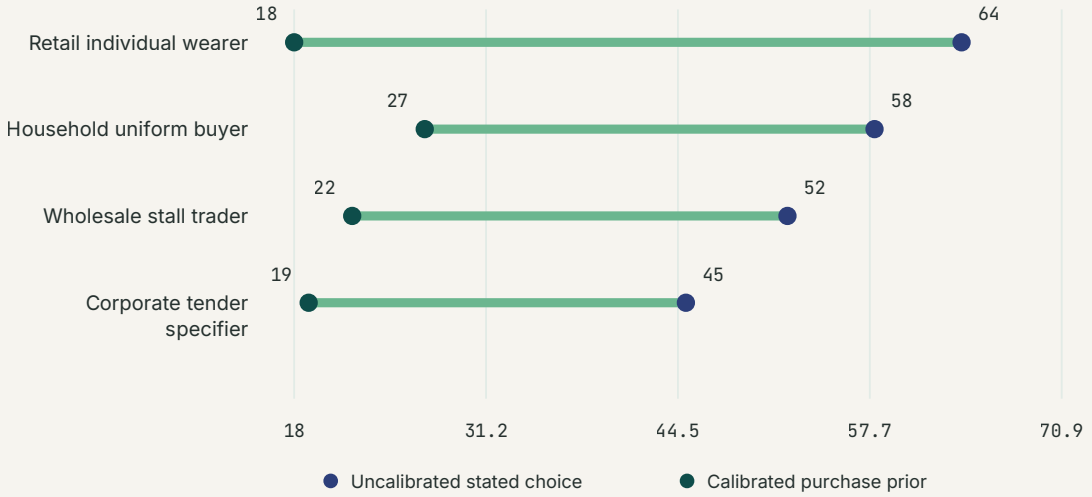
5.8 EX-ANTE DESIGN AND EX-POST CERTAINTY SCALING REDUCE HYPOTHETICAL INFLATION

GFT structures choice-to-purchase calibration in two stages: ex-ante protocol controls that discipline respondent behaviour during the interview, and ex-post probabilistic down-weighting that adjusts the mathematical prior afterward. [S31] [S0502] At the ex-ante stage, the instrument opens with a consequentiality and budget-reminder script (often termed a cheap-talk and budget-alignment protocol) that explains why overstating purchase intent leads to mispriced goods and asks the respondent to name their actual household or business clothing outlay in XAF over the preceding thirty days before viewing any choice cards. [S0502] Where regulations and field logistics allow incentive-aligned designs, respondents enter a binding lottery or pre-order voucher draw in which one randomly selected choice task is executed for real cash and physical garment delivery upon pilot release, directly linking stated preference to financial consequence. [S31] [S0502]

At the ex-post stage, every positive product selection is followed immediately by a numerical follow-up certainty scale (from 1, meaning very uncertain, to 10, meaning certain to pay immediately in cash at the stated price) alongside an availability-timing check. [S0501] [S0502] Empirical calibration rules do not treat a certainty score of 6 out of 10 as a 60 percent purchase probability; instead, responses below a strict threshold (typically 8 or 9 on a 10-point scale, combined with verified past-year category spend at or above the tested price tier) are recoded to the no-purchase outside option or assigned a discounted realisation kernel. [S31] [S0502]

EXHIBIT 24 / PERCENT OF SAMPLED SEGMENT

Calibration compresses raw stated choice into conservative purchase bounds



ILLUSTRATIVE [S31] [S0501] [S0502] Illustrative comparison of raw stated choice share against calibrated expected purchase probability across four buyer segments, demonstrating the realisation-probability compression described in S31, S0501 and S0502. Not an observed GFT operating result.

Exhibit 24, p. 86 demonstrates the magnitude of compression that an explicit calibration layer introduces across four illustrative buyer segments. [S31] [S0502] While raw stated choice shares in this illustrative pre-launch readout range from 45 percent among corporate tender specifiers to 64 percent among individual retail wearers, applying budget-consistency screens, follow-up certainty thresholds and segment-specific realisation multipliers compresses the expected purchase prior to 18 percent for individual retail wearers, 27 percent for household school-uniform buyers, 22 percent for wholesale stall traders and 19 percent for corporate specifiers. [S31] [S0501] Notice that calibration can reorder segments as well as shrink them: individual retail wearers show the highest raw enthusiasm (64 percent illustrative) but the steepest discount down to 18 percent because many respondents lack immediate cash liquidity or habitual purchase frequency at the tested price, whereas household school-uniform buyers face a mandatory seasonal purchase event that preserves a higher calibrated conversion share (27 percent illustrative). [S0501] [S0502] **Exhibit 24**, p. 86 illustrates why planning lot sizes from raw survey percentages is hazardous, though the exact compression ratio for any Cameroonian product family must ultimately be updated against real pilot transactions in Chapter 6. [S31] [S32]

METHOD

Mathematical specification of the choice-to-purchase calibration kernel

To keep the bridge from stated preference to the demand prior auditable, GFT specifies the expected unit demand $D_{i,s,h}$ for style i in buyer segment s over launch horizon h as the product of four explicit distributions rather than a single deterministic multiplier: [S14] [S31]

$$D_{i,s,h} = N_s \times A_{i,s,h} \times \sum_{k \in \mathcal{C}} \left(P^{\text{stated}}_{i,s,k} \times \rho_{s,k} \right) \times Q_{i,s,h}$$

1. **Reachable segment denominator (N_s)**: the count of active buying units in segment s within the venture's physical distribution footprint, bounded by demographic and channel covariates from Chapter 4. [S44] [S46]
2. **Planned availability and reach factor ($A_{i,s,h} \in [0,1]$)**: the fraction of segment s that will actually encounter style i in stock with complete size availability during horizon h, preventing the prior from assuming universal distribution on day one. [S06] [S07]
3. **Stated choice share by certainty cell ($P^{\text{stated}}_{i,s,k}$)**: the posterior probability from the hierarchical conjoint model that a respondent in segment s chooses style i over competing offers and the no-purchase option, partitioned into certainty and budget-verification tier $k \in \mathcal{C}$. [S28] [S48]
4. **Realisation probability distribution ($\rho_{s,k} \sim \text{Beta}(\alpha_{s,k}, \beta_{s,k})$)**: the stochastic conversion rate from stated choice in cell k to an actual cash transaction within horizon h, initialised from conservative literature priors and updated via beta-binomial conjugacy as pilot redemption vouchers or first-lot transactions arrive. [S28] [S31]
5. **Conditional order quantity ($Q_{i,s,h}$)**: the distribution of units purchased per converting buyer (for example, 1 to 2 pieces for an individual wearer versus 12 to 60 pieces for a market stallholder buying in dozen packs). [S06]

Monte Carlo propagation across $(\theta_{\text{conjoint}}, \rho_{s,k}, Q_{i,s,h})$ yields the nine standard predictive quantiles ($q_{0.005}$ through $q_{0.995}$) required by the forecast contract. [S14] [S16]

WORKED EXAMPLE

Illustrative calibration of a 200-respondent choice survey into a twelve-week initial production prior

Consider an illustrative pre-launch study for a men's 200 gsm combed cotton piqué polo shirt priced at XAF 6,500 retail (about US$10.70 at an illustrative reference rate of XAF 607 per US dollar, consistent with the XAF 655.957 euro peg) across two independent Douala boutique clusters representing a reachable active shopper base of $N_s = 8,000$ target buyers over a twelve-week launch window. [R32] [S31] Planned shelf and size-run availability across the pilot shops is set at $A_{i,s,h} = 0.75$, giving $8,000 \times 0.75 = 6,000$ effectively reached buyers. [S06]

Suppose an illustrative stratified sample of $n = 200$ verified category buyers completes the discrete choice task against imported alternatives priced at XAF 5,500 and XAF 8,000 plus an explicit no-purchase option. [S48] Out of 200 respondents, 110 select the new XAF 6,500 polo shirt, yielding an uncalibrated stated choice share of $110 / 200 = 55.0$ percent. [S0502] If a naive planner multiplied 6,000 reached buyers by 55.0 percent and an assumed 1.2 pieces per buyer, the uncalibrated forecast would claim 3,960 units.

Instead, the GFT calibration protocol partitions the 110 positive selections into three mutually exclusive verification cells based on follow-up certainty (1 to 10 scale) and verified past-year purchase receipts at or above XAF 6,000: [S31] [S0502]

- **Cell C1 (High certainty, verified budget)**: 36 respondents (18.0 percent of the 200-person sample) score 9 or 10 on certainty and report verified prior purchases at or above XAF 6,000. The pre-registered realisation prior for Cell C1 has a mean of $\bar{\rho}_1 = 0.55$ (illustrative 10th to 90th percentile interval: 0.40 to 0.70).
- **Cell C2 (Moderate certainty or unverified budget)**: 44 respondents (22.0 percent of the sample) score 7 or 8 on certainty. The realisation prior for Cell C2 has a mean of $\bar{\rho}_2 = 0.20$ (illustrative 10th to 90th percentile interval: 0.10 to 0.32).
- **Cell C3 (Low certainty or budget mismatch)**: 30 respondents (15.0 percent of the sample) score below 7 or report never spending above XAF 4,000 on a shirt. The realisation prior for Cell C3 has a mean of $\bar{\rho}_3 = 0.03$ (illustrative 10th to 90th percentile interval: 0.00 to 0.08).

Expected converting buyers per reached shopper is calculated in this illustrative case by summing across the three cells: $\sum_{k=1}^{3} P^{\text{stated}}_{k} \times \bar{\rho}_{k} = (0.180 \times 0.55) + (0.220 \times 0.20) + (0.150 \times 0.03) = 0.0990 + 0.0440 + 0.0045 = 0.1475$. [S31]

Thus, the calibrated central conversion rate is 14.75 percent of reached buyers, rather than the raw 55.0 percent stated share (an illustrative compression factor of 3.73 to 1). Multiplying 6,000 effectively reached buyers by 0.1475 yields 885 converting buyers. With an illustrative conditional purchase quantity of $Q = 1.15$ pieces per converting retail buyer, the central twelve-week demand prior is $885 \times 1.15 = 1,018$ units (gross retail value of about XAF 6.62 million at XAF 6,500 per unit). Propagating the 10th and 90th percentile bounds of ρ_k and sampling error yields an illustrative 80 percent predictive interval of 640 units ($q_{0.10}$) to 1,460 units ($q_{0.90}$). Production planners can now commit an initial pilot lot of 750 to 900 units with reserved greige fabric for a fast replenishment run if early sell-through tracks above the median, rather than overproducing 3,960 units on uncalibrated survey optimism. [S14] [S31]

5.9

Price research produces scenarios, not a universal elasticity

When an apparel venture lacks historical transaction records across price changes, managers frequently ask primary research to deliver a single price-elasticity coefficient (for example, claiming that a 10 percent price cut will raise demand by 18 percent across the market). [S03] [S31] Treating a survey-derived price slope as a universal structural elasticity is a category error. [S48] [S0502] Choice experiments measure willingness-to-pay trade-offs conditional on the competing profiles shown on the interview card, the respondent's immediate framing and the sampled channel; they do not observe general-equilibrium competitor retaliation, credit constraints among wholesale buyers, or seasonal liquidity shocks. [S06] [S31] [S48]

5.10 WILLINGNESS TO PAY IS BOUNDED BY REFERENCE PRICES AND INCOME TIERS

In Cameroon's clothing market, price response is highly non-linear and exhibits sharp threshold effects around psychological currency notes (such as XAF 2,000, XAF 5,000 and XAF 10,000) and existing channel reference prices. [R14] [R26] [R32] Against an average per-capita clothing and footwear expenditure of about US$94.70 a year (roughly XAF 57,500 a year at illustrative reference exchange rates) across roughly 8.0 pieces per person, an everyday garment priced near XAF 3,500 to XAF 5,000 competes in a mass replacement tier, whereas a garment priced above XAF 12,000 shifts into an occasional or formal discretionary tier. [R26] [R32] A conjoint model fitted across three discrete price levels (for instance, XAF 5,000, XAF 6,500 and XAF 8,000) identifies the local utility difference between those tested nodes for the sampled respondents. [S31] [S48] Extrapolating a linear log-price coefficient outside that tested interval, or applying a boutique shopper's price slope to an open-air market stallholder, fabricates precision where no evidence exists. [S08] [S0502]

Consequently, GFT forbids encoding primary price research as a continuous elasticity parameter inside the automated forecasting service. [S33] [S36] Instead, the discrete price tiers tested in the conjoint instrument are stored as explicit conditional scenarios (Low-Price Penetration Scenario, Baseline Margin Scenario, Premium Positioning Scenario), each carrying its own full quantile distribution ($q_{0.05}$, $q_{0.25}$, $q_{0.50}$, $q_{0.75}$, $q_{0.95}$) and sample provenance tag. [S08] [S16]

5.11 RECONCILING STATED PRICE SENSITIVITY WITH PILOT AVAILABILITY

A second hazard in pre-launch price analysis is ignoring the interaction between unit price, stock availability and substitution when the first pilot lots enter physical shops. [S06] [S07] Suppose a venture tests two price cells in a twelve-week pilot across matched retail outlets, placing a basic cotton T-shirt at XAF 4,500 in Cell A and XAF 5,500 in Cell B. [S32] If the XAF 4,500 outlets experience stockouts in medium and large sizes during week four because replenishment was sized to the central prior rather than the 80th percentile, raw sales volumes in Cell A will be censored downward. [S06] [S07] Comparing uncensored sales in Cell B against stockout-censored sales in Cell A would falsely imply that buyers are price-insensitive, when in reality latent demand at XAF 4,500 was truncated by empty shelves. [S07]

Therefore, no survey-based willingness-to-pay distribution is updated or replaced by pilot price-response estimates until the pilot transaction log has been audited for stockout timestamps, size-run completeness and concurrent promotional discounts as specified in Chapter 2 and Chapter 7. [S06] [S07] [S37] Until uncensored, randomized or quasi-experimental price variation is observed in live transactions, price remains a scenario selector owned by commercial leadership rather than an autonomous feature weight. [S32] [S36]

5.12

Delphi makes expert judgement anonymous and revisable

While buyer choice surveys capture end-user and retail shopper preferences, experienced wholesale merchants, garment production managers, institutional tender brokers and regional textile traders hold practical domain knowledge that a consumer questionnaire cannot reach. [S08] [S0501] These practitioners understand seasonal shipping delays, school-opening surges, fabric durability complaints and credit defaults in regional markets. However, eliciting their expertise through an unstructured roundtable meeting or a sales-force composite introduces severe cognitive and organisational biases. [S0501]

5.13 WHY OPEN-ROOM MEETINGS AND SALES COMPOSITES CORRUPT THE PRIOR

Forecasting literature documents three systematic failures when judgemental forecasts are collected without structural safeguards. [S08] [S0501] First, in open face-to-face meetings, group dynamics encourage deference to the highest-paid executive or the loudest commercial advocate, anchoring the entire room to an initial optimistic target and suppressing dissenting operational warnings. [S0501] Second, when salespeople or distributor account managers are asked to submit channel forecasts (a sales-force composite), the principle of segregating forecasters from users is violated: an account manager whose bonus depends on beating a quota has a direct incentive to under-forecast, whereas a project sponsor seeking capital approval has a direct incentive to over-forecast. [S0501] Third, unstructured experts rarely state the reference class or historical analogy behind their numbers, making it impossible for a planner to tell whether two panellists agree because they possess independent evidence or because both read the same headline target, such as Cameroon's 50 percent local-processing ambition by 2030. [R07] [S20] [S0501]

5.14 STRUCTURED DELPHI AND STRUCTURED ANALOGIES PRESERVE DISSENT

To extract domain expertise without groupthink or political distortion, GFT implements a two-round or three-round anonymous Delphi protocol combined with structured analogy documentation. [S29] [S0501] Originally developed at RAND and now standard in judgemental forecasting, the Delphi method replaces open committee debate with four formal controls: [S29] [S0501] 1. **Anonymity of panellists**: participants complete structured questionnaires independently (via an online form or sealed field sheet) so that no participant knows who authored a particular estimate or rationale. [S29] [S0501] 2. **Decomposition into specific quantities and quantiles**: instead of asking "How well will our workwear line sell?", the instrument asks panellists to estimate three explicit quantiles (10th percentile downside, 50th percentile median, and 90th percentile

upside) for defined metrics such as first-quarter wholesale reorders per active trader, size-return ratios, or institutional tender win rates. [S14] [S29] 3. **Mandatory structured analogy logging**: following empirical findings that experts who explicitly cite two or more concrete analogies with direct personal experience produce significantly more accurate forecasts, every panellist must name the specific historical product launches, fabric lines or regional rollouts on which their quantile estimates rest, rating the similarity of each analogy to the new venture on a documented 1-to-5 scale. [S0501] 4. **Controlled statistical feedback and revision**: after Round 1, an independent facilitator compiles the anonymous median, interquartile range and anonymised written rationales (especially for estimates outside the interquartile range) and returns them to the panel. [S29] [S0501] In Round 2, panellists may revise their quantiles in light of reasoning they had not previously considered, or they may retain their original numbers and record why the majority view overlooks a specific constraint. [S29]

Crucially, the GFT Delphi protocol never forces artificial consensus at the end of the final round. [S29] [S0501] When panellists remain divided after seeing each other's evidence (for example, when northern cotton-belt traders forecast strong seasonal demand after harvest payments while coastal urban retailers forecast sluggish turn due to competing imports), the spread across panellists' final distributions is preserved in full as epistemic uncertainty in the prior. [S08] [S14] [S29] Averaging away informed dissent before launch erases the exact tail scenarios that inventory buffers and flexible capacity allocations need to see. [S16] [S29]

GFT NOTE

Practical recommendation for small greenfield teams

A greenfield apparel manufacturer does not need expensive enterprise survey software or a standing econometrics department to run a disciplined primary-research prior. [S39] In GFT's Cameroon programme design, a small commercial and analytics team of two to three staff can execute the entire Phase 0 elicitation protocol using standard spreadsheet templates, offline-capable tablet forms and a version-controlled text repository: [S39] [S40]

- **Keep the Delphi panel small, diverse and external to project targets**: recruit 6 to 10 independent practitioners spanning wholesale fabric and garment traders, institutional procurement officers, garment production supervisors and retail store managers, explicitly excluding executives whose compensation or project approval depends on the launch volume forecast. [S29] [S0501]
- **Limit conjoint tasks to 8 to 10 choice screens per respondent**: each screen should show at most three deliverable garment profiles plus the explicit no-purchase or status-quo option, followed by the 1-to-10 certainty question and past-purchase verification. [S48] [S0502] Longer questionnaires induce respondent fatigue and random clicking after fifteen minutes.
- **Freeze the calibration arithmetic in an inspectable script**: store the certainty-cell realisation multipliers ($\rho_{s,k}$) in a plain configuration file alongside the raw anonymised survey rows so that any planner can re-run the calculation and verify how the prior changes under conservative versus central conversion assumptions. [S31] [S37]

5.15

Evidence synthesis must avoid counting one belief twice

By the close of pre-launch fieldwork, a greenfield venture holds three distinct streams of evidence: the analogue launch pool from Chapter 3, the macro and trade bounds from Chapter 4, and the calibrated choice surveys and expert Delphi distributions from the present chapter. [S08] [S28] [S31] Combining diverse forecasts is one of the most reliable ways to reduce variance and improve accuracy in forecasting practice, with over fifty years of empirical research showing that simple, transparent weighting schemes such as equal-weight or pre-fixed precision combinations consistently rival or beat complex optimised weights. [S20] However, forecast combination delivers its variance-reduction benefit only when the constituent forecasts contain independent information rather than echoes of the same underlying belief. [S08] [S20]

5.16 SHARED INFORMATION PATHS CREATE HIDDEN CORRELATION AMONG PRIORS

In a small industrial project, information contamination across evidence streams happens easily and silently unless it is policed. [S20] [S37] Consider three common failure modes:

- **Contaminated expert panels**: if the commercial team briefs the Delphi experts on the early results of the consumer conjoint survey or shows them the analogue launch table before Round 1, the experts' quantile estimates are no longer an independent judgemental source. [S20] [S0501] Combining the post-briefing Delphi distribution with the survey and analogue distributions counts the survey and analogue evidence twice, artificially narrowing the combined prediction interval. [S14] [S20]
- **Recycled reference classes**: if two wholesale distributors on the Delphi panel and the analyst who built the Chapter 3 analogue pool all base their estimates on the same single historical uniform contract or the same published CICAM or Camtext SA press announcement (such as CICAM's reported 5 percent domestic market share or Camtext SA's planned 12,000 tonnes a year capacity), their numbers share a single common error term. [R10] [R11] [S0501]
- **Selective post-hoc weighting**: if leadership inspects the analogue prior (say, a median of 800 units), the calibrated survey prior (1,018 units) and an optimistic distributor estimate (2,200 units), and then decides after seeing the numbers to assign a 70 percent weight to the distributor estimate so the business plan clears a financial hurdle, the combination rule has become a vehicle for confirmation bias. [S08] [S36] [S0503]

EXHIBIT 25 / PRE-LAUNCH SYNTHESIS SPECIFICATION

A prior register makes dependence and uncertainty visible

Evidence stream	Owner role	Sample or basis	Logged dependency	Pre-fixed weight
Analogue launch pool	Demand analyst	14 historical regional SKU launches	Shares category taxonomy with conjoint	0.35
Calibrated buyer conjoint	Field research lead	200 stratified category shoppers	Independent of Delphi; uses retail price tiers	0.35
Wholesale structured analogy	Commercial lead	8 stallholders rating 24 past lines	Blinded to survey; overlaps 3 pool SKUs	0.15
Anonymous Delphi panel	External facilitator	7 independent sector practitioners	Blinded to survey and internal targets	0.15

ILLUSTRATIVE [S08] [S20] [S29] [S31] [S36] Illustrative prior register structure showing how GFT records provenance, sample basis, shared information dependencies and pre-fixed combination weights across four pre-launch evidence streams before fieldwork readout.

To prevent double counting and post-hoc weight shopping, GFT requires every pre-launch evidence stream to be logged in a formal Prior Dependency Register before the individual readouts are combined, as illustrated in **Exhibit 25**, p. 93. [S20] [S36] For each stream, the register records the responsible owner role, the exact sample size or reference-class basis, any shared data or informational overlap with other streams, and the pre-fixed combination weight summing to 1.00 across eligible sources. [S20] [S36] In the illustrative configuration of **Exhibit 25**, p. 93, the analogue launch pool and the calibrated buyer conjoint survey each carry an equal weight of 0.35 because they rest on distinct empirical bases (historical launch curves versus prospective buyer choice calibration), while the wholesale structured-analogy interviews and the anonymous Delphi panel each carry 0.15 because both rely on practitioner memory and overlap partially with three historical lines in the analogue pool. [S20] [S28] [S31] While **Exhibit 25**, p. 93 makes assumptions and weights transparent for audit, the table itself does not prove that the chosen weights are optimal; their empirical skill can be scored only after pilot actuals arrive. [S14] [S20]

5.17 LINEAR OPINION POOLING PRESERVES TAIL DISPERSION ACROSS STREAMS

When combining probabilistic priors from multiple evidence streams, the mathematical aggregation rule matters as much as the weights. [S14] [S20] If an analyst averages the 10th percentiles, 50th percentiles and 90th percentiles across sources that disagree on the central level, or uses a naive precision-weighted normal conjugate update that assumes independent Gaussian errors, the resulting combined distribution becomes artificially narrow precisely when the sources strongly conflict. [S14] [S28] Instead, GFT combines pre-launch distributions via a finite mixture (linear opinion pool) of the component predictive densities, $F_{\text{prior}}(y) = \sum_{m=1}^{M} w_m \, F_m(y)$, where $\sum_{m=1}^{M} w_m = 1$ and $w_m \ge 0$. [S14] [S20]

Under linear opinion pooling, if the analogue launch pool ($w_1 = 0.35$) centres around an illustrative median of 800 units and the calibrated buyer survey ($w_2 = 0.35$) centres around 1,018 units while the wholesale analogy panel ($w_3 = 0.15$) warns of a downside cluster near 450 units, the combined prior distribution retains the variance between the streams as well as the variance within each stream. [S14] [S20] Disagreement among pre-launch sources widens the 80 percent and 95 percent predictive intervals rather than cancelling out into a false sense of certainty. [S14] [S16]

5.18

Pre-registration protects the prior from post hoc editing

The greatest threat to pre-launch primary research is not statistical sampling error; it is the human temptation to blurred boundaries between prediction and postdiction once the numbers are on the table. [S37] [S0503] Methodological research across the empirical sciences demonstrates that when analysis choices remain flexible after data collection, ordinary cognitive biases such as hindsight bias and confirmation bias lead analysts to select respondent exclusions, outlier rules, certainty cut-offs or subgroup splits that flatter a preferred commercial conclusion. [S0503] Defining the research questions, sampling frame, exclusion criteria and calibration formulas in a time-stamped, read-only document before observing the fieldwork outcomes, a discipline known as pre-registration, separates genuine predictive priors from post-hoc rationalisations. [S36] [S37] [S0503]

5.19 SEVEN MANDATORY FIELDS IN A PRE-REGISTERED RESEARCH PROTOCOL

Before a single questionnaire is printed or loaded onto a field tablet, the demand analyst and commercial owner sign and commit a pre-registration record into the venture's forecast repository. [S36] [S40] [S0503] That record locks seven operational specifications: 1. **Named decision and commitment horizon**: the exact production, assortment or channel decision the prior will inform (for example, first twelve-week cutting order for styles P01 to P04 across Douala retail and wholesale channels) and the date on which the commitment must be placed. [S08] [S36] 2. **Target population and stratification quotas**: the geographic locations, buying roles, channel definitions and minimum sample sizes per cell required for a readout to be accepted as decision-grade rather than exploratory. [S31] [S36] 3. **Deliverable attribute and price matrix**: the exact garment specifications, pack sizes and XAF price levels tested in the conjoint and Delphi instruments, signed off by production engineering as achievable within Phase 1 unit-cost and lead-time bounds. [S02] [S48] 4. **Respondent quality and exclusion rules**: objective, automated rules for dropping invalid interviews before looking at their product preferences, such as straight-lining across choice cards, completing a ten-screen choice task in under ninety seconds (illustrative speed-trap threshold), or failing an explicit attention-check screen. [S37] [S0503] 5. **Ex-post calibration formula and prior beta parameters**: the exact mapping from stated choice and follow-up certainty tiers to realisation probabilities ($\rho_{s,k}$), preventing anyone from raising the conversion multiplier after fieldwork to make a disappointing product look viable. [S31] [S0502] 6. **Blinding and dependency boundaries**: confirmation that Delphi panellists and wholesale analogy respondents will be elicited before they see

consumer survey tallies or internal business-plan targets, alongside the pre-fixed combination weights (w_m) in the Prior Dependency Register. [S20] [S29] 7. **Causal versus descriptive scope statement**: whenever the research compares competing commercial offers or price tiers, a clear statement of what is randomized within the survey versus what requires a controlled live pilot in Chapter 6 to establish a causal effect on actual purchases. [S32] [S0503]

> **WARNING**
>
> **Unresolved constraint: survey calibration cannot substitute for observed repeat purchase**
>
> Even a pre-registered, incentive-aligned choice study with a conservative realisation kernel informs only initial trial or first-order adoption over the stated launch window. [S30] [S31] [S0501] Stated-preference methods cannot measure how a garment's seams, dye fastness or collar shape perform after ten domestic washes in Cameroonian water and climate conditions, nor can they observe whether a wholesale trader's retail customers return for a second dozen pack six weeks later. [S01] [S06] If a planner treats a calibrated trial prior as a permanent steady-state weekly run rate, initial pipeline fill into boutiques or market stalls can be mistaken for sustained consumer sell-through. [S06] Resolving this constraint requires separating first-purchase trial from repeat-order velocity inside the live twelve-week pilot experiments of Chapter 6 and tracking batch-level return and reorder events in Chapter 7. [S06] [S32]

5.20 THE PRE-LAUNCH GATE HANDS AN AUDITABLE PRIOR TO THE PILOT TEAM

When fieldwork closes and the pre-registered script executes, the output is not a slide deck claiming that "72 percent of Cameroonians want our shirts." [S0502] The output is a versioned, machine-readable prior artefact containing the nine calibrated demand quantiles ($q_{0.005}$ to $q_{0.995}$) by style, channel and buyer segment, the full posterior intervals of the attribute utilities, the anonymised Delphi round distributions with their preserved dissenting rationales, and the completed Prior Dependency Register. [S14] [S16] [S29] If a style's calibrated 50th percentile demand sits below the factory's minimum economic batch size even before pilot testing, or if its viable price ceiling in the choice model lies below direct variable cost, the venture can redesign the garment, change the fabric specification or cancel the SKU before spending working capital on yarn and dye. [S08] [S31] For styles that clear the pre-launch gate, the explicit prior becomes the formal baseline hypothesis that the small-lot pilot experiments in Chapter 6 will test, update and gradually replace with owned purchase history. [S28] [S32]

Uncalibrated intent flatters the launch plan; a pre-registered prior protects the factory floor.

CHAPTER SOURCE NOTES

S31 INFORMS, Marketing Science; open manuscript from MIT. 2020.

R06 Cameroon Trade Hub. Publication date unrecorded; inherited source-text check.

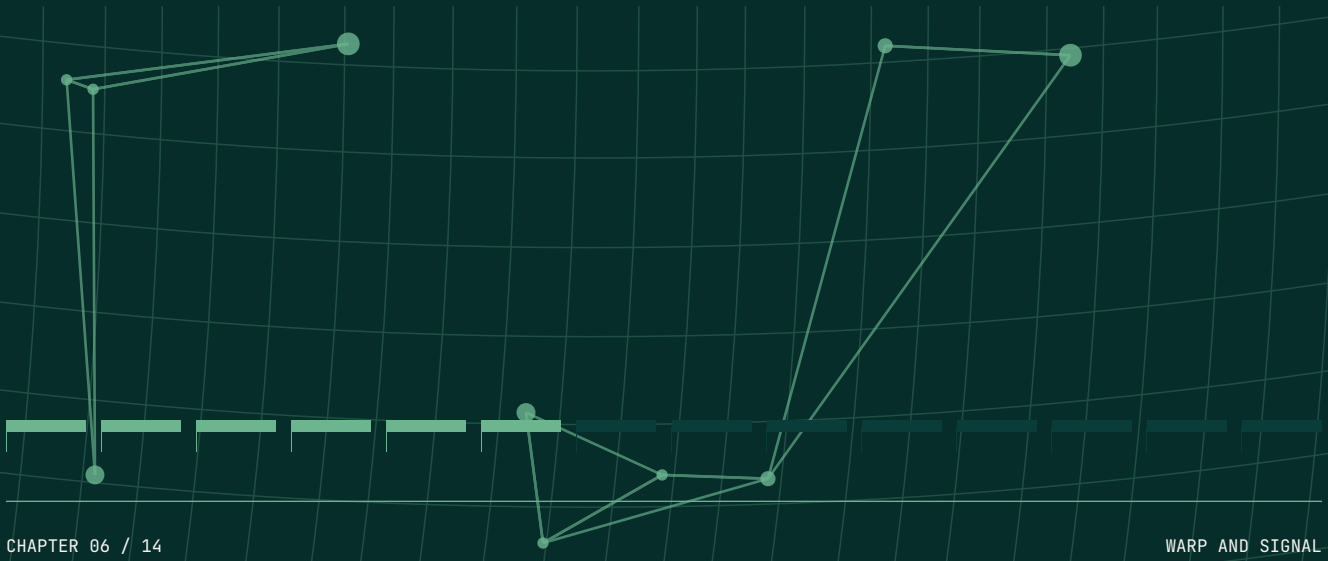
R26 Statista. Publication date unrecorded; inherited source-text check.
S08 International Journal of Forecasting, Elsevier; full text through arXiv. 2022.
S0501 OTexts, Monash University. 2026-10-30.
S0502 Centre for Evidence-Based Medicine, University of Oxford (Catalogue of Bias Collaboration). 2020.
S20 arXiv; journal version in International Journal of Forecasting, Elsevier. 2022-09-23.
S29 RAND Corporation. 2023-12-29.
S0503 Proceedings of the National Academy of Sciences of the United States of America, PubMed Central. 2018-03-13.
S01 InTech, International Journal of Engineering Business Management. 2013.
S06 Lancaster University Management School, Munich Personal RePEc Archive. 2018-11-19.
R14 WageIndicator. Publication date unrecorded; inherited source-text check.
R15 wage.is. Publication date unrecorded; inherited source-text check.
R16 WageIndicator. Publication date unrecorded; inherited source-text check.
S02 arXiv. 2019-06-27.
S48 Elsevier, Technological Forecasting and Social Change. 2006.
S37 Cell Press, Patterns; open full text from PubMed Central. 2023-08-04.
S03 arXiv. 2020-07-10.
S28 Journal of Industrial Ecology, Wiley. 2022-10-17.
S36 National Institute of Standards and Technology, U.S. Department of Commerce. 2023.
R32 Wise; XE. Publication date unrecorded; inherited source-text check.
S14 Journal of the American Statistical Association, American Statistical Association. 2007.
S16 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2022.
S32 Springer Nature, Data Mining and Knowledge Discovery. 2008-07-30.
S44 World Bank, World Development Indicators. 2026-10-30.
S46 World Bank, World Development Indicators. 2026-10-30.
S07 arXiv. 2026.
S33 Google for Developers. 2025-08-25.
R07 Ecofin Agency. Publication date unrecorded; inherited source-text check.
S39 Neural Information Processing Systems Foundation, NIPS 2015. 2015.
S40 Google Cloud Documentation. 2024-08-28.
R10 Ecofin Agency. Publication date unrecorded; inherited source-text check.
R11 Business in Cameroon; MMI News. Publication date unrecorded; inherited source-text check.
S30 Management Science, INFORMS; abstract through EconPapers. 1969.

06

PART II: EARN EVIDENCE BEFORE SCALE / LAUNCH EXPERIMENTS

Small pilots convert hypotheses into purchase evidence

The first pilot should be designed to reduce a named uncertainty, not merely to produce encouraging sales.



CHAPTER 06
DECISION TEST

The first pilot should be designed to reduce a named uncertainty, not merely to produce encouraging sales.

A small pilot earns its place by changing a decision that the current prior cannot settle. It should test a defined offer with identifiable customers, record whether each offer was available, preserve returns and substitutions, and end at a pre-agreed stop, adapt, repeat or scale gate. Encouraging sales alone are not enough. A release that sells through because stock was scarce, a promotion reached a self-selected audience, or one channel supplied every order may leave the launch forecast just as uncertain as before. Controlled assignment can separate an offer effect from selection where operations permit it, while sequential forecast versions show exactly how the new evidence changed the prior. [S32] [S28]

The pilot therefore has two products. One is a bounded commercial transaction: customers receive a real garment, price and service proposition. The other is an owned evidence record that a later model can use without guessing what happened. GFT would design both together. Lot size, release timing and channel exposure limit the production commitment, while availability logs and cell assignments protect the meaning of purchases. The first release is not a miniature national launch and does not need to settle every uncertainty. It needs to answer the next consequential question well enough to justify, delay or reverse the following commitment.

6.1

A pilot tests a decision rather than decorating the launch plan

The pilot charter begins with a decision sentence: if the evidence crosses a stated boundary, the venture will change a named commitment. That commitment may concern the opening assortment, the price architecture, the first channel, the size curve, the next production lot or the timing of a wider release. A charter that merely says “test the market” gives every result a plausible interpretation and none a required consequence. Controlled-experiment practice starts with changes whose influence on observable behaviour can be evaluated, which makes the decision and its measurement rule inseparable from the test. [S32]

A useful charter names the current prior before fieldwork starts. The prior might say that one silhouette should lead the opening range, that a particular price band is credible, or that direct sales should precede distributor expansion. The pilot then identifies the evidence that would weaken or strengthen that position. This prevents a favourable result from being used to support a different hypothesis after the event. It also exposes a pilot that cannot possibly change the decision, however many transactions it generates.

The unit of learning must match the uncertainty. A choice between two prices needs comparable offer cells and a common availability rule. A choice between channels needs separate channel readouts and clear definitions of shipment, order and final purchase. A choice about repeat purchasing needs an identifiable first-purchase cohort and enough observation time for a repeat opportunity. None of these questions is answered reliably by one aggregate sell-through figure. The charter therefore links each hypothesis to its target event, grain, exposure denominator and decision owner.

Scope should remain deliberately narrow. The first pilot should not simultaneously change garment specification, price, promotion, delivery promise and channel unless the decision concerns that whole bundle. A bundle test can be legitimate, but its result applies to the bundle and does not identify which component produced the response. Where the team needs component-level learning, it should hold other conditions stable or use a design whose cells can support the intended comparison. The write-up must not claim more resolution than the design contains.

A decision threshold also needs an evidence-quality threshold. A strong purchase result can still be unusable if assignment failed, listing time differed between cells, stock records are incomplete, or returns have not had time to arrive. The charter distinguishes the commercial readout from the permission to learn from it. This avoids rewarding a cell for favourable conditions that the future launch cannot reproduce. It also makes a failed measurement exercise visible before its numbers enter the forecast.

The pre-launch register records the hypothesis, commitment at risk, offer cells, assignment unit, eligible population, primary outcome, guardrails, availability rules, observation window, analysis owner and gate actions. Any amendment receives a date and reason. Changes made after outcomes are visible remain in the record, but they are labelled exploratory rather than silently becoming the original plan. This is the practical boundary between a forecasting experiment and launch communication.

> **GFT NOTE**
>
> GFT would refuse to call a release a forecasting pilot unless its charter can complete one sentence: “We will change this named commitment if this defined evidence crosses this agreed boundary.” The same charter must state what happens when the evidence is inconclusive or the measurement fails. If no commitment can change, the activity may still have commercial value, but its sales should not be presented as a controlled reduction in forecast uncertainty.

6.2

Randomisation separates offer effects from self-selected demand

Customers who choose a channel, location or launch date can differ before they see the offer. A pilot that lets the most enthusiastic buyers select the test cell can confuse that enthusiasm with the effect of the garment or price. Controlled experiments address this problem by assigning comparable units to different conditions, supporting a causal interpretation when the assignment and measurement are implemented correctly. [S32] For an apparel launch, the unit might be an eligible customer, a selling location, a distributor account or a defined time cell, depending on how the offer can actually be delivered.

The assignment unit must prevent the test from leaking across cells. A customer who can see two prices on consecutive visits may not provide independent responses to either. A shop whose staff promote the test garment more strongly is no longer delivering only the assigned product condition. Locations that share customers, stock or

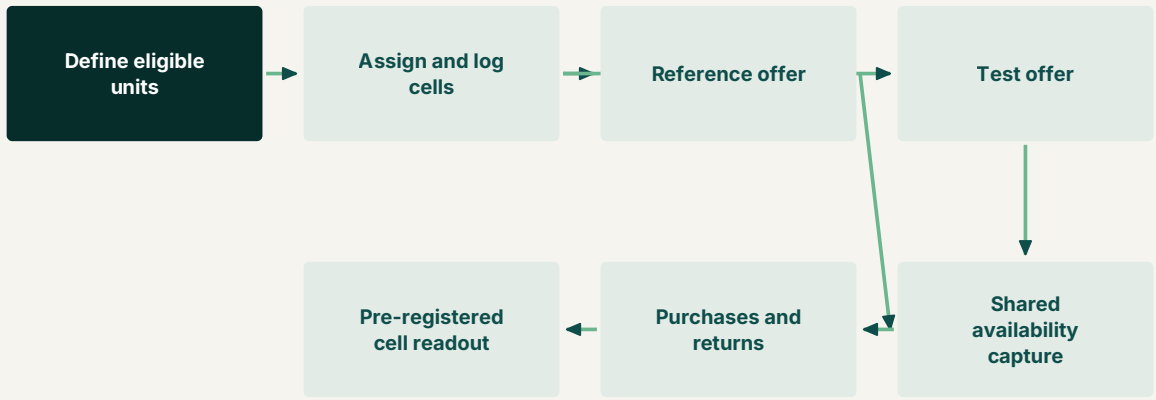
sales staff can contaminate one another. The operational design should choose a unit that keeps the assigned experience coherent, then record exceptions rather than assume that the assignment code solved every practical problem.

Randomisation does not replace comparability checks. Before outcomes are read, the team examines whether cells received the intended numbers and types of eligible units, whether listing and availability conditions were equivalent, and whether any external event affected one cell alone. These checks diagnose implementation; they do not justify deleting inconvenient units after seeing purchases. The primary readout follows assigned cells, while an exposed-only analysis can be secondary and clearly labelled. This keeps refusal, non-collection and missed exposure from quietly reshaping the comparison.

The measurement sequence in **Exhibit 26**, p. 100 makes those dependencies visible. It treats availability and returns as part of the experiment rather than fields to be appended after a winner has been declared. The flow is a GFT method design derived from controlled-experiment and censored-demand principles, not evidence that a particular launch used it. [S32] [S06] [S07]

EXHIBIT 26 / ASSIGNMENT AND EVIDENCE SEQUENCE

Comparable test cells keep the launch readout interpretable



DERIVED [S32] [S06] [S07] Derived GFT control sequence from controlled-experiment guidance and censored-demand research. It specifies a proposed method and does not claim an implemented pilot or a measured effect.

The primary outcome should be observable and consequential. Gross orders, fulfilled purchases, net purchases after cancellations, returns and repeat orders answer different questions. A pilot about initial purchase demand might use fulfilled net purchases per valid offer exposure, while retaining gross orders and returns beside it. A pilot about institutional interest should not call a non-binding expression of interest a sale. The denominator likewise needs care: visitors, eligible customers, account contacts and offer impressions represent different exposure opportunities.

WORKED EXAMPLE: COMPARE NET PURCHASE EVIDENCE, NOT ORDERS ALONE

Assume an illustrative randomised pilot with 500 valid offer exposures in each of two cells. Cell A records 60 gross orders and 6 cancellations before the readout, leaving 54 net purchases. Cell B records 45 gross orders and 5 cancellations, leaving 40 net purchases. The illustrative net purchase rates are therefore 54 ÷ 500 = 10.8 percent for A and 40 ÷ 500 = 8.0 percent for B. The arithmetic difference is 2.8 percentage points. It describes this constructed pilot only. It is not a universal price effect, a forecast for Cameroon, or proof that the difference will persist at scale. The review would still check assignment, availability, return maturity and independent support before allowing the result to alter the prior.

Exhibit 27, p. 101 preserves that worked calculation in a compact readout. The important comparison is not merely A against B. It is the movement from a common exposure denominator to gross orders, cancellations and net purchases, with the same definitions in both cells. Counts and rates remain illustrative and do not carry a significance claim.

EXHIBIT 27 / ILLUSTRATIVE COUNTS AND PERCENT OF VALID EXPOSURES

A common denominator keeps the pilot comparison honest

Measure	Cell A	Cell B	Readout role
Valid offer exposures	500	500	Common denominator
Gross orders	60	45	Initial response
Cancellations	6	5	Outcome adjustment
Net purchases	54	40	Gross less cancellations
Net purchase rate	10.8	8	Percent of exposures

ILLUSTRATIVE [S32] Illustrative worked example. A net purchases = 60 - 6 = 54; B net purchases = 45 - 5 = 40. Rates = 54 / 500 = 10.8 percent and 40 / 500 = 8.0 percent. Difference = 2.8 percentage points. No confidence interval, causal effect estimate or scale forecast is claimed.

Small cells require restraint even when the arithmetic is exact. A handful of large accounts, repeated orders from one buyer, or an event affecting one location can dominate the result. The readout should show the number of independent buying units beside transactions and identify concentration without turning every repeated garment into independent evidence. Where the design has too few assignment units for a stable comparison, the appropriate conclusion is inconclusive. The team can retain the observations as descriptive history and run another test, rather than promote an exploratory difference into a causal law.

Where randomisation is impossible, the charter records how units entered each condition. A distributor may have selected its own assortment, a location may have been chosen because of expected demand, or early buyers may have received a different price from later buyers. Those results can still inform a forecast, but the language changes. The report describes observed purchases under selected conditions, exam-

ines plausible selection mechanisms and avoids attributing the whole difference to the offer. A transparent non-randomised pilot is more useful than a supposedly controlled one whose assignment failed.

6.3

Limited releases identify the ramp while containing exposure

A limited release allows the venture to buy information with a bounded production commitment. The operative word is bounded. The lot, selling window, eligible channels and replenishment conditions are fixed before the result is known, so an enthusiastic first response cannot automatically trigger an unplanned expansion inside the same evidence window. This separates learning from the pressure to make the launch look successful. It also preserves a meaningful distinction between the initial release and the later ramp that the pilot is meant to inform.

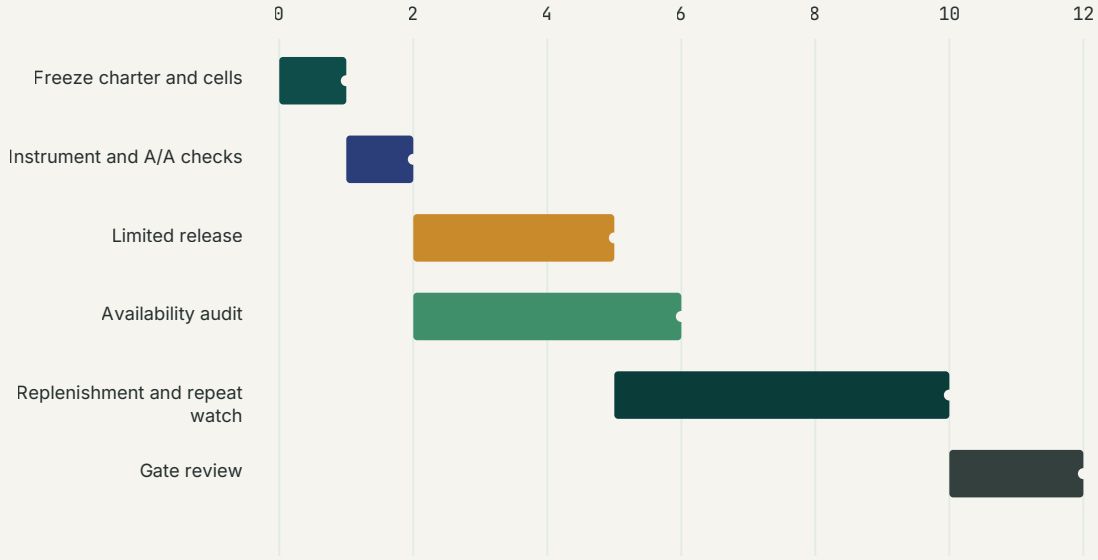
The release should be large enough to exercise the real buying and fulfilment process, not merely to display samples. Customers need a genuine offer, a defined price, a delivery or collection mechanism and a route for returns. Production needs to execute the intended specification rather than a handmade substitute whose quality and cost cannot be repeated. At the same time, the pilot need not cover every size, region or channel. Uncovered conditions remain explicit transfer questions for a later release.

Release timing is part of the treatment. A lot released in several planned tranches can test whether early purchases persist after initial novelty, while a single lot can address first-window response when production flexibility is limited. The charter should not switch between those designs because early sales are disappointing. If commercial safety or customer harm requires interruption, the human stop rule takes precedence and the analysis records the deviation. The statistical plan never obliges the team to continue an unsafe or misleading offer.

The illustrative calendar in **Exhibit 28**, p. 103 sequences one possible first release over twelve weeks. It allows instrument checks before sale, availability audit during sale, and a later window for replenishment and repeat observation. The durations are planning assumptions, not a universal minimum, a GFT client commitment or evidence that twelve weeks can resolve every apparel question.

EXHIBIT 28 / ILLUSTRATIVE WEEKS FROM CHARTER APPROVAL

A twelve-week pilot can sequence learning without rushing scale



ILLUSTRATIVE [S32] [S06] Illustrative planning calendar. Tasks and durations are assumed to show sequencing and overlap. They are not observed milestones, a critical-path calculation or a required duration for an apparel launch.

The calendar protects different learning clocks. Initial conversion can be read before repeat purchase, but a gate that requires repeat evidence cannot be cleared early merely because the first lot sold through. Returns and cancellations also need their own maturity rules. A readout taken before customers have had a realistic opportunity to return or repeat should be labelled provisional. Later observations arrive as new evidence versions, not as silent edits to the first conclusion.

A small lot creates a further risk: scarcity can become the story. Selling out may indicate strong demand, insufficient stock, delayed replenishment or some combination of all three. The release plan therefore separates the demand test from the replenishment test. During the demand window, availability should be as comparable as the design requires. During a later replenishment window, the team can test response time and continuing purchases without pretending that the two phases share identical supply conditions.

6.4

Availability must be designed into the pilot

Low sales are interpretable only when the customer could have bought. Retail research shows that demand estimates based on sales alone can be negatively biased when inventory constrains observation, and that stockouts can shift purchases towards substitutes. [S06] A pilot built to create purchase history must therefore measure the opportunity to purchase with the same care as the purchase itself. Otherwise, the model learns the venture's stock and listing failures as if they were customer preferences.

Availability begins at the actionable product and location grain. The record identifies when each variant was listed, when it could be fulfilled, when available inventory reached zero, and when listing or fulfilment resumed. A style can appear available while the requested size is not. An online listing can remain visible while delivery is impossible. A distributor can report a shipment while the garment is not yet offered to the final customer. The pilot dictionary must choose the availability event that matches the demand target and preserve the others as context.

Stockout timestamps need verification, not just a system flag. Inventory records can be late or inaccurate, so the protocol includes physical or transactional checks proportionate to the pilot's size. Staff record suspected gaps, customer requests that could not be fulfilled, manual overrides and unexpected delisting. These observations do not automatically become demand quantities. They tell the analyst where the sales series is censored and where a zero cannot support the interpretation of no interest. Stockout-annotated retail research makes that distinction central to latent-demand recovery, although its fresh-retail setting and hourly granularity should not be imported uncritically into apparel. [S07]

Substitutes belong in the same observation window. When the requested variant is unavailable, a customer may buy another size, colour or product, delay the purchase or leave. The pilot should retain which alternatives were visible and available, together with their outcomes. It should not assume that all unmet demand transferred, vanished or waited. Those are different behaviours with different consequences for the focal product and the substitute. The analysis can test a correction only if the collected fields support its assumptions.

Returns and cancellations complete the availability record. A fulfilled order that is later returned is not the same evidence as a retained purchase, and a cancelled order may never have entered physical demand fulfilment. The readout keeps gross order, fulfilment, cancellation, return and net purchase events separate, with their dates and reasons. It also marks outcomes still immature at the information cut-off. Deleting the original transaction after a return would destroy the sequence needed to understand both product response and data quality.

The availability rule can require a minimum valid exposure condition without inventing a universal threshold. The charter sets that condition for each channel and cell before outcomes are read. Intervals failing it are excluded from the primary availability-valid analysis, retained in a separate contaminated record and included in sensitivity reporting. Exclusion is not permission to remove weak sales selectively. It applies to defined measurement failures under a frozen rule, with counts and consequences visible.

A further distinction separates intended scarcity from operational failure. A deliberately small lot can constrain exposure by design, but unplanned early exhaustion can still prevent the intended comparison. Likewise, a promotion that drives an unexpected stockout changes both demand and observation. The pilot report should describe the mechanism rather than label every sold-out release a success. Where censoring is material, the forecast may retain a broad range, present alternative correction scenarios or await another release. More arithmetic cannot replace missing demand observations.

6.5

Channel and geography readouts remain separate

A pooled launch total can hide the only result that matters for the next commitment. Direct retail, online sales, distributor accounts and institutional buying create different purchase events, exposure denominators, inventory responsibilities and return paths. Multichannel fashion research evaluates physical and non-store channels separately and calls for region- and channel-specific testing of candidate predictors. [S04] The pilot should preserve that separation even when the venture ultimately wants one production forecast.

Channel separation starts with the event boundary. A distributor's opening shipment is not consumer sell-through, and an institutional award is not the same as wearer uptake. Direct customer purchases can be counted against valid offer exposures, while a distributor readout may focus on qualified accounts, orders, subsequent replenishment and documented downstream availability. These measures can coexist without being added as if they were identical demand events. The forecast contract assigns each one a role and an aggregation rule.

Geography adds a second boundary. Selling locations can differ in customer mix, calendar, access, substitutes, delivery reliability and staff execution. The pilot register records which of those conditions were observed and which are merely suspected. It does not invent a regional adjustment factor from a successful location. A result from one selling context becomes evidence for another only through an explicit transfer assumption, normally accompanied by wider uncertainty or another test.

The readout should be hierarchical rather than collapsed. Product variants roll into styles, selling locations roll into markets, and channels roll into the venture total only after their component evidence has been inspected. The M5 retail evaluation was explicitly organised around hierarchical unit sales and required forecasts at the lowest cross-sectional level to be summed to higher levels, which provides a useful evaluation principle rather than an apparel effect size. [S11] For the pilot, component and total views should reconcile while retaining their different sample support.

A small pilot will have sparse cells. The answer is not to pool them until every apparent problem disappears. Each readout carries its independent assignment units, valid exposures, availability coverage and outcome maturity. Cells below the charter's support condition are labelled preliminary or inconclusive. Partial pooling may later borrow strength across comparable groups, but that modelling choice belongs after the raw channel and geography evidence has been preserved. It must not erase a failed channel by hiding it inside a favourable total.

The practical output is a channel and geography transfer register. For every cell it records the offer, event definition, assignment unit, selling window, availability condition, net outcome, concentration, returns maturity and permitted next use. One cell may justify another direct-sale lot but no distributor expansion. Another may justify repeating the same test in a second location before any wider commitment. The pilot's value lies in these specific permissions and prohibitions, not in one launch-wide approval.

This separation also protects planning conversations. Production may need a total range, but commercial owners need to know which channels and places created it. Reviewing both levels reveals whether the total depends on one account, one city, one promotion or one size curve. The team can then direct the next limited release towards the uncertainty that blocks expansion rather than repeat the easiest selling condition and call the accumulated transactions new evidence.

6.6

Sequential updates preserve the prior and every revision

The pilot should update a forecast, not overwrite its history. Before launch, the venture holds a versioned prior built from analogues, market evidence, choice research and structured judgement. Each pilot readout then creates a new forecast version with an information cut-off, the evidence admitted, the update rule and the resulting range. Hierarchical Bayesian modelling offers one candidate for combining shared and local evidence, but its supporting demand application concerns copper rather than apparel, so the method and its assumptions still require launch-level validation. [S28]

An update register makes the transition inspectable. It stores the prior identifier, analogue corpus and research versions, pilot cell definitions, valid exposure counts, availability treatment, cancellations and returns cut-off, support by channel, model or weighting rule, posterior or revised quantiles, reviewer and approval status. A planner can then reconstruct why the central estimate moved and which new observations caused the change. Without that chain, each weekly spreadsheet becomes a new story with no way to test whether learning actually occurred.

Evidence weight should not equal transaction count. Repeated purchases from one account, several sizes bought in one order, and transactions sharing a promotion can carry dependent information. Conversely, a smaller number of well-assigned, availability-valid purchases may answer the named uncertainty more directly. The update rule therefore considers assignment quality, independence, censoring, offer stability and channel relevance. These are not decorative metadata fields. They determine whether the new history deserves to displace the prior.

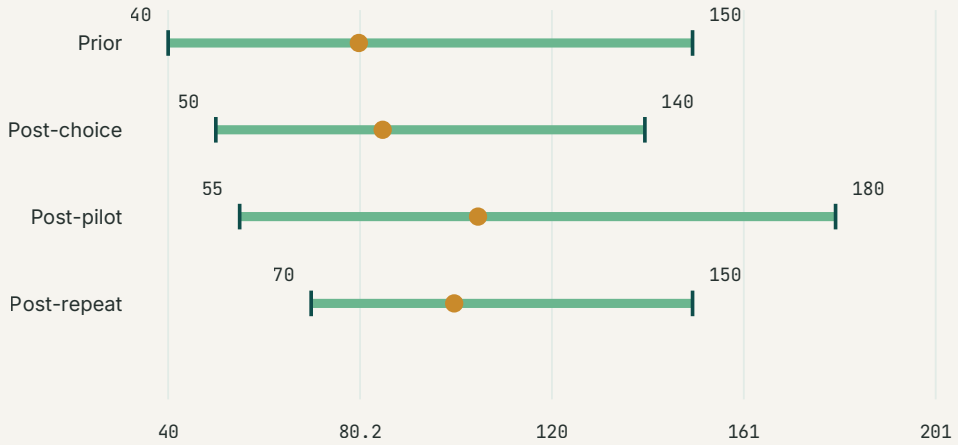
Double counting is a related danger. If the same buyers contributed to choice research and the pilot, or if an expert revised the prior after seeing early sales, those inputs are no longer independent confirmations. The evidence register records such dependencies. A combined forecast should not present them as separate votes for the same outcome. Where dependence is unclear, the review can compare a conservative update with the proposed one and retain additional uncertainty rather than claim a precise solution.

More data can legitimately widen the forecast. The pilot may reveal that the target channel behaves unlike the analogue channel, that the size curve is unstable, or that availability failures are more consequential than expected. A revision that responds by widening the range is not a failure of learning. It is evidence that the previous confidence was not earned. The update protocol should permit restoration of an earlier prior when new records fail validation, rather than forcing every version to move forward with narrower bounds.

Exhibit 29, p. 107 illustrates that non-monotonic path. The constructed post-pilot range widens when conditions depart from plan, then narrows only after repeat evidence becomes usable. Values are indexed illustrations, not garment sales, fitted posterior quantiles or calibrated coverage statements. Their purpose is to show the reporting form: prior, evidence stage and residual doubt remain visible together.

EXHIBIT 29 / ILLUSTRATIVE INDEXED DEMAND PER LAUNCH WINDOW

Forecast revisions should show both learning and residual doubt



ILLUSTRATIVE [S28] [S06] Illustrative bounds and midpoints. Prior = 40, 80, 150; post-choice = 50, 85, 140; post-pilot = 55, 105, 180 after an assumed availability and transfer exception; post-repeat = 70, 100, 150 after assumed valid repeat evidence. These are constructed ranges, not statistical intervals, observed demand or guaranteed uncertainty reductions.

The midpoint in that illustration does not tell the whole story. A revised central value can rise while the upper and lower bounds both move outwards, signalling disagreement between the first purchases and the transfer assumptions. Reviewers should therefore inspect the full distribution and the evidence exceptions, not just the direction of the point forecast. They should also compare the update with the frozen pre-launch rule. If the rule is changed after seeing the result, the new version is a new modelling proposal that requires separate evaluation.

Sequential updating needs disciplined timing. The charter states when initial purchase, return and repeat windows close, and which cut-off applies to each forecast release. Late-arriving events enter the next version with their original event times and their arrival times. This preserves both the commercial sequence and the information sequence. A later backtest can then reproduce what the venture actually knew at each decision instead of using a cleaned final history that was unavailable at the time.

Finally, the update pack should be readable by commercial and production owners without reference to model code. It states what was expected, what was observed, which observations were excluded and why, how the range changed, and which commitment the revision supports. Technical reviewers retain the detailed model diagnostics. The planning version keeps the decision chain visible so that a forecast revision can be challenged before it becomes a material or capacity commitment.

6.7

Stop, adapt and scale rules are fixed before results arrive

A pilot needs more than a success threshold. It needs a complete set of authorised outcomes before anyone knows whether the preferred offer is winning. The charter should distinguish stop, adapt, repeat and scale actions, together with the evidence-quality and commercial conditions attached to each. Human owners set the metrics and thresholds in the context of the actual decision rather than importing a universal benchmark. NIST's risk framework specifically places human judgement behind the choice of trustworthiness metrics and their precise thresholds. [S36]

Stop rules protect customers, operations and evidence. A misleading offer, an unresolved fulfilment fault, compromised assignment or a material safety concern can justify immediate interruption even when purchases are strong. The stop owner does not need to wait for statistical maturity. The event and its timing are recorded so that later analysis can distinguish commercial response from the conditions under which the test ended. Continuing a faulty experiment merely to complete a sample would subordinate the operation to the measurement plan.

Adapt rules allow a defined correction without turning every disappointing result into a new hypothesis. An adaptation may repair an availability failure, correct a listing error, or repeat a cell whose assignment mechanism failed. It should not quietly change the price, audience or product after weak sales and then count the later response as the original test. Material changes create a new pilot version and a fresh evidence window, while the original outcome remains in the register.

Repeat rules address inconclusive evidence. The trigger may be insufficient independent buying units, excessive censoring, immature returns, conflicting channel readouts or a difference too unstable to support the intended commitment. Repetition should have a purpose: more comparable locations, cleaner availability, a longer repeat-purchase window or a revised lot condition. Simply repeating the easiest cell until it produces an encouraging number would make the gate a selection device rather than a learning rule.

Scale rules join statistical evidence to operating consequences. The review asks whether the observed offer can be produced and fulfilled repeatedly, whether availability remained valid, whether returns and repeat behaviour are sufficiently mature, and whether the revised forecast changes the plan for the better. Where a forecast candidate is also under review, the gate retains its pre-specified accuracy and calibration conditions beside the operational and commercial tests. Practical machine-learning guidance warns that an improvement in model loss is not acceptable if system performance deteriorates. [S33] For a pilot, the equivalent rule is that an attractive purchase metric cannot compensate for a production or fulfilment condition that would fail at the proposed scale.

METHOD

Use a gate sheet with four evidence blocks. The assignment block records eligible units, cell integrity and independent support. The availability block records listing, stock, substitutes, cancellations, returns and outcome maturity. The forecast block records the frozen prior, update rule, revised range and sensitivity cases. The decision block records the production and service conditions, the authorised action, the accountable owner and the next reversal trigger. No block can be borrowed from another to conceal missing evidence.

Scale itself should remain staged. Clearing the pilot gate authorises only the commitment named in the charter, such as another lot, an additional comparable location or a controlled channel extension. It does not automatically authorise every product, geography and production horizon. The next release carries its own monitoring and reversal conditions. If the transfer assumptions weaken, the venture can return to the approved prior, repeat a narrower test or stop the expansion without rewriting the meaning of the earlier evidence.

A complete pilot handover therefore contains the charter, assignment and exposure logs, availability audit, cell readouts, update versions, gate decision and next-release conditions. It distinguishes observed purchase evidence from illustrative scenarios and from the owner's judgement. It also states what the pilot did not test. Those limits keep the first owned history useful after the launch team changes and prevent a small, context-specific success from becoming an unsupported general rule.

A pilot earns scale only when its evidence can change the plan and its limits can stop it.

CHAPTER SOURCE NOTES

S32 Springer Nature, Data Mining and Knowledge Discovery. 2008-07-30.

S28 Journal of Industrial Ecology, Wiley. 2022-10-17.

S06 Lancaster University Management School, Munich Personal RePEc Archive. 2018-11-19.

S07 arXiv. 2026.

S04 MDPI, Processes. 2021-09-03.

S11 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2022.

S36 National Institute of Standards and Technology, U.S. Department of Commerce. 2023.

S33 Google for Developers. 2025-08-25.

07

PART II: EARN EVIDENCE BEFORE SCALE / DEMAND DATA

Clean launch records protect the history the model will learn

The first sales history becomes valuable when availability, product context and time are captured with the transaction.

CHAPTER 07
DECISION TEST

The first sales history becomes valuable when availability, product context and time are captured with the transaction.

A clean launch record preserves the conditions under which a sale occurred, not merely the fact that a transaction was posted. It joins stable event identities to availability intervals, the exact product offer, fulfilment and return events, and the information that was available at each decision time. This is the minimum evidence structure for learning demand. A daily sales extract can describe cash or units leaving the business, but it cannot distinguish weak demand from an unavailable size, an old photograph from the offer actually seen, or a revised external series from the value known at launch. Retail research shows why sales-only demand estimates can be biased, while leakage research shows why collection and preprocessing decisions can invalidate apparently strong models. [S06] [S37]

The design consequence is direct: capture first, correct cautiously and model later. Corrections should remain reversible interpretations of preserved events. They should never replace the raw history with a cleaned number whose origin cannot be inspected. GFT would treat this record as the demand side of the production data spine in its Cameroon textile programme, where the inherited sector baseline shows roughly 95 percent of cotton fibre exported raw against 5 percent processed locally alongside a national target of 50 percent local processing by 2030. [R06] [R07] A new apparel venture operating in that environment cannot rely on a large engineering department or decades of archival point-of-sale logs. As machine-learning systems research notes, practical expense and failure often stem not from a lack of new statistical models, but from missing workflows for data preparation, labelling, productionisation and monitoring. [S0702] The objective is therefore a small, testable capture contract that lets every later forecast replay what the venture knew, what customers could buy and what subsequently changed.

7.1

Every sales event needs a stable identity and timestamp

A learnable ledger begins with an event identifier that never changes when a report is rerun. Order creation, payment confirmation, dispatch, delivery, cancellation and return are separate event types. Each carries its own event time, the time the business received the record, and links to the order line, product variant, selling location and channel. This preserves the sequence that a flattened daily total destroys. It also lets the forecast contract select fulfilled purchases, net purchases or another target without pretending that those measures are interchangeable. [S37]

Two clocks matter in every commercial log. Event time states when something happened in the commercial process. Arrival time states when the data became available to the forecasting system. A return can belong commercially to an earlier order while becoming known only after a forecast has been issued. Both facts should coexist. If arrival overwrites event time, a replay can accidentally use knowledge that did not exist at the forecast origin. If event time is ignored, the ledger cannot explain the operating sequence. Temporal and preprocessing leakage are not abstract research concerns here: leakage can arise from collection, sampling and preprocessing choices, so the capture schema is part of model validity. [S37]

Stable identities also need stable relationships across the order lifecycle. An order header can identify the commercial commitment, but each line needs its own variant, quantity and offer reference. A split dispatch should point to the original line rather than create an unrelated sale. A return should point to the dispatched line where the operation permits it. A substituted item should retain both the requested and delivered variants. These links allow later analysis to follow demand through fulfilment without double counting the order, the shipment and the payment as separate acts of demand.

Validation should reject impossible sequences from the released dataset while retaining the offending records in quarantine. Payment after cancellation may be legitimate in an unusual process, so the rule should route an exception for review rather than automatically erase it. A dispatch before an order, a return before a dispatch, or a quantity exceeding the original line requires a declared resolution rule. Late corrections arrive as superseding versions with their own arrival times, while the previous version remains available for audit and for any forecast that legitimately used it. [S40]

Exhibit 30, p. 113 sets out the resulting record boundary. It is a GFT design derived from attribute, availability and leakage requirements, not a claim that this architecture has already produced a forecast result. The important feature is the connection between layers: an event without its offer and availability context is not yet a reliable training observation. [S02] [S06] [S37]

EXHIBIT 30 / RECORD RESPONSIBILITIES

A learnable sales record connects six event and context layers



DERIVED [S02] [S06] [S37] Derived GFT design from new-item attribute requirements, censored-demand research and leakage controls. It specifies capture responsibilities, not an implemented client system or a measured forecasting benefit.

The identity layer should be straightforward to operate across both digital and physical channels. Source systems need deterministic export keys, and manual wholesale or showroom channels need a controlled identifier issued before the event form is completed. Reimports should recognise an existing event instead of duplicating it. A changed source identifier requires a crosswalk, not a new commercial history. These controls are more valuable than an elaborate feature store whose underlying events cannot be reconciled. [S39] [S0702]

GFT NOTE

GFT would make four fields non-negotiable before any launch model trains: persistent event ID, event time, arrival time and source-record version. [S37] Product, channel and offer links sit beside them. If any is absent, the record may remain useful for accounting or operations, but it receives a restricted forecasting status. The team repairs the capture path rather than filling the gap with an undocumented assumption.

7.2

Availability and stockout timestamps are demand data

A zero sales row has at least three distinct meanings in an apparel launch. It can be an observed zero while a valid offer was available, a structural zero when no offer existed, or a missing observation when the channel failed to report. Only the first supports a direct interpretation of no purchase demand. Intermittent-demand research defines sporadic series by periods showing no demand at all, and subsequent work introduces bias corrections for intermittent estimators. [S18] [S19] If a capture system feeds stockout zeros or unlisted periods into Croston-type or Syntetos-Boylan estimators as if they were genuine no-demand intervals, the estimated inter-demand interval is upward biased from the start. Retail forecasting research similarly identifies the negative bias that follows when sales alone stand in for demand under stockouts, and stockout-annotated research makes the same collapse of the sales-equals-demand assumption explicit, although its fresh-retail evidence should not be treated as an apparel effect size. [S06] [S07]

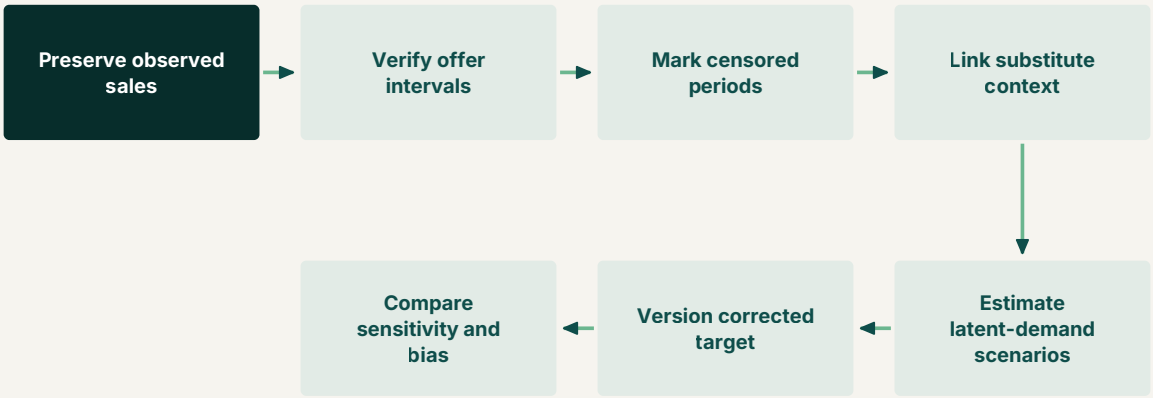
Availability therefore needs intervals, not a weekly flag added after the sales total is known. The ledger records when a variant became visible, when it could actually be fulfilled, when inventory ceased to support purchase, and when a valid offer resumed. The relevant condition depends on the channel. A visible online page with no deliverable stock is not an available offer. A distributor shipment waiting in a back room is not necessarily customer availability. A style-level stock balance cannot prove that every size was purchasable, because apparel buyers purchase specific sizes rather than style averages. [S02] [S06]

Each interval should carry its evidence source and verification status. A system inventory transition may be the primary signal, while a physical count, failed basket, staff exception or customer request supplies a check. Conflicting signals remain visible until resolved. The purpose is not to manufacture certainty about every stock movement. It is to identify where the sales observation is censored and where a zero is genuinely informative. [S07]

Exhibit 31, p. 115 places correction after that capture sequence. The corrected target is an analytical output with a method and version, not a replacement for observed sales. This distinction follows from the cited censoring research and prevents a later analyst from mistaking an imputed value for a directly observed transaction. [S06] [S07]

EXHIBIT 31 / EVIDENCE AND DECISION SEQUENCE

Stockout correction starts before model training



DERIVED [S06] [S07] Derived GFT method from retail censoring and stockout-annotation research. Latent-demand estimation is conditional on recorded availability and substitution assumptions. No universal correction formula or recovered-demand accuracy is claimed.

Substitution records make the observation richer but not automatically complete. A customer whose first choice is unavailable may choose another size, colour or garment, postpone the purchase or leave. The ledger should record which alternatives were visible and available, any substitution offer and the eventual outcome. It should not assume that every substitute sale contains one unit of transferred focal demand. Research on retail stockouts warns that substitute sales can themselves overstate the substitute's underlying demand, which is why both focal and alternative histories need context. [S06]

> **WORKED EXAMPLE: A STOCKOUT CREATES A RANGE, NOT A RECOVERED FACT**
>
> Assume an illustrative offer is valid for a ten-hour selling window. It records 80 fulfilled units in the first four hours, then remains unavailable for six hours, with no returns, cancellations or substitute purchases in this constructed case. Observed sales are 80 units. A no-further-demand scenario leaves latent demand at 80 units. A constant-rate scenario calculates 80 ÷ 4 = 20 units an hour, then 20 × 10 = 200 units across the full window. The resulting 80 to 200 unit range demonstrates sensitivity to an assumption. It is not a statistical interval, an observed launch result or a recommended correction for an apparel product.

Exhibit 32, p. 116 preserves the illustrative calculation and its assumptions. The correct operational response is to improve the next availability record and compare defensible scenarios, not to select the illustrative 200-unit figure because it makes demand look stronger. [S06] [S07] Real launch data would also need to address intraday traffic patterns, returns, substitution and the reliability of the stockout timestamp.

EXHIBIT 32 / ILLUSTRATIVE UNITS IN ONE SELLING WINDOW

A censored sales record supports scenarios rather than a unique recovery

Record or scenario	Calculation	Units
Observed fulfilled sales	Preserved events	80
No-further-demand scenario	80 + 0	80
Constant-rate scenario	(80 / 4) x 10	200

ILLUSTRATIVE [S06] [S07] Illustrative arithmetic from the adjacent case. Observed sales = 80. Pre-stockout rate = 80 / 4 = 20 units per hour. Constant-rate scenario = 20 × 10 = 200. No-further-demand scenario = 80. Assumptions are constructed and no recovery accuracy is claimed.

Availability coverage should influence admissibility before model fitting. Intervals with unresolved listing or stock evidence can be excluded from a primary availability-valid training set under a frozen rule, retained for sensitivity analysis and reported as a share of the launch history. Documenting these collection rules and exclusions in a structured dataset datasheet ensures that downstream modellers understand why specific intervals are marked censored or restricted. [S0701] This is not permission to delete inconvenient low-sales periods. It is a declared measurement-quality decision whose effect on the forecast must remain inspectable.

7.3

Product attributes and images must be versioned with the SKU

An attribute model can use only the product description that the capture system preserves. New-item fashion research infers demand from clothing, footwear and merchandising attributes learned from other products, so inconsistent category names, unversioned materials or ambiguous colour labels weaken the input before modelling begins. [S02] The launch record therefore separates the enduring SKU identity from the descriptions that can change. Category, silhouette, material, colour, size, occasion and merchandising group each have a controlled value, effective interval and source.

The taxonomy should be small enough for the launch team to apply consistently. A useful category is one that different people and channels can assign in the same way, with an explicit unknown value when evidence is absent. Splitting and merging categories require mapping tables and new taxonomy versions. Backfilling a new classification across old products is permissible only as a labelled transformation. It must not silently change the attributes used by an earlier forecast. [S37]

Variant identity deserves particular care. A style can remain commercially stable while its size curve, colour set or specification changes. The SKU key should identify the sellable variant at the grain used for stock and purchase, while a separate style key supports aggregation. Product codes reused after discontinuation would corrupt both analogue selection and availability analysis. Where channels use different codes, the crosswalk records the mapping and its validity period rather than relying on a spreadsheet maintained after launch.

Images follow the same effective-dating rule. Visual-similarity research transfers corrected demand from clusters of historical products to new items, which makes the link between image and historical offer consequential. [S05] The record should identify the image set customers could actually see, its publication interval, channel and associated offer version. A later studio photograph may describe the garment more clearly, but it is not the visual signal that accompanied the first purchases.

Image versioning does not require a large content platform at the outset. A controlled file reference, checksum, capture date, publication date and SKU link can preserve the evidence. Cropped marketplace images, swatches, lifestyle photographs and technical drawings should not be treated as interchangeable inputs without a declared rule. Where an offer used several images, the record retains the set and its sequence rather than selecting one after the sales result is known. [S05] [S37]

> **METHOD**
>
> Run an attribute and image audit before every pilot release. Two reviewers independently apply the taxonomy to a small sample and resolve disagreements in the dictionary. Reconcile every transaction line to one SKU and one effective offer version. [S02] Reconcile every published offer to the image set and price condition active at that time. [S03] [S05] Unknown remains a valid captured state; guessed values do not become model inputs.

This discipline supports later analogue work without repeating it. The capture layer supplies trustworthy descriptors and visual provenance. The separate modelling decision determines which attributes transfer, how similarity is calculated and whether the reference class is relevant. Clean records cannot prove that visual resemblance predicts demand. They prevent a model from learning from descriptions and images that customers never encountered. [S05]

7.4

Price and promotion fields require an effective customer offer

List price is a commercial intention, not necessarily the price condition that produced a purchase. The event line should preserve list price, paid price, explicit discount, bundle membership, campaign reference and delivery terms as separate fields. Fashion forecasting research includes selling price alongside age, attributes, temporal and store features, which makes a reliable price history a candidate model input rather than administrative detail. [S03] The evidence comes from an established retailer, so no coefficient or effect should be imported into the new venture.

Paid price belongs at the order-line grain where possible. A basket-level discount needs a documented allocation method if analysis requires variant prices. The rule should state how rounding and returned quantities are treated, and the allocated components should reconcile to the amount paid. A bundle should retain its constituent products and the bundle price condition. Otherwise, a lower apparent garment price can be confused with a change in demand for that garment alone.

Promotion capture extends beyond the discount percentage. The offer register records eligibility, audience, channel, placement, start and end times, and any coupon or membership requirement. A campaign that only some customers saw cannot be represented by a flag applied to every sale in the same week. Exposure and redemption remain distinct. A transaction can carry the promotion it used, while the offer ledger carries the conditions under which it was available. [S03]

Delivery and collection terms form part of the effective proposition. A customer may respond to the combined garment, fee, timing and collection condition rather than to the ticket price alone. The record should therefore retain the terms attached to the offer and the terms actually applied at fulfilment. Changes after purchase are later events, not reasons to rewrite the original selling condition.

Calendar and event variables need the same provenance. Multichannel fashion research recommends testing holidays, commercially relevant events, extraordinary events and economic conditions for specific regions and channels. [S04] That is a research agenda, not permission to assume that every named day affects every launch. The capture register identifies the calendar source, geography and event classification, then leaves incremental value to chronological validation.

Price and promotion validation should focus on commercial meaning. Negative paid prices, discounts without a defined basis, promotion dates that exclude the transaction, and basket allocations that fail to reconcile all require review. The team should also test the opposite error: a supposedly full-price sale carrying an implicit bundle or delivery concession. These checks keep a demand model from learning administrative artefacts as customer response. [S37]

7.5

Returns, cancellations and substitutions stay visible

A return does not erase the history that led to it. The original order, payment and dispatch remain events, while the return adds its own time, quantity, condition and reason. Cancellation similarly needs a position in the process. Cancelling before dispatch and returning after delivery have different fulfilment and working-capital consequences. The forecast contract can later define net demand at a stated cut-off, but the ledger should not precompute that target by deleting the original sale. [S06] [S37]

Return reasons serve product learning as well as target construction. Fit, perceived quality, damage, late delivery and an unwanted substitute describe different failure mechanisms. They should use a controlled taxonomy with an unknown state and room for a verified note. A reason recorded by the customer, inferred by staff and assigned after inspection should not carry the same status. The forecast target can remain net fulfilled purchases while the reason fields inform separate product and service reviews.

Outcome maturity is essential at every evaluation cut-off. At any forecast origin, some orders have not had a realistic opportunity to be delivered, cancelled or returned. The record should mark those outcomes as open rather than treating them as retained

sales. Later arrivals create a new target version with both original event times and arrival times. This prevents a replay from benefiting from a final net-sales value that was unknowable when the earlier decision was made. [S37]

Substitution needs a chain of custody. The requested variant, suggested alternatives, customer acceptance, delivered variant and subsequent return should remain linked where the operation supports that detail. Availability and substitution are linked in retail demand analysis, so the alternative set at the moment of purchase belongs beside the transaction. [S06] A delivered substitute should count as demand for the delivered product under the physical-sales contract, while any inferred transfer to the requested product remains a modelling assumption.

The aggregation rule must be consistent across reporting views. Gross orders, fulfilled units, retained units and net units after cancellation answer different planning questions. They can reconcile without being collapsed into one field. A weekly report should identify the cut-off and maturity rule applied to each. The raw ledger remains event based, so another defensible target can be constructed later without requesting a new extraction whose historical definition has changed. [S0701]

This approach also exposes operational feedback across the garment range. A poor size offer can create stockouts in some variants and returns in others. A substitution rule can lift recorded sales while weakening satisfaction. Those are hypotheses to investigate with linked evidence, not conclusions to assert from a return rate alone. The value of the ledger is that it preserves the sequence needed to test them.

7.6

External features need as-of timestamps to prevent leakage

An external feature needs a value time and a knowledge time. The value time identifies the period the observation describes. The knowledge time records when that value became available to the venture. Publication and ingestion timestamps, source version and revision history complete the chain. Without them, a backtest can join a launch to a later revised series and call the result foresight. Leakage research specifically covers spurious relationships created through data collection and preprocessing, so as-of capture belongs in the raw evidence layer rather than in a final modelling disclaimer. [S37]

Search data illustrates the required restraint. Research on consumer behaviour found that search queries can guide near-future prediction in some settings, while results varied by application and baseline. [S34] A launch register should therefore preserve query definition, geography, collection method, observation interval and retrieval time. A later change in query grouping or geographic mapping creates a new feature version. The model must test incremental value against the existing prior; the capture system must not assume that a larger search number represents additional purchases.

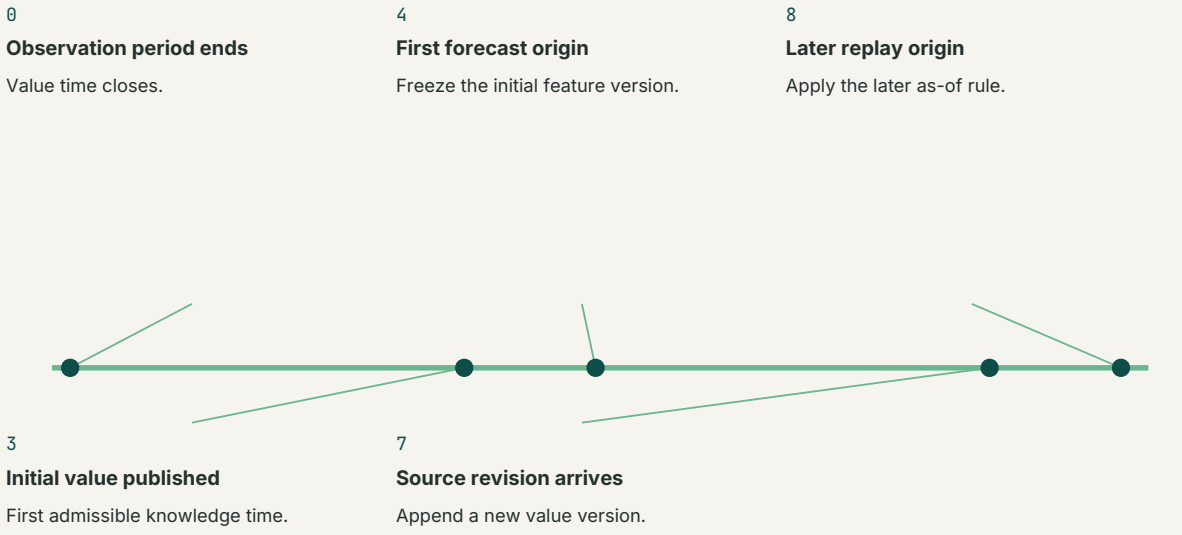
Weather, macroeconomic and trade inputs follow the same contract even when their publication rhythms differ. Empirical work on retail sales found that unusual weather has a modest role in explaining monthly fluctuations, which justifies recording weather observations and forecast vintages separately rather than substituting realised

weather for what was predicted before the selling week. [S35] Trade series require similar care: country-reported flows are submitted to international compilers on varying schedules and often show bilateral mirror asymmetries between reporting partners. [S41] [S42] Each record therefore states the source, reporter, partner, original period, unit, transformation and revision status. Missing means not yet published or not successfully collected, never zero by default. Where a source issues a correction, the earlier release remains attached to any forecast origin at which it was the available value.

Exhibit 33, p. 120 shows the intended replay rule with a constructed publication sequence. The first forecast can use only the initial release available at its origin. A later replay may use the revision if its own origin permits it. The dates are illustrative planning positions, not a claim about a particular publisher's lag. [S37] [S34]

EXHIBIT 33 / ILLUSTRATIVE DAYS AFTER THE OBSERVED PERIOD ENDS

A forecast replay must use the feature value known at its origin



ILLUSTRATIVE [S37] [S34] Illustrative publication and revision sequence. Initial release on day 3, first forecast origin on day 4, source revision on day 7 and later replay origin on day 8. The first forecast uses the initial value; the later replay may use the revision. These are constructed timings, not observed publication lags.

The exhibit also clarifies what a feature snapshot must contain. A file of current values is insufficient. The snapshot pins every source record and transformation used at the forecast origin, including missingness and any imputation fitted on the training evidence. Preprocessing fitted across the full history can leak later information even when each individual value carries a plausible date. The capture layer makes the necessary separation possible; the evaluation protocol must still enforce it. [S37]

A practical registry gives every external feature an owner and a permitted forecasting role. The owner monitors failed retrieval, changed definitions and unexpected revisions. The permitted role states whether the feature may inform a near-term covariate, an analogue screen or only a contextual note. Features that cannot be reproduced as of their claimed origin stay out of launch replay. This is stricter than retaining every interesting signal, and it protects the small launch corpus from hindsight.

7.7

A small team needs validation, ownership and recovery before scale

The control system should be proportionate to the decisions at risk. Research on hidden technical debt warns against small accuracy gains purchased with large increases in system complexity, and infrastructure research highlights the cost of maintaining end-to-end data preparation and monitoring pipelines. [S39] [S0702] For a cold-start team, the answer is not to omit controls. It is to automate a short set of high-consequence checks, assign clear ownership and make recovery routine. More pipelines and more conditional exceptions would create the maintenance burden that a small forecasting service cannot afford.

Every critical field needs four recorded elements: an accountable role, a validation rule, a completeness check and a recovery procedure. [S36] [S40] The demand owner protects target and availability definitions. The commercial owner resolves product, price and promotion exceptions. The data owner maintains identifiers, arrival times and source versions. Risk-management guidance emphasises that human judgement must set the specific metrics and threshold values used to govern trustworthiness, so named roles approve admissibility rules rather than leaving them to silent defaults. [S36] One person may hold more than one role in a lean venture, but the approvals should remain separate records so that commercial pressure does not silently redefine the evidence.

Validation should proceed in layers. Schema checks establish required fields, types and controlled values. Sequence checks identify impossible event orders. Reconciliation checks compare order lines with payment, dispatch and return quantities. Availability checks test sales against valid offer intervals. As-of checks prevent records arriving after the forecast origin from entering its snapshot. A failure receives a reason code and affected scope, not just a warning light.

Exhibit 34, p. 122 illustrates the review form by field and sales week. Values are constructed percentages of expected records that are missing or late, not benchmarks from a real launch. The purpose is to locate a failure before it contaminates a training target. Thresholds must be set in the pilot contract from operational tolerance and decision risk, not copied from this illustration. [S36] [S40]

EXHIBIT 34 / ILLUSTRATIVE PERCENT OF EXPECTED RECORDS MISSING OR LATE

Completeness checks belong by field and sales week

	Week 1	Week 2	Week 3	Week 4
Paid price	2	1	0	0
Availability	8	12	5	3
Returns	0	4	9	6
Attributes	5	2	1	1
Channel ID	1	0	0	2

Cell scale: 0 to 12

ILLUSTRATIVE [S40] Illustrative data-quality grid. Values are constructed to demonstrate field-by-week review, with 0 meaning no detected missing or late records in this example. They are not observed completeness, acceptable thresholds or a GFT operating result.

The constructed availability failure in the second week matters because its consequence differs from a late return reason. Missing availability can change the meaning of zero sales and may block primary training use. An incomplete return reason may leave net units intact while restricting product diagnostics. The review should therefore attach actions to failure modes, not pursue one aggregate completeness score that treats every field as equivalent.

Recovery begins by containing use. The affected records receive a visible status, and the forecast release applies the pre-agreed rule: restrict the target, retain the prior, use the last approved data version or withhold publication. Backfills arrive as new record versions with original event times preserved. A successful repair clears the quarantine only after reconciliation. It does not rewrite the evidence snapshot used by an earlier forecast. [S37] [S40]

Machine-learning operations guidance calls for tracking data summaries and monitoring performance so that notifications or rollback can follow deviations from expectations. [S40] Combining that operational log with a dataset datasheet gives every release a durable record of collection rules, known gaps and recommended uses. [S0701] It records exceptions, affected products and weeks, owner, repair action, resolution time and forecasting consequence. Model rollback then has a data explanation rather than becoming a generic response to any disappointing result.

GFT NOTE

GFT recommends a minimum launch-control pack before model selection: an event dictionary, identifier and crosswalk register, availability log, versioned offer catalogue, as-of feature registry, dataset datasheet, field-level validation report and recovery ledger. [S40] [S0701] Keep the pack in a small number of inspectable tables. Release the forecasting dataset only when its critical exceptions have named dispositions. If the evidence fails, retain the prior or approved baseline and repair capture before adding model complexity.

The final acceptance test is reconstructive. A reviewer who was not present at launch should be able to select a forecast origin and rebuild the events, offers, availability and external features then known. The reviewer should also be able to explain every later correction without destroying that earlier view. When this test passes, the first transactions become a durable evidence asset. When it fails, additional modelling would only conceal the uncertainty that capture has left unresolved.

Preserve what happened, what was knowable and what constrained the sale; the model can learn only the history those records defend.

CHAPTER SOURCE NOTES

S06 Lancaster University Management School, Munich Personal RePEc Archive. 2018-11-19.

S37 Cell Press, Patterns; open full text from PubMed Central. 2023-08-04.

R06 Cameroon Trade Hub. Publication date unrecorded; inherited source-text check.

R07 Ecofin Agency. Publication date unrecorded; inherited source-text check.

S0702 arXiv. 2017-05-22.

S40 Google Cloud Documentation. 2024-08-28.

S02 arXiv. 2019-06-27.

S39 Neural Information Processing Systems Foundation, NIPS 2015. 2015.

S18 International Journal of Forecasting, Elsevier; abstract through Lancaster EPrints. 2005.

S19 International Journal of Production Economics, Elsevier; abstract through Lancaster EPrints. 2001.

S07 arXiv. 2026.

S0701 arXiv. 2021-12-01.

S05 Springer Nature, International Journal of Data Science and Analytics. 2026-01-08.

S03 arXiv. 2020-07-10.

S04 MDPI, Processes. 2021-09-03.

S34 Proceedings of the National Academy of Sciences. 2010.

S35 Board of Governors of the Federal Reserve System, Finance and Economics Discussion Series. 2000.

S41 United Nations Statistics Division, UN Comtrade Help Center. 2026-10-30.

S42 United Nations Statistics Division, UN Comtrade Help Center. 2026-10-30.

S36 National Institute of Standards and Technology, U.S. Department of Commerce. 2023.

GREEN FUTURE TECH

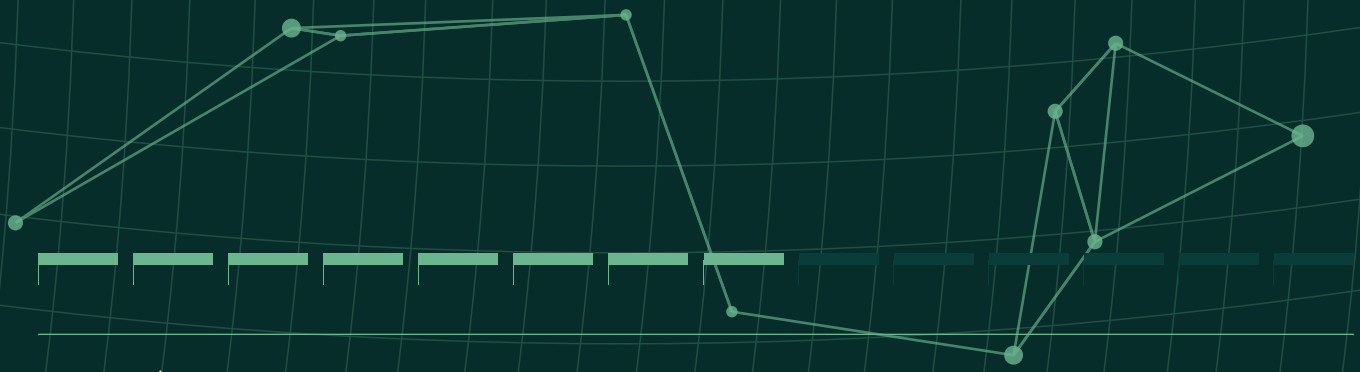
LIMITED / HONG KONG SAR

08

PART III: MODEL WITH SCARCITY AND UNCERTAINTY / BASELINE LADDER

Simple models remain the cold-start benchmark

A modest model with a visible assumption is the correct starting rival for every more elaborate candidate.



CHAPTER 08
DECISION TEST

A modest model with a visible assumption is the correct starting rival for every more elaborate candidate.

Simple models are the floor because they make a cold-start forecast falsifiable. Before the first target observation, the reference is an analogue or attribute prior with stated provenance. As observations arrive, naive, smoothing and intermittent-demand forecasts expose whether additional structure has learned anything useful. None of these methods is automatically the best production model. Each is a control against unsupported confidence. A more elaborate candidate earns release only by beating the eligible baselines in the same launch replay, with acceptable calibration, planning consequences and operating cost.

This position does not confuse simplicity with carelessness. A baseline still needs the correct demand event, product grain, availability rule and information cut-off. It also needs an explicit eligibility condition. A seasonal repetition is meaningless before a relevant seasonal cycle exists. A smoothing model cannot repair a stockout-censored target. An attribute regression cannot learn from a catalogue the venture does not possess. The baseline ladder therefore begins with evidence conditions, not a software leaderboard.

8.1

Naive and seasonal-naive forecasts expose false sophistication

A naive forecast carries the latest valid observation forward. A seasonal-naive forecast carries forward the observation from the corresponding point in the previous seasonal cycle. Their value is not that demand should remain unchanged. Their value is that they impose no unverified story about why it should change. If a candidate cannot improve on that refusal to speculate, its additional parameters have not demonstrated practical skill.

The empirical warning is specific rather than universal. In a comparison of monthly series from the M3 competition, a seasonal random walk was more accurate than half of the tested machine-learning methods. The authors also reported greater computational requirements for the machine-learning methods. This does not establish that statistical methods will win on every apparel launch. It does remove any presumption that a learned method is superior before a fair comparison. [S09]

The M4 results strengthen that caution. The six pure machine-learning entries in the reported comparison did not beat the combination benchmark, and only one beat Naive2. Combinations of mostly statistical approaches dominated the leading group. Again, the correct transfer is a test protocol, not an imported performance claim: every apparel challenger must face naive and combination controls under the venture's own launch conditions. [S10]

8.2 THE NO-HISTORY BASELINE COMES FROM OUTSIDE THE TARGET SERIES

A completely new product has no latest target observation to carry forward. Its baseline is therefore the frozen analogue prior developed before launch, or a pre-specified attribute model trained on other products. Calling this a baseline matters. Otherwise, a later model could claim an improvement against a weaker comparator selected after the results were known. Forecasting practice allows judgemental methods when the item of interest has no observations, which makes a documented prior a legitimate starting control rather than an absence of method. [S08]

The pre-launch baseline record should contain the analogue inclusion rules, availability corrections, attribute transformations, central estimate and dispersion. It should also state which decisions the range is allowed to inform. A later naive model does not erase this record. It supplies a new rung whose results can be compared with the original outside view at the same forecast origin.

8.3 THE SEASONAL BASELINE WAITS FOR A REPEATED CYCLE

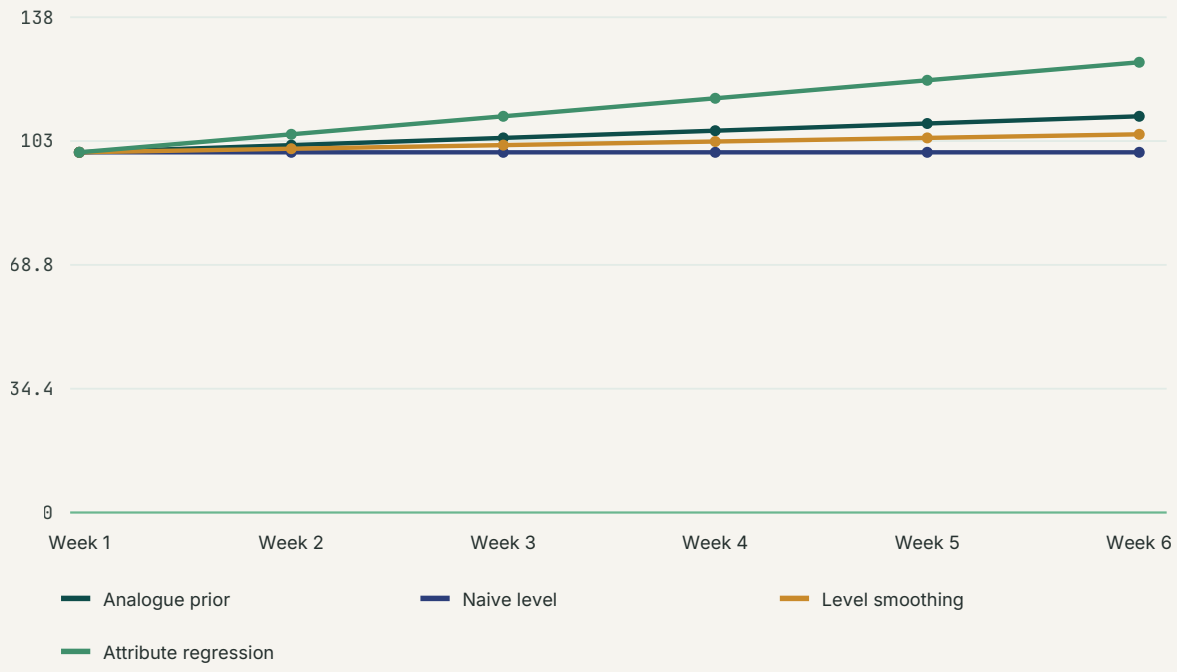
Seasonality is an eligibility claim, not a calendar label. A weekly repeating pattern cannot be established merely because a planner expects a weekend effect. The seasonal-naive rung becomes available only when the corresponding earlier observation exists, was measured under the same target definition and belongs to a cycle the planning team is prepared to defend. Changes in listing duration, availability, channel or promotional exposure can make that earlier observation a poor reference even when the calendar matches.

This discipline prevents two common errors. The first is fitting seasonal parameters to a launch ramp and calling the resulting shape seasonality. The second is selecting the seasonal period after inspecting the full outcome. Both errors let hindsight enter the baseline. The seasonal length and any exceptions belong in the frozen forecast contract before the replay begins.

The first baseline fan in **Exhibit 35**, p. 127 is deliberately illustrative. All methods share an assumed starting level of 100 units so that their disagreement can be seen without pretending to represent a real garment. The paths diverge as the horizon extends. They are not prediction intervals and carry no claim about a GFT engagement or a Cameroon apparel result.

EXHIBIT 35 / ILLUSTRATIVE UNITS PER WEEK

A baseline fan should widen where launch evidence is weak



ILLUSTRATIVE [S08] [S02] Illustrative model comparison. Every path starts at an assumed 100 units. The analogue path adds 2 units a week, smoothing adds 1 after the origin, and the attribute path adds 5. The naive path remains flat. These are constructed paths, not observed demand or calibrated uncertainty.

The useful output is the disagreement itself. A wide gap between defensible baselines tells the planning team that model choice still matters. It is a reason to retain a broad range and gather launch evidence, not to select the most reassuring path. Once actuals arrive, the same frozen paths become scored forecasts rather than scenarios.

8.4

Exponential smoothing becomes a durable baseline as history emerges

Simple exponential smoothing replaces the single latest observation with a recursively updated level. Recent observations receive more weight, but the forecast remains flat when the model contains only a level. This makes it a natural next rung after the naive forecast: it tests whether a filtered level is more useful than the last valid point without introducing a trend or seasonal narrative.

The broader forecasting review describes exponential smoothing as a continuing benchmark and notes that simple approaches are less prone to overfitting than complex methods. That is why smoothing remains useful both alone and inside combinations. Its role here is deliberately modest. It estimates only the components that the available launch evidence can support. [S08]

8.5 LEVEL ENTERS BEFORE TREND, AND TREND ENTERS BEFORE SEASONALITY

The ladder begins with a level-only model. A trend component is introduced only when successive, availability-aware observations support a direction that persists across forecast origins. A seasonal component comes later still, after enough comparable cycles exist to distinguish repetition from a one-off launch pattern. Adding all components because the software permits them would confuse model capacity with evidence.

A trend also needs a planning interpretation. Early sell-through can rise because distribution expands, a stockout ends or a promotion starts. None of these mechanisms guarantees an underlying demand trend. The baseline record should therefore carry known offer changes beside the series. Where those changes explain the ramp, a trend model is not eligible to convert them into an automatic continuation.

Damping deserves the same scrutiny. An unconstrained trend can extend a short launch pattern far beyond its evidence. A damped variant may be tested as a separate candidate, but damping is not a licence to fit a trend earlier. It is one more assumption to replay, with its parameter selected only on information available at each origin.

8.6 A WORKED UPDATE SHOWS WHAT THE MODEL ACTUALLY KNOWS

> **WORKED EXAMPLE: ONE SMOOTHING UPDATE**
>
> Assume an illustrative previous level of 100 units, a new valid observation of 120 units and a smoothing weight of 0.25. The updated level is 0.25 times 120 plus 0.75 times 100, which equals 105 units. A level-only forecast for the next period is therefore 105 units. These figures demonstrate the recursion only. They are not observed apparel demand, an estimated smoothing weight or a recommended production quantity.

The calculation makes the model inspectable. A planner can see that the new observation moved the level only part of the way towards the latest sales value. That restraint can help when a single week is noisy. It can hurt when the series has genuinely shifted. The replay should show both cases rather than describing smoothing as intrinsically conservative or responsive.

> **METHOD**
>
> For every smoothing rung, freeze the component set, initialisation rule, parameter-search range and missing-data treatment before evaluation. Fit parameters inside each training fold. Record stockout and listing changes beside the target. Compare forecasts at every required horizon, then retain the simplest variant whose performance remains stable across launch origins and product groups.

A smoothing baseline also needs a failure output. If there are too few valid observations for the approved component set, the service should return the previous eligible rung and mark the reason. Silent substitution of a more flexible model would destroy the comparability that the ladder is meant to protect.

8.7

Intermittent-demand methods address zero-heavy series with care

Sparse apparel variants can record no sales in many periods even when the listing and availability records are valid. Intermittent-demand methods are designed for that pattern. Croston-type approaches separate the size of nonzero demand from the interval between demand occurrences, rather than forcing a smooth level through alternating zeros and purchases. The research definition is general, while the empirical comparison in the cited study concerns automotive demand, not clothing. The method is therefore a candidate baseline whose value must be established on apparel cells. [S18]

The variant matters as much as the family name. Syntetos and Boylan identified bias in the original Croston estimate and introduced a correction that gives approximately unbiased demand-per-period estimates. A baseline labelled only “Croston” is incomplete. The register must name the estimator, smoothing parameters, initialisation and treatment of intervals with no observed event. [S19]

> **WORKED EXAMPLE: NAMING THE INTERMITTENT ESTIMATOR**
>
> Take an illustrative nonzero demand size of 10 units and an illustrative mean interval of 2.5 periods. The unadjusted Croston rate is 10 divided by 2.5, or 4 units a period. With an assumed smoothing weight of 0.2, the Syntetos and Boylan correction multiplies that rate by 1 minus 0.2 divided by 2, giving 0.9 times 4, or 3.6 units a period. The arithmetic demonstrates why the variant belongs in the forecast record. It is not an estimate for any apparel product.

8.8 A ZERO BECOMES USEFUL ONLY AFTER ITS CAUSE IS KNOWN

The intermittent rung is eligible only for observed zeros under a valid offer. A blank row can mean that the product was not listed, the size was unavailable or the channel failed to report. Treating those conditions as zero demand would teach the model that commercial absence is customer rejection. The data contract must resolve that distinction before any intermittent estimator runs.

Aggregation also needs a stated rule. Sparse size-level demand may be better evaluated alongside a style-level total, but pooling away every zero would hide the inventory problem the variant forecast must inform. The readout should therefore preserve occurrence and size errors at the actionable grain, then show how those forecasts add up at the style and channel levels. No aggregate score should conceal a variant whose zero pattern has been misclassified.

The illustrative register in **Exhibit 36**, p. 130 converts this requirement into a compact audit record. Its values repeat the worked example and make the correction inspectable. A production register would add the actual sequence, availability flags and fitted parameters for each forecast origin.

EXHIBIT 36 / ILLUSTRATIVE UNITS AND PERIODS

The intermittent baseline must expose its correction and inputs

Register item	Value	Decision meaning
Nonzero demand size	10	Assumed event size
Mean demand interval	2.5	Assumed periods between events
Croston rate	4	Size divided by interval
Correction multiplier	0.9	Named bias correction
Corrected rate	3.6	Baseline after correction

ILLUSTRATIVE [S18] [S19] Illustrative arithmetic from the adjacent case. Croston rate = 10 / 2.5 = 4. Syntetos-Boylan multiplier = 1 - 0.2 / 2 = 0.9. Corrected rate = 4 × 0.9 = 3.6. No observed apparel sequence is implied.

8.9

Diffusion curves describe adoption only under strong assumptions

Diffusion models answer a different question from weekly variant forecasting. The Bass model concerns the timing of initial purchases and relates that timing to previous buyers through innovative and imitative behaviour. The original study developed and tested the model on consumer durables. That scope is useful for category adoption scenarios, but it is not direct evidence for short-life apparel sales or repeat purchases. [S30]

A diffusion curve can nevertheless serve as an explicit scenario when the decision concerns the shape of first-time adoption for a defined market. Its market potential and adoption parameters must come from declared analogue evidence or structured judgement, with sensitivity ranges. Until target observations identify them, they are assumptions rather than fitted facts. The scenario should also state whether the market boundary, channel access and product definition remain stable.

8.10 LAUNCH MECHANICS CAN MIMIC ADOPTION WITHOUT PROVING IT

A new apparel line often changes distribution, availability and communication during its first selling periods. Those actions can produce an accelerating sales path without demonstrating social imitation. Conversely, constrained stock can produce a flat path even when underlying interest is growing. A curve fitted to the visible path would confound commercial exposure with adoption unless those conditions enter the account.

Repeat purchasing creates another boundary. A model of initial-purchase timing should not silently become a forecast of replenishment, returns or size-level sales. Where those outcomes matter, they need their own target definitions and baselines. The diffusion scenario stays above the operating forecast unless launch replay demonstrates a defensible mapping between the two.

The strongest pre-launch use is therefore comparative rather than automatic. Planners can ask how a slower or faster adoption assumption changes a category-level range while holding the analogue evidence visible. They should not promote a diffusion curve to the main forecast merely because it produces a smooth and plausible shape.

8.11

Attribute regression links new items to analogue demand

Attribute regression offers a more direct bridge to an unseen item. Instead of asking only what previous sales did, it asks how category, material, colour, price and merchandising characteristics relate to outcomes for other products. The cited fashion study built generalised models from clothing and footwear attributes and merchandising factors to forecast new items. Its essential requirement is historical data from those other products, which makes this a transfer baseline rather than a solution from no evidence anywhere. [S02]

Product age, selling price, temporal features and store context can extend the model while keeping its inputs visible. The product-age study models fashion sales as a function of those fields in an established retail setting. For a new manufacturer, the same fields should be treated as candidate predictors whose definitions and transfer must be checked, not as coefficients that can be imported directly. [S03]

8.12 A SMALL REGRESSION EARNS TRUST THROUGH EXPLICIT CONSTRAINTS

The first attribute baseline should use a limited taxonomy and a regularised functional form that the planning team can inspect. Each coefficient or effect needs a unit, sign and training population. An effect for a price band, for example, is meaningful only alongside the currency, price date, channel and included product set. A prediction for a garment outside the training range should carry an extrapolation flag rather than a silent point estimate.

The target transformation is equally important. A model of expected units, a model of purchase occurrence and a model of nonzero order size answer different questions. Sparse variants may require separate occurrence and size components, while a sufficiently continuous style total may support a direct level regression. The choice belongs in the baseline specification and must remain consistent between training and replay.

Residual analysis supplies the practical check. Errors should be inspected by product age, price band, channel, product family and availability condition. A good overall fit can conceal a systematic failure on the exact type of launch now being considered. Where an attribute group has weak support, the model should widen its range or defer to the broader analogue prior rather than manufacture precision.

8.13 A NEW FIRM NEEDS AN OWNED ANALOGUE CORPUS

An established retailer can train on its own catalogue. A greenfield maker may have to assemble licensed, public or partner launch records whose access and provenance vary. That difference changes the baseline's evidentiary status. The attribute model

remains simple, but its reference class may be noisy, incomplete or weakly comparable. The forecast register must show who supplied each launch, which fields were harmonised and why the target item belongs in that population.

The eligibility map in **Exhibit 37**, p. 132 summarises the resulting choice. Its axes are GFT's rule-based assessment of target history and transferable evidence, not measured accuracy. A method moves along the history axis only when its required observations exist. No position implies a performance score.

EXHIBIT 37 / GFT RULE-BASED ELIGIBILITY COORDINATES

Evidence conditions determine which baseline is eligible



DERIVED [S02] [S08] [S18] [S19] [S30] Derived conceptual mapping from the cited method requirements. The horizontal coordinate rises from no target history at 0 to repeated target cycles at 10. The vertical coordinate rises from no transferable analogue or attribute evidence at 0 to strong documented evidence at 10. Positions are eligibility guidance, not accuracy measurements.

The map prevents a common sequencing error. Transfer models dominate the upper-left region because the target lacks history. Local time-series methods become eligible as the horizontal evidence grows. Combinations become attractive when both forms are defensible. The axes describe information availability only; launch replay still decides which eligible model performs well enough to use.

8.14

Equal-weight combinations are a serious candidate

A combination baseline asks whether several defensible errors can offset one another. Equal weighting is the correct first version because it adds no fitted weight-estimation problem to an already small launch corpus. A long review of forecast combinations finds that equally weighted averages can perform well against more elaborate weighting schemes. The same review stresses that results depend on the quality and diversity of the component forecasts, not on averaging alone. [S20]

Diversity must come from information or mechanism. An analogue prior, a naive level and an attribute regression can disagree for useful reasons. Three parameter settings of the same model usually do not provide the same diversity. The combination register should name what each component contributes and exclude any forecast whose target, availability rule or information cut-off differs from the agreed contract.

> **WORKED EXAMPLE: AN EQUAL-WEIGHT LAUNCH FORECAST**
>
> Assume illustrative forecasts of 80 units from the naive baseline, 100 units from the analogue prior and 140 units from attribute regression for the same product, channel, week and information cut-off. Their equal-weight forecast is 80 plus 100 plus 140, divided by 3, which gives 106.7 units after rounding to one decimal place. The calculation is a constructed demonstration. It is not an observed launch result, and rounding does not make the underlying evidence precise.

8.15 WEIGHTS REMAIN FROZEN UNTIL THE REPLAY HAS EARNED A CHANGE

The combination rule belongs in the pre-launch specification. If planners choose weights after seeing which component came closest, the result is no longer an out-of-sample forecast. Any later weighting scheme must be trained and assessed inside each replay fold, using only forecasts and outcomes available at that origin. With few independent launches, equal weights often remain the more honest comparator because their behaviour does not depend on an unstable covariance estimate.

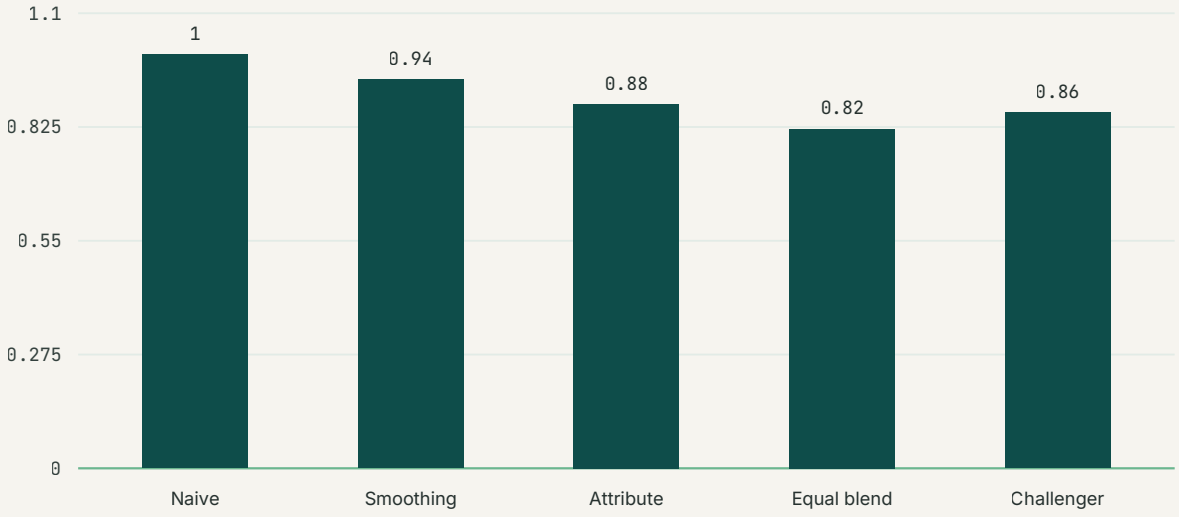
A combination also needs component-level diagnostics. The readout retains each input forecast, its eligibility status and its error by horizon and product group. If one component fails a data check, the service applies the pre-agreed rule: omit and renormalise, revert to the previous approved combination, or withhold the forecast. An unnoticed change in membership would make successive combination scores incomparable.

Point averaging should not be mistaken for uncertainty calibration. A narrow cluster of component forecasts can still share the same wrong assumption, while a wide cluster does not by itself define probabilities. Distribution construction and calibration belong to a separate evaluation step. The equal-weight baseline supplies a strong central forecast and a provenance record, not an automatic probability statement.

The illustrative scores in **Exhibit 38**, p. 134 show the required reporting form. They compare candidates on one defined product group with a nonzero naive scaling denominator. The values are constructed to demonstrate interpretation, not results from a launch replay.

EXHIBIT 38 / ILLUSTRATIVE MASE FOR ONE PRODUCT GROUP

Relative error should be reported against each eligible baseline



ILLUSTRATIVE [S12] [S20] Illustrative values for a reporting template, not empirical results. All candidates share the same as-of naive denominator, assumed to exist and be nonzero. MASE = candidate mean absolute error divided by the training-fold naive mean absolute error. Values below 1 beat that scaling reference in this constructed case.

MASE supports comparisons across heterogeneous series by scaling absolute error with a naive benchmark. In this chapter's launch protocol, that denominator must use only information available at the forecast origin. It is undefined for a no-history or all-zero training series and must not be coerced into a number. Where it is unavailable, the report uses MAE, bias and aggregate relative absolute error against the frozen analogue baseline, with the number of eligible launches shown beside the result. [S12]

The chart also shows why the release rule must name every comparator. In the constructed example, the challenger beats the naive reference but not the equal blend. A report containing only the first comparison would overstate its achievement. The baseline pack travels with every candidate so that selective comparison cannot turn a modest result into a release case.

8.16

The model ladder promotes complexity only on evidence

The ladder is a governance sequence, not a ranking of algorithms by age or technical fashion. Each rung adds one identifiable capability: transfer from analogues, filtering of a local level, explicit treatment of intermittent occurrence, use of product attributes, or combination across methods. The candidate then repeats the same launch replay against every eligible lower rung. Complexity that does not improve the decision remains in research.

This is consistent with practical machine-learning guidance to keep the first model simple and establish baseline behaviour before adding complexity. It also reflects a systems principle: a statistical gain is insufficient if the change degrades realised

system performance. For apparel planning, the release review therefore includes forecast error, calibration, planning outcome and the reliability of the surrounding pipeline. [S33]

8.17 EVERY RUNG REPEATS THE SAME EVIDENCE CONTRACT

A fair replay holds constant the target event, product and channel grain, horizons, availability treatment, launch exclusions and information cut-offs. Preprocessing, attribute selection and parameter tuning occur inside the training fold. Whole target launches remain out of training when the claim concerns an unseen item. These conditions make the baseline win meaningful rather than an artefact of unequal information.

The result pack separates product families, channels, horizons and availability conditions. It reports central error and bias beside uncertainty quality. It also reports how often each method was eligible and how often it failed back to a lower rung. A candidate that performs well only after quietly dropping difficult launches has not beaten the baseline on the same decision set.

Support size limits the conclusion. When independent launches are few, the correct gate outcome is to retain the approved prior or simple baseline and collect more evidence. A close result should not be resolved by adding metrics until the challenger wins. The release protocol fixes the primary comparisons and the required planning consequence before model scores are computed.

8.18 OPERATING BURDEN BELONGS INSIDE THE PROMOTION TEST

Machine-learning systems incur maintenance costs through data dependencies, configuration, feedback and monitoring. Research on hidden technical debt specifically warns against tiny accuracy benefits purchased with large increases in system complexity. A small industrial team should apply that warning directly: every additional feature, pipeline and model dependency must justify its continued operation. [S39]

The complexity budget records runtime, failure modes, data owners, reproducibility, planner interpretability and rollback effort. It does not impose a universal threshold. Instead, it asks whether the observed improvement is large and persistent enough to change a production decision after those burdens are included. A statistically attractive result that does not alter a viable plan has little claim to deployment.

8.19 THE BASELINE PACK NEEDS TESTS BEFORE IT NEEDS OPTIMISATION

A baseline service should pass small, deterministic tests before it enters a replay. A flat valid series should produce a stable naive forecast. A level shift should affect naive and smoothing forecasts in the expected order. A sparse sequence should trigger the named intermittent estimator only after its zeros pass the availability test. A repeated cycle should make the seasonal rung eligible only at the approved lag. An unseen item with attributes but no target rows should use the transfer path, not an empty local model.

These tests check meaning as well as code. Each one states the input records, information cut-off, expected eligibility, forecast and fallback. When a result changes, reviewers can tell whether the cause was a corrected rule, a changed parameter

search or an accidental alteration to the target. This is especially important for simple methods because their transparency is valuable only when the implemented calculation matches the written definition.

The pack also needs negative tests. Missing availability should prevent a zero-demand interpretation. An absent analogue identifier should prevent an attribute forecast from borrowing an unrecorded reference class. A zero naive denominator should produce an undefined MASE rather than an apparently excellent score. A failed component in a combination should trigger the frozen membership rule. Each failure should leave a visible status in the forecast register.

Ownership completes the control. The demand owner approves target and availability definitions. The forecasting owner approves model specifications and replay results. The planning owner confirms that the output can change a real commitment. These roles may be performed by a small team, but the approvals remain distinct records. A simple model then becomes a stable reference against which both future models and future data changes can be assessed.

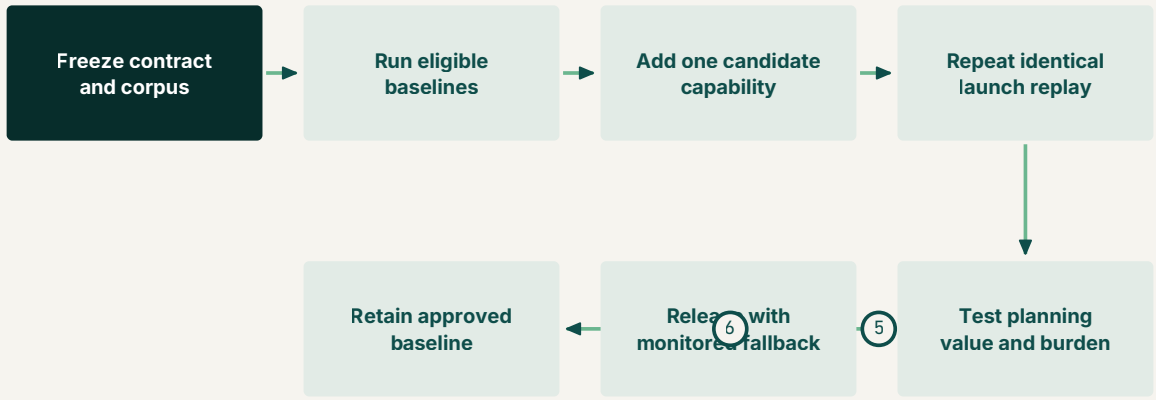
GFT NOTE

GFT recommends shipping a versioned baseline pack before any advanced challenger. It contains the frozen analogue prior, eligible naive and smoothing models, a named intermittent-demand variant, a regularised attribute model where a suitable corpus exists, and an equal-weight combination of diverse approved forecasts. Each forecast carries its information cut-off, eligibility reason and fallback. Promotion requires a launch-replay win, acceptable uncertainty, a better planning outcome and a workable operating burden. Failure returns to the last approved baseline without an emergency modelling exercise.

The derived control sequence in **Exhibit 39**, p. 137 makes this recommendation operational. It does not claim that the sequence has produced a result at a GFT client. It specifies the evidence path that a cold-start service should implement and test.

EXHIBIT 39 /

The model ladder releases complexity only through a baseline gate



5. Evidence clears gate
6. Evidence insufficient

DERIVED [S09] [S10] [S20] [S33] [S39] Derived GFT control sequence from the cited benchmark, combination, systems-value and technical-debt principles. It is a proposed operating method, not a completed implementation or measured client result.

A baseline is consequently a permanent operating asset, not a disposable stage before serious modelling begins. It anchors evaluation when data are scarce, provides a fallback when advanced models drift and gives planners a visible reason to trust or reject a revision. The forecasting service becomes more capable only when the evidence, rather than the architecture, justifies the next rung.

Let every forecast earn its complexity against a model whose assumptions can be seen.

CHAPTER SOURCE NOTES

S09 PLOS ONE, Public Library of Science. 2018-03-27.

S10 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2018.

S08 International Journal of Forecasting, Elsevier; full text through arXiv. 2022.

S02 arXiv. 2019-06-27.

S18 International Journal of Forecasting, Elsevier; abstract through Lancaster EPrints. 2005.

S19 International Journal of Production Economics, Elsevier; abstract through Lancaster EPrints. 2001.

S30 Management Science, INFORMS; abstract through EconPapers. 1969.

S03 arXiv. 2020-07-10.

S20 arXiv; journal version in International Journal of Forecasting, Elsevier. 2022-09-23.

S12 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2006.

S33 Google for Developers. 2025-08-25.

S39 Neural Information Processing Systems Foundation, NIPS 2015. 2015.

GREEN FUTURE TECH

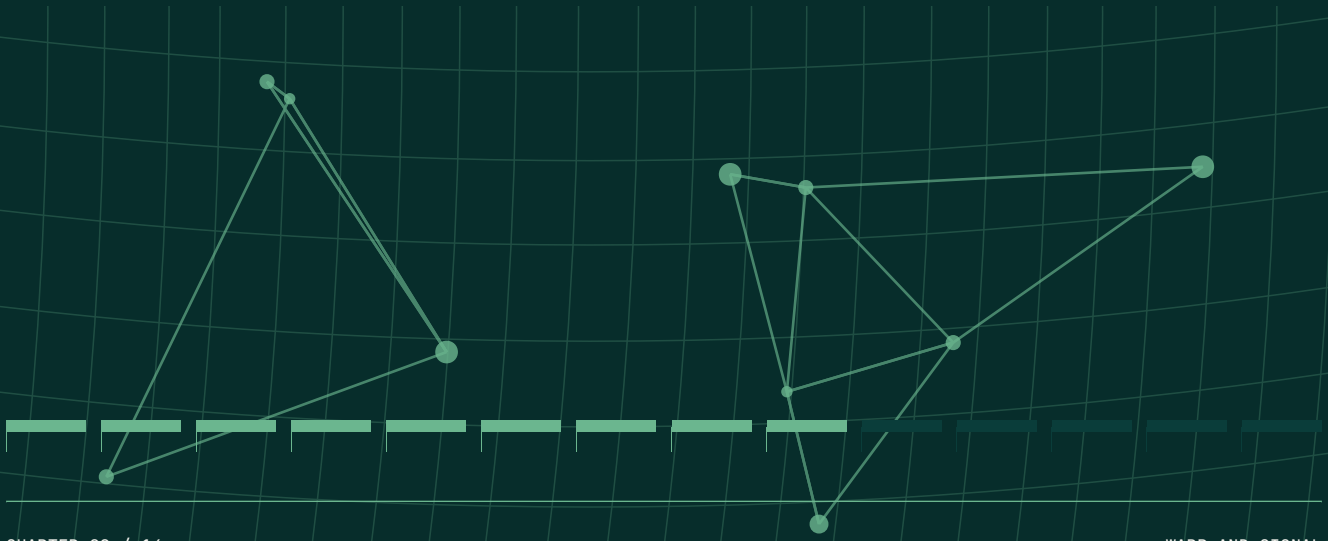
LIMITED / HONG KONG SAR

09

PART III: MODEL WITH SCARCITY AND UNCERTAINTY / CROSS-SERIES LEARNING

Global models borrow strength across products and attributes

Cross-series learning can reduce reliance on one short history, but only when the shared evidence is relevant and the transfer is tested.



CHAPTER 09
DECISION TEST

Cross-series learning can reduce reliance on one short history, but only when the shared evidence is relevant and the transfer is tested.

Global models borrow strength by learning shared structure across products while using attributes and emerging local observations to distinguish each launch. For a greenfield apparel maker, this is a candidate response to short target history, not an exemption from the need for relevant training evidence. The shared population must contain products, channels and selling conditions from which transfer is defensible. The model must then forecast launches that were wholly absent from its fitting process. Release depends on beating the frozen analogue, attribute and simple time-series baselines, with calibrated uncertainty and a planning benefit that justifies the operating burden.

The important distinction is between evidence and rows. A large table assembled from a small number of related launches is still a small launch corpus. A global architecture can use the rows efficiently without creating independent confirmation of its transfer claim. This chapter therefore treats model choice and evidence design as one problem: define what may be shared, preserve what remains local, deny the evaluation every future or target-specific advantage, and retain a simpler fallback when shared learning does not survive that test.

9.1

A global model learns one function across many series

A local model fits a separate function to each series. A global model fits one function across the group, so an observation from one product can influence the parameters used to forecast another. The formal distinction comes directly from work on locality and globality in groups of time series. That work also warns against treating global models as useful only for similar series: shared fitting can apply more broadly, but its practical value still depends on the forecasting problem and evaluation design. [S22]

For an apparel launch, the shared function might learn how early sales evolve after release, how price bands alter the level, or how a channel changes the ramp. Each product need not have enough observations to identify all those relationships alone. The global fit pools statistical evidence for common structure. Product identity, static attributes or a local component can then retain differences that the common structure should not erase.

This mechanism is not the same as averaging every historical series into one launch curve. Complete pooling would impose a common answer. A useful global model can represent product-specific inputs, varying scales and different launch dates while sharing parameters across the corpus. The design question is which structures are shared and which are allowed to vary. That choice should appear in the model register rather than remain implicit in a software default.

9.2 POOLING CHANGES THE EVIDENCE UNIT, NOT THE DEMAND DEFINITION

The training unit is usually a product-time observation. The transfer claim, however, concerns a new product, launch, channel or market. Those units are not interchangeable. Hundreds of weekly rows from one successful style do not provide hundreds of independent examples of how new styles perform. The evaluation register therefore counts independent launches alongside row counts and records family, channel and seasonal dependence among them.

The target definition must also remain constant across the shared corpus. One source cannot contribute gross orders while another contributes fulfilled sales unless the model is explicitly designed for separate targets. Returns, cancellations, listing exposure and stockouts require the same discipline. Otherwise, the global learner may be excellent at predicting differences in data collection rather than differences in demand.

Scale creates a related risk. High-volume products can dominate a loss calculated over raw units, while low-volume variants become numerically irrelevant. Normalisation, weighting or a relative loss can alter that influence, but each remedy introduces an assumption. The chosen scheme should reflect the planning decisions that matter. It should not be selected merely because it produces a flattering aggregate score.

9.3 SHARED LEARNING NEEDS A VISIBLE ROUTE FROM CORPUS TO FORECAST

The derived architecture in **Exhibit 40**, p. 141 separates the evidence corpus, shared learner, local launch state, predictive output and evaluation control. It is a proposed GFT design, not a claim that this service has already produced a measured result. Its main safeguard is that the launch-replay gate sits outside the learner and can reject the entire transfer path.

EXHIBIT 40 /

A global launch model combines shared learning with local evidence



DERIVED [S02] [S21] [S22] [S28] Derived GFT architecture from cited cross-series, attribute, probabilistic and partial-pooling methods. Layers define responsibilities and control points, not an implemented GFT system or measured client outcome.

The local launch state deserves particular attention. Before the first sale, it contains the product's attributes, offer, channel and frozen prior. After valid observations arrive, it can also contain an updated level, residual pattern or posterior state. A model that silently replaces the pre-launch logic with a different post-launch rule makes its own backtest difficult to interpret. The service should version both states and record when each became eligible.

9.4 RELEVANCE IS A CONDITION OF BORROWING, NOT A PROPERTY OF MODEL SIZE

A large external corpus can still be irrelevant to the target launch. Products may differ in use occasion, price, distribution, return policy or selling season. Even technically correct histories can teach the wrong relationship when the customer and channel differ. Relevance therefore begins with a documented reference population, not with the number of parameters or observations.

The inclusion register should state why each source launch belongs, which attributes were harmonised and which exclusions were applied before outcomes were examined. Failed and withdrawn products belong in the same frame when their records are valid. Otherwise, shared learning inherits a survivor-heavy account of launch demand. Where suitable analogues are scarce, the honest output is a wide prior and an evidence-collection plan, not a more flexible learner.

9.5

Static product attributes give unseen items a model entry point

An unseen item cannot contribute its own lagged sales at launch. Static attributes give a global model another way to place that item within the learned population. Category, fabric, colour, silhouette, price band, occasion and merchandising descriptors can connect the new offer to outcomes observed for other products. The cited fashion-retail study used clothing, footwear and merchandising attributes to build generalised new-item demand models, but its method still depended on historical data from those other products. [S02]

This distinction matters for a new firm. An established retailer may possess a clean catalogue and years of linked outcomes. A greenfield maker may hold neither. It must assemble an analogue corpus whose access rights, definitions and selling conditions can be inspected. Attribute-based transfer is therefore not a model from no evidence. It is a model whose evidence lives outside the target series.

9.6 EVERY FEATURE NEEDS A TEMPORAL ROLE

Temporal Fusion Transformer research makes a useful input distinction among static covariates, known future inputs and exogenous series observed only historically. [S24] The distinction applies even when the chosen model is not a transformer. A fabric label is static for the offer. A planned launch date or approved price can be known in advance. Actual sales, realised promotion and measured stock availability become known only as time passes.

A feature register should record what was knowable at each forecast origin. Planned promotion may enter a pre-launch forecast only if the plan itself was fixed and available then. Realised promotion cannot replace it in a historical replay. A revised product description cannot stand in for the description customers saw at launch. These are information-contract questions before they are modelling questions.

The derived register in **Exhibit 41**, p. 143 assigns common launch fields to temporal roles and permitted uses. Its final column also records the failure that the role is meant to prevent. The assignments are methodological choices derived from the cited input and leakage principles, not empirical effect estimates.

EXHIBIT 41 /

A feature becomes valid only when its temporal role is explicit

Feature family	Role at forecast origin	Permitted use	Control against misuse
Fabric and silhouette	Static	Place item in training population	Version taxonomy and missing values
Planned price and launch	Known future	Condition the intended offer	Freeze the as-of plan
Past valid sales	Observed history	Update local launch state	Exclude future and censored rows
Realised availability	Observed history	Correct target interpretation	Retain stockout timestamps
Product image	Static versioned input	Find visual analogues	Use the image offered at launch
Channel and region	Static or planned	Segment transfer and readout	Do not pool unlike selling systems

DERIVED [S02] [S05] [S24] [S37] Derived GFT feature-role contract from cited attribute, visual-transfer, mixed-input and leakage sources. Roles are design rules. No feature is claimed to improve apparel accuracy before launch replay.

9.7 VISUAL SIMILARITY SUPPLEMENTS ATTRIBUTES WITHOUT REPLACING COMPARABILITY

Visual models can represent appearance when a structured taxonomy is incomplete. A recent cold-start fashion study combines corrected historical demand, visual embeddings and clusters, then transfers demand signals from similar historical products to a new item. [S05] That is a relevant architecture candidate because it addresses both missing target history and the visual character of fashion. It is not direct evidence that the same clusters or demand relationships hold for a Cameroon apparel venture.

Visual resemblance does not establish equivalent fit, willingness to pay, quality perception or channel exposure. Two similar images can represent products with different prices and delivery promises. A studio photograph can also encode background, styling or brand conventions that have little to do with the garment. The visual pipeline therefore needs structured offer fields beside it and an audit of whether neighbours are comparable in ways the planner considers material.

Versioning is essential. The training record should link each image to the offer actually presented, including later replacements as new versions. Similarity search and cluster construction must be fitted without the held-out target's demand. Otherwise, the target can influence its own analogue set before evaluation and the apparent cold-start result will be optimistic.

9.8 BORROWED HISTORIES STILL NEED AVAILABILITY CORRECTION

A global learner trained on recorded sales can learn supply restriction as if it were weak demand. The visual cold-start study addresses this problem by correcting censored historical sales before transferring cluster-level signals. [S05] The broader lesson is more important than the particular correction algorithm: every source series needs availability and substitution context, and every correction needs its own assumptions and validation.

A new firm should not import corrected values as if they were raw observations. The corpus should preserve the original sale, stock status, substitute context, correction method and corrected value as separate fields. Sensitivity analysis can then compare transfer with and without the correction. If the forecast changes materially, the planning team should see that dependence rather than receive a single apparently precise result.

> **METHOD**
>
> Build the shared corpus as an evidence register before fitting any global model. For every launch, store the target definition, product and channel identifiers, complete observation window, source rights, availability fields, attribute versions and inclusion reason. Record independent launch counts by product family and channel. Fit imputation, scaling, attribute selection, clustering and embeddings inside the training fold. Keep the target launch and all its derived representations outside every fitting step.

9.9

DeepAR offers a probabilistic cross-series candidate

DeepAR trains an autoregressive recurrent network across related time series and produces probabilistic forecasts. Its shared fitting makes it a natural candidate when many products provide examples of temporal behaviour, while its distribution output can represent more than a central demand estimate. [S21] Those capabilities address two genuine launch needs: borrowing temporal structure and carrying uncertainty into planning.

The autoregressive part still matters. At a forecast step, the model conditions on past target values or on samples generated from its own predictive distribution. A completely unseen product therefore needs a defensible starting state and any available static or dynamic covariates. A global fit does not remove the question of what the model conditions on before the target's first valid observation. The implementation register should state that initialisation explicitly.

9.10 THE TRAINING POPULATION MUST MATCH THE CLAIMED LAUNCH CONDITION

DeepAR's description refers to a large number of related series. [S21] In a cold-start apparel corpus, relatedness needs an operational definition. It could mean shared category and comparable channel, or a broader population with attributes that let the model distinguish groups. It should not mean whatever launch records happened to be easiest to obtain.

The evaluation must hold out whole products and launches. Randomly dividing product-week rows would allow later observations from a target item into training while earlier rows are scored. That test may measure forecasting for an item already represented in the corpus. It does not measure the claimed unseen-launch condition. Leakage research identifies both temporal leakage and non-independence between training and test samples as threats to valid performance claims. [S37]

The holdout must extend to preprocessing. Product scaling, vocabulary mappings, cluster formation and any demand correction fitted from outcomes belong inside each training fold. The target product's static attributes may be used only in the way they would have been available at launch. Its future sales cannot influence a normalisation constant, nearest-neighbour set or model selection decision.

9.11 PROBABILITY OUTPUT IS A HYPOTHESIS TO CALIBRATE

A likelihood-based model generates a distribution under its chosen assumptions. That does not make the resulting intervals calibrated for new apparel launches. Heavy zeros, small counts, stockout censoring and abrupt offer changes can violate a convenient likelihood form. The output should be checked for empirical coverage, interval width and error by product group, horizon and availability condition.

Sampling adds another control requirement. Forecast generation should retain the random seed, number of samples, likelihood family and quantile algorithm so that a released distribution can be reproduced. Variation caused only by Monte Carlo noise should not be mistaken for a change in demand evidence. Where sampling noise is material at the planning quantiles, the team should increase computational precision or simplify the output before comparing models.

The most useful DeepAR test is consequently narrow. Ask whether shared recurrent learning improves on the attribute prior and simpler global alternatives for held-out launches, then ask whether its distributions remain useful after calibration. A positive answer on one dimension does not settle the other.

9.12

N-BEATS tests a generic deep architecture without covariate assumptions

N-BEATS was proposed for univariate point forecasting through backward and forward residual links and a deep stack of fully connected layers. Its generic configuration avoids time-series-specific components, and the authors evaluated it across varied target domains. [S23] This makes it a valuable architectural comparator: it tests how much forecast skill can come from sequence modelling alone.

That same input restriction defines its limitation at launch. A univariate model receives a target series, not the full attribute, availability and channel contract described earlier. Before any target observations exist, it has no target sequence from which to forecast. An N-BEATS candidate for a genuinely unseen item would therefore need an additional transfer mechanism, such as an analogue-supplied path or a separately learned initial state. That mechanism becomes part of the model and must be validated with it.

9.13 A LATER-STAGE COMPARATOR CAN STILL BE USEFUL

Once a short valid history exists, N-BEATS can test whether a flexible sequence representation beats naive, smoothing or recurrent alternatives at the required horizons. The comparison should use exactly the same observations and availability corrections. Giving the deep candidate a cleaned target while the baseline receives raw sales would confound architecture with data preparation.

The absence of covariates also creates a clean ablation. If the univariate model performs as well as a mixed-input alternative, the additional attributes and exogenous pipelines have not demonstrated enough value to justify their operation. If it fails where availability or price changes drive the series, the result shows why the launch context cannot be discarded merely to fit an architecture.

Generic interpretation should not become a causal story. A basis component or attention-like visualisation may describe how a model formed a prediction. It does not establish that a fabric, promotion or calendar event caused demand to change. The planning review should keep predictive diagnostics separate from claims about commercial mechanisms.

9.14 COLD-START ELIGIBILITY BELONGS IN THE MODEL CARD

The N-BEATS register should state the minimum valid history, treatment of leading zeros, required sequence length, forecast horizons and fallback when those conditions fail. A model that returns a number despite an empty or availability-corrupted input is not being flexible. It is concealing an unmet assumption.

For the pre-launch decision, the eligible control remains the frozen analogue or attribute prior. N-BEATS enters the ladder only when its required sequence exists or when a complete, separately validated transfer design gives it a legitimate entry point. This keeps the comparison honest about which problem each architecture can actually solve.

9.15

Temporal fusion adds input flexibility and operating burden

Temporal Fusion Transformers combine static covariates, known future inputs and historical exogenous series for multi-horizon forecasting. The architecture also includes feature-selection and gating components intended to suppress unnecessary inputs. [S24] On paper, that input structure fits the launch contract well: attributes describe the item, planned offer fields describe the intended future, and observed sales or availability update the launch state.

The fit is conceptual until validated. Feature selection inside a high-capacity model does not prove that a field was legitimately available, nor that its learned relationship transfers to a new population. Every input still needs an owner, temporal role, missingness rule and as-of timestamp. The more flexible interface increases the number of ways an implementation error can enter training or serving.

9.16 MULTI-HORIZON OUTPUT SHOULD FOLLOW SEPARATE PLANNING CLOCKS

An apparel venture may need an assortment scenario before launch, a first-lot range and a later replenishment forecast. A multi-horizon model can produce these outputs together, but their decision consequences differ. Early horizon error may alter the initial production commitment, while later horizon error may affect material procurement or channel allocation. The evaluation should preserve those distinctions rather than average all horizons into one convenient score.

Known future inputs require equal care. A price or promotion plan can change between the forecast origin and the realised period. The replay must use the plan that existed at the origin, not the final observed offer disguised as a known input. Where the plan itself is uncertain, scenarios should vary it explicitly and remain separate from the model's statistical prediction interval.

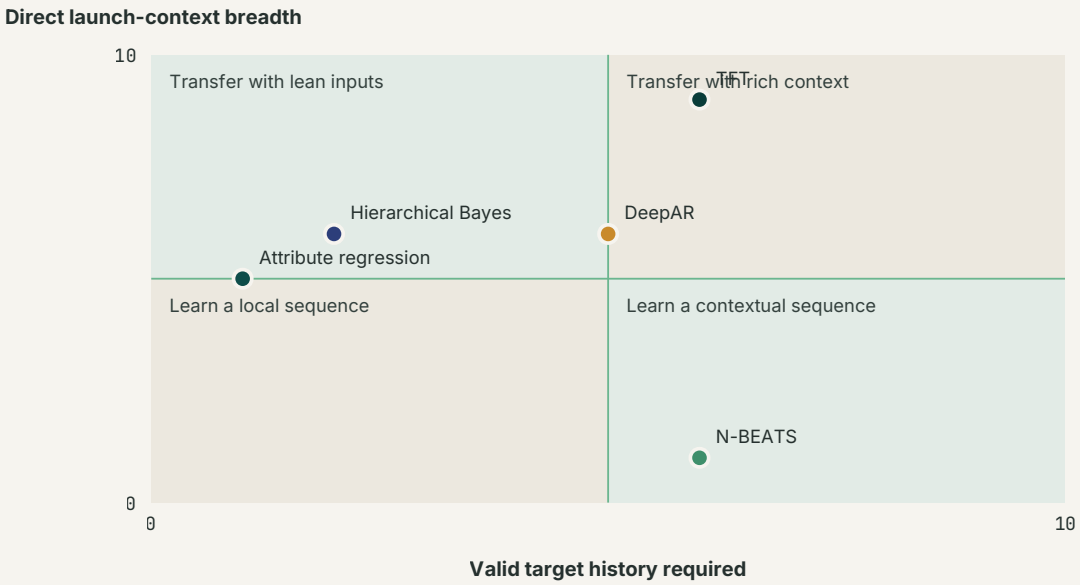
Interpretability outputs can help a reviewer inspect which inputs and time points influenced a forecast. They should not substitute for error analysis or causal identification. A planner needs to know whether the model is wrong for a specific product family, channel or stock condition. That evidence comes from segmented held-out outcomes, not from a plausible attention map.

9.17 ARCHITECTURE CHOICE FOLLOWS EVIDENCE AND OPERATING CAPACITY

The matrix in **Exhibit 42**, p. 148 places the candidate families on two rule-based dimensions: the target history their standard launch setup requires and the breadth of launch context they can represent directly. Positions are derived eligibility guidance, not accuracy scores. A model farther to the left is not automatically better, and a model higher on the chart has not proved that its extra inputs help.

EXHIBIT 42 / GFT RULE-BASED ELIGIBILITY COORDINATES

Architecture choice follows evidence and input capacity



DERIVED [S02] [S21] [S23] [S24] [S28] Derived conceptual mapping from cited method inputs. Horizontal coordinate rises from no target history at 0 to substantial valid target history at 10. Vertical coordinate rises from univariate input at 0 to mixed static, known-future and observed context at 10. Positions are not accuracy, calibration or implementation-cost measurements.

The output contract supplies the remaining distinctions. Attribute regression offers an inspectable transfer baseline and can produce uncertainty when specified accordingly. DeepAR is explicitly probabilistic. Standard N-BEATS addresses point forecasting. TFT supports multi-horizon mixed inputs. Hierarchical Bayes represents shared and local uncertainty directly. These are model-family properties, not evidence that one candidate will outperform another on the venture's launches. [S02] [S21] [S23] [S24] [S28]

Maintenance burden also differs by implementation rather than by name alone. Mixed-input models require more feature contracts and serving checks. Probabilistic models require calibration and reproducible sampling. Bayesian models require prior review and posterior computation. Univariate deep models require sequence and training controls even without covariate pipelines. The review should price those obligations in staff time, failure modes and recovery effort, not assume that a smaller feature list always means a smaller system.

Machine-learning systems research warns that small accuracy gains can be unwise when they require large increases in complexity. [S39] A small industrial team should apply a complexity budget to every candidate. It covers data dependencies, configuration, runtime, monitoring, reproducibility, planner comprehension and rollback. A candidate that cannot be operated reliably by the available team remains a research artefact even if its replay score is attractive.

GFT NOTE

GFT recommends beginning the global-model comparison with two challengers: a regularised attribute model for the true no-history case and one shared probabilistic model for launches with enough valid context. Add a mixed-input architecture only when the feature register is complete and an ablation shows that its extra inputs improve held-out launches. Keep N-BEATS as a univariate comparator when its history condition is met. Retain hierarchical Bayes where local-versus-shared parameter estimates would help planners inspect transfer. Every candidate faces the same frozen baselines, and none graduates without a planning benefit and an operable fallback.

9.18

Bayesian pooling balances shared and local parameters

Hierarchical Bayes makes the amount of borrowing an explicit modelling problem. Local parameters are drawn from a shared distribution whose own parameters are estimated from the groups. A sparse local series can therefore borrow from the population without being forced to equal it. The cited demand study estimates local and global parameters together and shows how hierarchical modelling shares information across heterogeneous sites. Its empirical application is copper demand, so it supports the pooling mechanism rather than an apparel effect size. [S28]

Three cases clarify the choice. No pooling fits each product independently and can produce unstable local estimates. Complete pooling fits one common response and can hide genuine product or channel differences. Partial pooling estimates a compromise, with greater shrinkage when local evidence is weak or groups appear similar. The method is attractive for cold starts because a new group can begin with the shared distribution and update as valid local observations arrive.

9.19 SHRINKAGE SHOULD BE CALCULATED, NOT INVOKED AS A METAPHOR

A simple normal approximation makes the mechanism visible. The local estimate moves towards the shared mean by a weight determined by local uncertainty and estimated between-group variation. Greater local noise produces greater borrowing. Greater genuine dispersion between groups permits local estimates to remain further apart. A fitted hierarchical model estimates these quantities jointly, but the approximation is useful for reviewing whether the implemented behaviour has the intended direction.

WORKED EXAMPLE: ONE PARTIAL-POOLING UPDATE

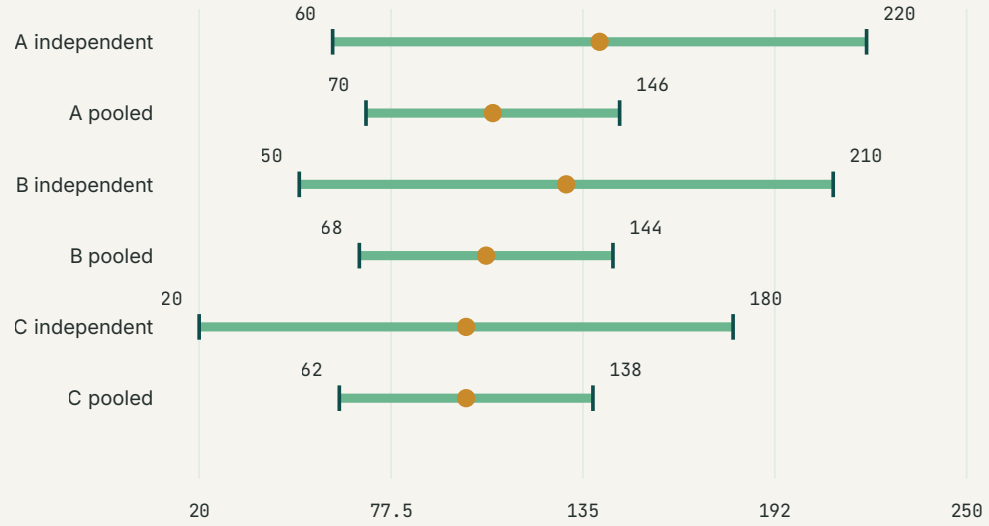
Assume an illustrative shared launch estimate of 100 units, a noisy local estimate of 140 units, a local standard error of 40 units and between-group standard deviation of 20 units. Under a simple normal approximation, the local weight is 20 squared divided by the sum of 20 squared and 40 squared, which is 400 divided by 2,000, or 0.2. The partially pooled estimate is 0.8 times 100 plus 0.2 times 140, giving 108 units. Every input is constructed to demonstrate shrinkage. This is not an apparel observation, a fitted posterior or a recommended production quantity.

The worked example also exposes an assumption. The shared mean is treated as relevant to the local product. If the target belongs to a different channel or offer population, shrinking towards that mean can create confident bias. Partial pooling is not a universal safety mechanism. Its groups and priors require the same comparability work as a global neural model.

The constructed intervals in **Exhibit 43**, p. 150 compare independent and partially pooled estimates for three illustrative groups. They show direction and dispersion only. A production exhibit would replace them with posterior intervals, effective sample sizes and sensitivity results from the fitted launch corpus.

EXHIBIT 43 / ILLUSTRATIVE DEMAND UNITS

Partial pooling should reveal both shrinkage and residual uncertainty



ILLUSTRATIVE [S28] Illustrative normal-approximation example, not fitted results. Shared mean = 100; local estimates A = 140, B = 130, C = 100; local standard error = 40; between-group standard deviation = 20. Local weight = 400 / (400 + 1,600) = 0.2, giving pooled midpoints A = 108, B = 106, C = 100. Bounds are constructed for display and are not calibrated credible or prediction intervals.

9.20 PRIORS AND POSTERIOR CHECKS BELONG BESIDE THE FORECAST

A small corpus can allow the prior and group structure to dominate the answer. The release pack should therefore show alternative defensible priors, the estimated between-group variation and the resulting change in launch forecasts. Sensitivity is not a search for the prior that gives the preferred production quantity. It is a test of whether the conclusion remains stable under reasonable modelling choices.

Posterior predictive checks ask whether data simulated from the fitted model resemble important features of the observed corpus, such as zero frequency, launch dispersion and early ramp shape. They can reveal a likelihood or grouping assumption that a central error metric misses. The cited hierarchical study uses posterior-predictive checking and prior sensitivity precisely because its local data are limited. [S28]

The planning forecast still needs a separate predictive layer. A posterior interval for a shared parameter is not automatically a prediction interval for next week's units. It must include the relevant observation noise, local uncertainty and any scenario assumptions. Calibration is then assessed on held-out launches rather than inferred from the fact that the model is Bayesian.

9.21

Similarity and global learning fail when the analogue mix shifts

A global model learns the population it is given. When the target assortment, price structure, channel or market moves away from that population, old relationships can become unreliable even if the data pipeline remains technically healthy. Concept-drift research distinguishes changes in the relationship between inputs and target from merely observing new input values. [S38] Both matter, but they call for different diagnostics.

Visual transfer has an additional dependency: the historical catalogue must remain relevant, and its demand signals must remain correctly interpreted. The cited visual cold-start architecture transfers corrected demand from similar products, so a shift in the neighbour population or correction assumptions can alter the forecast before any local sales trend becomes obvious. [S05] Similarity monitoring therefore belongs beside sales and error monitoring.

9.22 THE DRIFT REGISTER WATCHES TRANSFER CONDITIONS, NOT JUST AVERAGE SALES

The first group of measures describes the incoming population. It tracks attribute coverage, distance to training examples, price-band mix, channel mix, launch timing and the share of items outside documented training ranges. These measures can warn that the model is extrapolating. They do not prove that accuracy has already deteriorated.

The second group describes observation quality. It tracks stockouts, listing exposure, missing attributes, image versions, returns and late-arriving corrections. A change in any of these can alter the target or its measurement without a change in underlying customer interest. Data-quality alarms should prevent automatic retraining until the demand owner has interpreted the cause.

The third group describes realised performance. It tracks bias, error, interval coverage and planning outcomes by product family, channel, horizon and availability condition. Aggregate sales can remain stable while the model fails on a newly important segment. Segmented results are therefore essential to locating a transfer failure.

9.23 RELEARNING IS CONDITIONAL, AND ROLLBACK REMAINS NORMAL

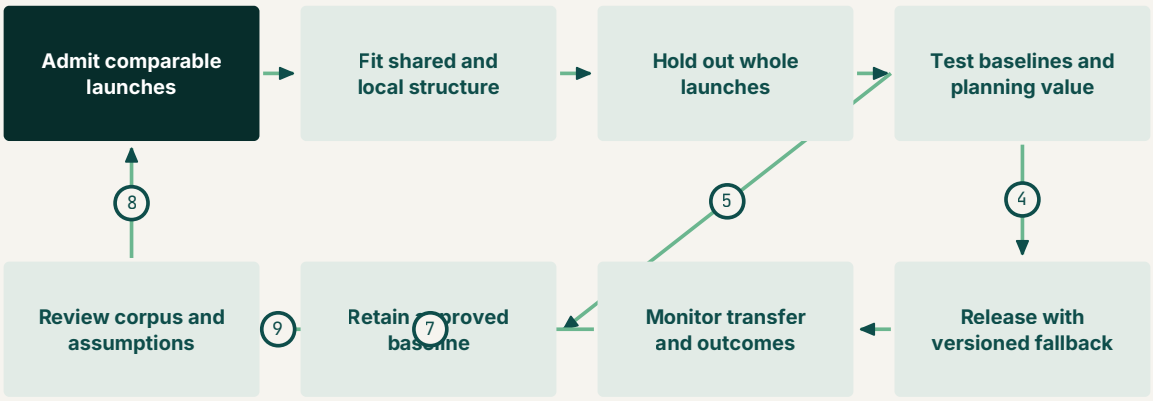
A drift alarm should open a review rather than trigger unquestioned retraining. Adding recent data can help when the new launches belong to the population the service is meant to serve. It can hurt when those launches are stockout-contaminated, structurally different or too few to support a stable update. The response may be to reweight the corpus, revise a feature, widen uncertainty, gather a pilot or revert to the frozen baseline.

Every candidate must also clear the same leakage controls used during its initial replay. A new feature selected after inspecting the held-out result cannot remain in an unchanged performance claim. A new analogue set requires a new frozen evaluation. The leakage literature's requirement for clean separation and legitimate features applies to each revision, not only to the first model. [S37]

The derived control path in **Exhibit 44**, p. 152 connects corpus admission, unseen-launch evaluation, planning value and monitored fallback. It treats rejection and reversion as successful controls when evidence is inadequate. It is a proposed GFT operating sequence, not evidence of a completed implementation.

EXHIBIT 44 /

Shared learning remains conditional from corpus admission to rollback



4. Evidence clears gate
5. Evidence insufficient
7. Transfer changes
8. Rebuild and replay
9. Revert or collect evidence

DERIVED [S05] [S33] [S37] [S38] [S39] Derived GFT control sequence from cited visual-transfer, system-value, leakage, drift and technical-debt sources. It specifies a proposed governance path, not a measured GFT result or an automatic retraining rule.

Statistical improvement remains only one admission condition. Practical machine-learning guidance prioritises realised system performance when a model change improves loss but degrades the resulting system. [S33] For a launch forecast, the final review asks whether the candidate improved a defined planning decision after accounting for availability, overrides and operating failures. If not, the baseline remains the better service.

The chapter's answer is consequently conditional. Global models are valuable when they encode a transferable population, give unseen products a legitimate entry point and preserve a route for local learning. Their strength is shared evidence under

inspection. When that population shifts or cannot be established, the disciplined action is to stop borrowing and return to a prior or baseline whose uncertainty the planning team can still defend.

Borrow strength only from evidence whose resemblance can survive examination.

CHAPTER SOURCE NOTES

S22 arXiv. 2021-03-26.

S02 arXiv. 2019-06-27.

S21 arXiv; journal version in International Journal of Forecasting, Elsevier. 2019-02-22.

S28 Journal of Industrial Ecology, Wiley. 2022-10-17.

S24 arXiv. 2020-09-27.

S05 Springer Nature, International Journal of Data Science and Analytics. 2026-01-08.

S37 Cell Press, Patterns; open full text from PubMed Central. 2023-08-04.

S23 arXiv. 2020-02-20.

S39 Neural Information Processing Systems Foundation, NIPS 2015. 2015.

S38 ACM Computing Surveys; open record from Bournemouth University. 2014-04-30.

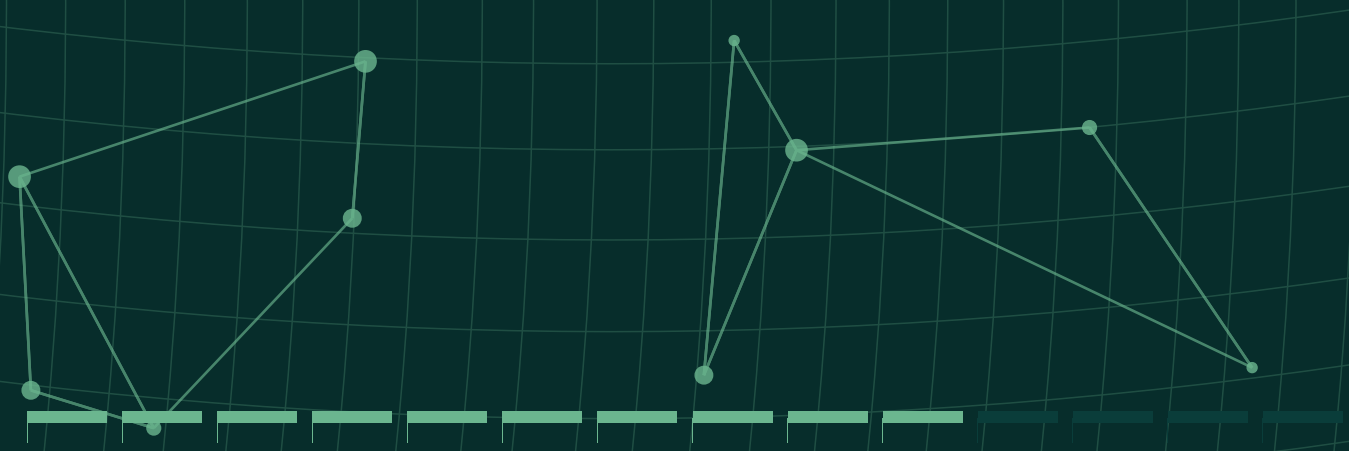
S33 Google for Developers. 2025-08-25.

10

PART III: MODEL WITH SCARCITY AND UNCERTAINTY / FOUNDATION MODELS

Pretrained forecasters must earn a place beside the baselines

Zero-shot benchmark performance is a reason to test a pretrained forecaster, not a reason to skip the evidence contract.



CHAPTER 10
DECISION TEST

Zero-shot benchmark performance is a reason to test a pretrained forecaster, not a reason to skip the evidence contract.

A pretrained forecaster belongs in a cold-start apparel programme only as a challenger with a reversible route to production. Its external training can supply useful evidence when the target product has no sales history, but it cannot establish that the evidence transfers to a particular garment, channel and market. Chronos, TimesFM and Lag-Llama therefore enter beside the analogue prior, simple baselines and global attribute models, not above them. Each must reproduce the same launch conditions, preserve the demand contract, improve planning-relevant results and operate within the team's means. A public benchmark can justify an experiment. It cannot approve a production commitment.

This position separates two propositions that are often compressed into one claim. The first is that a model can produce a forecast without fitting parameters on the target series. The second is that the resulting forecast is useful for the decision at hand. The first proposition is a plausible reason for inclusion in a test harness. The second requires target-domain evidence, calibrated uncertainty where distributions are used, acceptable decision consequences and a working fallback. The remainder of this chapter specifies that test without dismissing the genuine convenience of pretrained inference.

10.1

Zero-shot means an unseen dataset, not no evidence anywhere

Zero-shot describes the relationship between a model and an evaluation task. It does not describe an absence of training evidence. Chronos was pretrained on a large collection of public datasets and a synthetic dataset generated through Gaussian processes. Its reported benchmark covers 42 datasets, with comparable and occasionally superior zero-shot performance on new datasets relative to methods trained specifically on them. That is evidence of transfer across the study's tasks, not evidence learned without data and not a result on apparel launches. [S25]

TimesFM likewise uses pretraining on a large time-series corpus. Its authors report out-of-the-box zero-shot performance on public datasets that comes close to supervised forecasting models fitted for each dataset. The same abstract describes operation across different history lengths, prediction lengths and temporal granularities. These are useful candidate properties. They do not change the population on which the reported performance was observed. [S26]

10.2 THREE EVIDENCE LABELS PREVENT A CATEGORY ERROR

Every result should carry three distinct labels. **Pretraining evidence** identifies the external series and synthetic processes used to fit the model, to the extent that the model provider discloses them. **Evaluation evidence** identifies the datasets and forecast origins used to report skill. **Target evidence** identifies what the apparel venture itself supplies at the decision moment, including attributes, availability and any valid launch observations. A zero target-history condition concerns only the third label. It cannot erase the other two.

This separation also prevents a misleading comparison with a local baseline. The pretrained candidate may have learned from extensive external histories before seeing the target, while a naive target-series baseline has access only to the launch record.

The comparison remains useful because it represents the actual deployment options, but the report must not describe the former as evidence-free or the latter as the only model with assumptions. Both have evidence lineages that need inspection.

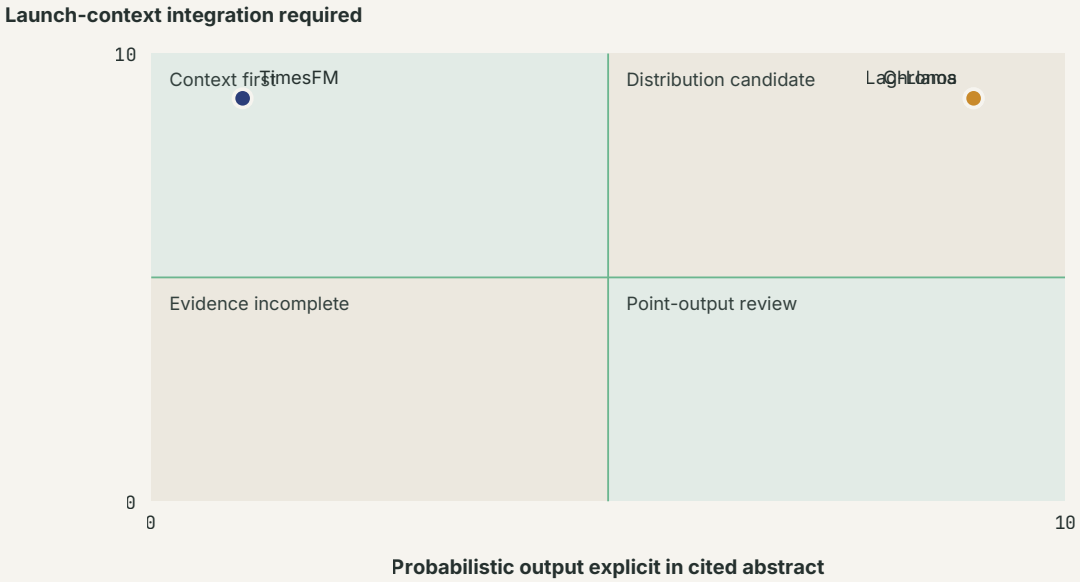
The evaluation label needs particular care when a model is used without target fitting. An unseen dataset in a public benchmark can still come from a domain, frequency or measurement process represented in pretraining. Unless the provider's dataset inventory and chronology establish otherwise, the prudent description is that the evaluation dataset was not used for fitting in the reported experiment, not that its domain was wholly unfamiliar. Where that inventory is incomplete, the limitation stays open rather than being resolved by assumption.

10.3 AN EVIDENCE MATRIX RECORDS CAPABILITY WITHOUT IMPLYING TRANSFER

Exhibit 45, p. 156 converts the opened model descriptions into a GFT evidence register. The horizontal coordinate records whether the cited abstract explicitly establishes probabilistic output. It does not rank model quality. The vertical coordinate records the integration work required by this paper's launch contract, whose product, price, availability and channel fields sit outside a univariate series. Positions are rule-based interpretations of the sources, not benchmark measurements.

EXHIBIT 45 / GFT RULE-BASED EVIDENCE COORDINATES FROM 0 TO 10

Zero-shot claims need precise evidence labels



DERIVED [S25] [S26] [S27] Derived qualitative mapping from the cited abstracts. Horizontal: 0 means probabilistic output is not established by the opened abstract and 10 means it is explicit. TimesFM's low position is an evidence-register limitation, not a claim that every version lacks distribution output. Vertical: 0 means no context integration is needed and 10 means a separate path is required for all launch-contract context. Positions are not accuracy scores.

The matrix makes two reviews unavoidable. A model whose probabilistic status is unclear needs an output-contract review before it can be compared on distribution scores. Every model needs a context review before its output can represent the launch rather than merely its sales sequence. Neither review is settled by calling the model a foundation model.

A model register should therefore store the exact model and checkpoint identifiers, provider revision, pretraining documentation, benchmark claim, input form, output form and unresolved evidence gaps. It should also state whether the candidate will be used zero-shot, adapted on reference launches or fine-tuned on target history. Those are different interventions with different validation requirements. Combining their results under one product name would hide what the evidence actually supports.

10.4

Chronos and Lag-Llama provide probabilistic candidates

Chronos converts scaled and quantised series values into tokens from a fixed vocabulary, then trains transformer-based language-model architectures with cross-entropy loss. The reported models use the T5 family and range from 20M to 710M parameters. This mechanism matters more than the linguistic analogy: the model learns a distribution over token sequences, not an understanding of garments, prices or customer intentions. [S25]

Tokenisation creates specific checks for the evaluation harness. The team should record the scaling rule, treatment of zeros and missing intervals, context length, quantisation behaviour and conversion from generated tokens back into demand units. A coarse representation may be acceptable for a noisy launch, but that is an empirical question. Very small counts, abrupt order spikes and long zero runs should be inspected in their original units so that a smooth token sequence cannot conceal an implausible operating forecast.

Sampling from a probabilistic model supplies a predictive distribution, but it does not certify calibration. The forecast service must retain enough sampled paths or quantiles to reproduce the published output, then assess interval coverage, width and quantile loss on held-out launches. A narrow distribution can be confidently wrong. A wide one can be unhelpful even when its coverage looks acceptable. Proper scoring rules provide a principled way to compare the full forecasts rather than selecting whichever quantile happens to flatter the candidate. [S14]

10.5 CHRONOS EARNS A TEST THROUGH TRANSFER, NOT THROUGH SCALE

The Chronos results distinguish datasets included in training from new zero-shot datasets. The authors report stronger performance in the former setting and comparable, occasionally superior performance in the latter. That pattern is exactly why the apparel evaluation should preserve separate readouts for reference launches used in any adaptation and wholly unseen target launches. It also prevents the strongest reported setting from becoming the default headline for a different use case. [S25]

Model size is a deployment variable, not a proxy for cold-start relevance. The parameter range reported for Chronos gives the evaluator several possible operating points, but the smallest adequate candidate is the one that belongs in a small team's complexity budget. The selection record should show accuracy and distribution quality beside runtime and failure behaviour for each tested variant. A larger checkpoint earns its additional burden only through the same launch replay. [S25]

10.6 LAG-LLAMA ADDS LAG-BASED PROBABILISTIC TRANSFER

Lag-Llama is presented as a general-purpose model for univariate probabilistic forecasting, using a decoder-only transformer with lag covariates. Its authors describe pretraining on a diverse corpus and zero-shot generalisation on downstream datasets across domains. They also report gains from fine-tuning on relatively small fractions of previously unseen datasets. These findings support two distinct candidate modes: direct inference and documented adaptation. [S27]

The distinction controls leakage. Direct inference can be evaluated on a held-out launch with no parameter update from that launch. Fine-tuning requires the target launch, or a defined portion of its early history, to enter training. The evaluation must then replay exactly when those observations became available and keep later outcomes out of every fitting and preprocessing step. A fine-tuned result cannot be presented as a pre-launch forecast merely because the adaptation was small.

Lag covariates also have a precise operational meaning. They encode earlier values at selected offsets; they do not automatically encode product attributes, stock availability or future promotions. A forecaster that performs well after seeing a stable weekly rhythm may still have no valid representation of a first production lot whose sizes are only partly available. The input register should list the lags actually used and confirm that each exists at the forecast origin.

For both probabilistic candidates, output validation should test count plausibility as well as statistical loss. Negative demand, non-monotone quantiles, unstable samples under a fixed seed and excessive mass on impossible order sizes require explicit handling. The team should not silently clip every implausible result and call the repaired distribution a model success. Any transformation belongs in the candidate specification and is applied identically during replay and operation.

10.7

TimesFM tests decoder-only transfer across forecasting settings

TimesFM uses a patched-decoder attention model pretrained on a large time-series corpus. Its design addresses variation in history length, prediction length and temporal granularity, while the reported zero-shot results concern a variety of public datasets. This makes it a practical challenger where launch records do not arrive in one neat shape. It does not make every shape statistically meaningful. [S26]

A flexible input interface should not redefine the forecast contract. A size-level weekly series, a style-level monthly series and an institutional order series may all fit the same technical input format while representing different events. The evaluation still fixes the target event, product grain, channel and horizon before inference. Padding, resampling and missing-value rules must preserve that meaning. Technical compatibility is not permission to aggregate away a decision.

10.8 VARIABLE HISTORY DOES NOT REMOVE THE LAUNCH BOUNDARY

A new product can have a very short valid history even when the model accepts a much longer one. The service should pass only observations that existed at the forecast origin and met the availability rule. It should not borrow future rows from another

channel to lengthen the sequence unless that transfer is a declared feature with its own validation. Convenience in assembling a tensor can otherwise become a route for hindsight.

Prediction-length flexibility needs the same discipline. The planning team may require several horizons, but each has its own information set and decision consequence. An assortment commitment and a short-cycle replenishment decision should not share one undifferentiated accuracy score. The replay records each requested horizon separately and tests whether performance persists across origins rather than relying on one favourable launch window.

Temporal granularity is also an analytical choice. Weekly apparel demand can contain operational events that a monthly total hides, including stockouts, listing gaps and a single institutional order. Aggregating the series may improve a score while removing the variation that the production decision must manage. The chosen grain therefore follows the commitment, with any alternative grain treated as a separate candidate rather than a preprocessing default.

10.9 ZERO-SHOT AND ADAPTED TIMESFM RESULTS NEED SEPARATE CLAIMS

The public result supports a zero-shot trial on the venture's reference corpus. If the team later adapts the model, that new candidate receives a new identifier, training record and evaluation. The report should not carry over the original zero-shot claim, because the adapted model has consumed additional evidence and acquired new failure modes. Any fine-tuning, calibration or context layer is fitted inside the appropriate replay fold. [S26] [S37]

This is especially important for institutional uniforms and size-level launches. Neither setting is validated by a general statement about temporal granularities. An institutional order can reflect a contract event rather than consumer sell-through, while a size variant can be censored by uneven availability. TimesFM may still produce a useful challenger for either target, but only a target-specific replay can establish that result. The public benchmark supplies the hypothesis, not the verdict.

10.10

Univariate pretrained models do not use the full launch context by default

The three reviewed approaches primarily consume series values, patches, tokens or lags. Chronos tokenises scaled values, TimesFM learns from time-series patches and Lag-Llama is explicitly univariate with lag covariates. None of those input descriptions, by itself, carries style, size, price, stock availability, promotion or channel into the model. Those fields require an explicit integration path rather than an assumption that pretraining has inferred them. [S25] [S26] [S27]

The launch contract should not be impoverished to suit the challenger. Availability remains necessary for interpreting observed sales, product attributes remain necessary for comparing unseen items, and channel remains necessary for explaining the selling mechanism. If the pretrained model receives only a cleaned sales vector, the discarded fields still belong in the evaluation record, the analogue prior and the competing global model. A narrow model input is not a narrow decision context.

10.11 FOUR INTEGRATION ROUTES CARRY DIFFERENT BURDENS

The first route is **univariate challenger only**. The model forecasts the valid target series and is compared with every eligible baseline on the same launches. This is the cleanest initial test because it adds no correction cascade. Context still informs eligibility, censoring and result segmentation. Failure returns directly to the approved baseline.

The second route is **combination with a context model**. A pretrained series forecast is combined with an analogue or global attribute forecast under weights fixed before evaluation. This can retain both temporal adaptation and product context, but it introduces dependence between components. The report must show each component, the combination rule and the result when either input fails. Combination weights are learned only inside training folds if they are learned at all.

The third route is **residual or correction modelling**. A context model learns to adjust the pretrained output using attributes, price or availability. This can target systematic errors, but it creates another model whose target, training population and information cut-off need validation. Research on machine-learning systems warns that correction cascades create dependencies and can make later improvements harder to assess. A small team should accept that burden only when the combined system beats both components in replay. [S39]

The fourth route is **native context-capable adaptation** where a documented model version supports additional inputs. The team then verifies the actual input schema, temporal role and implementation rather than inferring capability from the model family name. Static attributes, observed historical variables and known future inputs are not interchangeable. A promotion planned for next week may be legitimate for a forecast made today; the realised promotion result is not.

10.12 AVAILABILITY CORRECTION PRECEDES MODEL SELECTION

A univariate model cannot tell whether a zero means no buyer or no stock unless that distinction is encoded upstream. The target pipeline therefore preserves listing and availability intervals, marks censored observations and applies only an approved demand-correction method. Where the evidence cannot support a correction, the evaluation can exclude the affected interval under a frozen rule or retain it as a separate sensitivity case. It should not insert an invented demand value and reward the model for fitting it.

This order matters for every candidate, not only pretrained ones. The difference is that a convenient external model can make the cleaned vector look complete and authoritative. The forecast register should therefore link each inference input to the raw events, availability version and correction version from which it was built. A reviewer must be able to reconstruct why the model saw a particular zero, spike or missing interval.

Context also belongs in the error analysis. Results should be segmented by product family, channel, launch age, availability condition and horizon. An aggregate gain may come entirely from high-volume style totals while size-level forecasts deteriorate. Conversely, a model may improve sparse variants but fail on institutional orders. Those are different deployment decisions, and neither should be concealed by a single average across incompatible cells.

METHOD

For each pretrained candidate, publish an input-and-context sheet before replay. Name the target event, valid history rule, context length, missingness treatment, output form and every context field excluded from the model. Then choose one integration route and freeze its combination or correction rule. Fit all adaptations inside the training fold. Retain the excluded fields for eligibility, leakage checks, segmented scoring and planner review. If the input sheet and the production service differ, the candidate has not been evaluated as deployed.

10.13

Public benchmark rank is weak evidence under domain shift

A leaderboard compresses performance across datasets into a convenient ordering. A cold-start launch asks a narrower and harder question: does this model improve the forecast for the next decision, under this measurement process, with only the information available at this origin? Fashion research describes the coexistence of demand uncertainty, missing history and seasonal effects, together with short selling seasons and fragmented product-level demand. Those conditions make a general rank an imperfect guide to a particular apparel cell. [S01]

Domain shift can enter through the product, the channel, the target and the observation process. A model trained partly on regular utility or traffic series may recognise temporal structure without ever encountering a size curve shaped by stock availability. A distributor's order pattern may differ from consumer purchases even when both are recorded weekly. A new firm may also change its prices, listings and replenishment rules during the launch, altering the process that generated the series. The transfer claim should name these differences rather than refer to domain shift as one undifferentiated risk.

10.14 BENCHMARK PROVENANCE BELONGS IN THE REVIEW

The evaluator should request the model's pretraining dataset list, benchmark list, forecast origins, comparison methods and any overlap analysis that the provider can supply. This is not an accusation of contamination. It is a normal attempt to understand what zero-shot means for the cited result. If the necessary information is unavailable, the review records that limit and places more weight on the venture's own held-out launches.

The target corpus needs equal provenance. Whole launches remain held out when the claim concerns unseen products. Scaling, imputation, feature selection, residual correction and calibration are fitted only on the training side of each replay. Leakage research defines the problem as a spurious relationship created by data collection, sampling or preprocessing, and shows how it can inflate apparent model performance. Pretraining does not exempt the surrounding pipeline from that failure mode. [S37]

A fair readout also preserves launch independence. Many weekly rows from one product do not constitute many independent proofs of transfer. Closely related colour and size launches may share a campaign, stock pool or institutional order. The report

should therefore show support by independent launch group beside the row count, with sparse cells marked preliminary. A model that wins after difficult launches are removed has not passed on the deployment population.

10.15 DRIFT MONITORING STARTS AT CANDIDATE SELECTION

Concept-drift research concerns changes in the relationship between inputs and targets over time and reviews strategies for adaptation and evaluation. For a pretrained forecaster, drift includes changes between the external pretraining population and the launch population as well as changes after release. The initial model review is therefore the first drift analysis, not a separate administrative exercise. [S38]

The selection pack records the reference population, expected deployment population and known mismatches. After release, monitoring tracks input validity, forecast levels and dispersion, errors by launch group, availability conditions and the frequency of fallback. A shift in assortment or channel can invalidate transfer before average sales move enough to trigger a conventional alarm. The response may be renewed evaluation, a wider range or reversion to the approved baseline, depending on the evidence.

This approach preserves the value of public benchmarks without granting them the final vote. They help identify candidates whose learned representations may transfer. The local replay then tests whether that potential survives the actual product, channel, measurement and decision conditions. A poor local result is not repaired by citing a higher public rank.

10.16

Compute, latency and vendor dependence belong in selection

A pretrained model is a service component with an operating footprint, not merely a checkpoint that produces a score. The review must measure inference time on the intended hardware, memory use, batch behaviour, failure rate and the effort required to reproduce output. It must also record hosting, licence terms, support arrangements, model revision policy and an exit path. These are procurement and engineering requirements whose values come from the candidate under test, not universal assumptions about foundation models.

The decision clock determines acceptable latency. A weekly batch forecast can tolerate a different runtime from an interactive planning tool, but both need a defined timeout and fallback. Precomputing forecasts can reduce interactive delay while creating staleness if new availability or order evidence arrives. The operating design should state when a cached forecast expires and who can request a refresh. Faster inference is not useful if it publishes an obsolete information set.

Vendor dependence requires particular attention when weights, code or service terms are outside the team's control. A provider revision can change outputs even when the input record is unchanged. The forecast register therefore pins the deployed version, retains the validation artefacts and tests any replacement before release. Where the

service is hosted externally, the review also records the data sent, retention terms and the route for removing it. Those checks concern operational control; they do not replace legal review of the applicable terms.

10.17 THE OPERATING REGISTER EXPOSES THE FULL CANDIDATE

Exhibit 46, p. 163 is the minimum evidence sheet for a release review. Its entries are deliberately unpopulated because no candidate has been measured in this chapter. The values must come from tests on the intended deployment route and from documents actually supplied by the provider. An empty field is a blocker to approval, not a zero cost.

EXHIBIT 46 / REQUIRED RELEASE-REVIEW ENTRIES

A pretrained candidate is approved with its operating evidence

Review field	Required evidence	Release question
Model identity	Checkpoint, code, configuration and provider revision	Can the validated output be reproduced?
Runtime	Batch and interactive latency on intended hardware	Does the forecast meet its decision clock?
Resources	Memory, accelerator need and batch capacity	Can the team operate it reliably?
Cost	Measured inference, hosting, integration and support basis	Does planning value exceed full operating cost?
Rights and support	Licence, permitted use, support and revision terms	Can the service remain compliant and maintainable?
Data route	Fields sent, hosting location, retention and deletion path	Is information exposure understood and approved?
Exit	Version retention, baseline fallback and migration test	Can planners continue without this candidate?

ILLUSTRATIVE [S39] [S40] Illustrative GFT register derived from the cited technical-debt and monitoring guidance. No runtime, price, memory or vendor performance is claimed. Every value must be measured or documented for the exact candidate and deployment route before release.

Machine-learning technical-debt research warns against tiny accuracy benefits purchased through large increases in system complexity. The relevant comparison is therefore not candidate inference cost alone. It includes data preparation, correction layers, monitoring, version management, planner training and the work needed to recover from failure. If a baseline or smaller global model produces equivalent planning value, the additional burden has not earned approval. [S39]

Monitoring guidance likewise treats data statistics, online performance, notification and rollback as connected controls. A candidate without those controls is an experiment, regardless of its offline score. The release pack should name the observed signal, the person or role responsible for review, the action threshold and the forecast that planners receive when the threshold is crossed. [S40]

10.18

A challenger must beat the prior and the global baseline in launch replay

The acceptance test is a paired comparison, not a demonstration that the pretrained service can return plausible numbers. Every candidate enters the same rolling-origin, held-out-launch replay as the analogue prior, eligible simple models and the global attribute baseline. The target, availability correction, horizons and information cut-offs remain fixed. Only the declared model capability changes. This makes any improvement attributable to the candidate rather than to a more favourable dataset or preprocessing route.

Point accuracy begins with baseline-relative measurement. MASE scales absolute error by a naive reference and supports comparison across heterogeneous series. In this paper, its denominator must exist at the forecast origin and cannot be invented for a no-history or all-zero training series. Where MASE is undefined, the report retains MAE, bias and aggregate relative absolute error against the frozen analogue prior, together with the number of eligible launches. [S12]

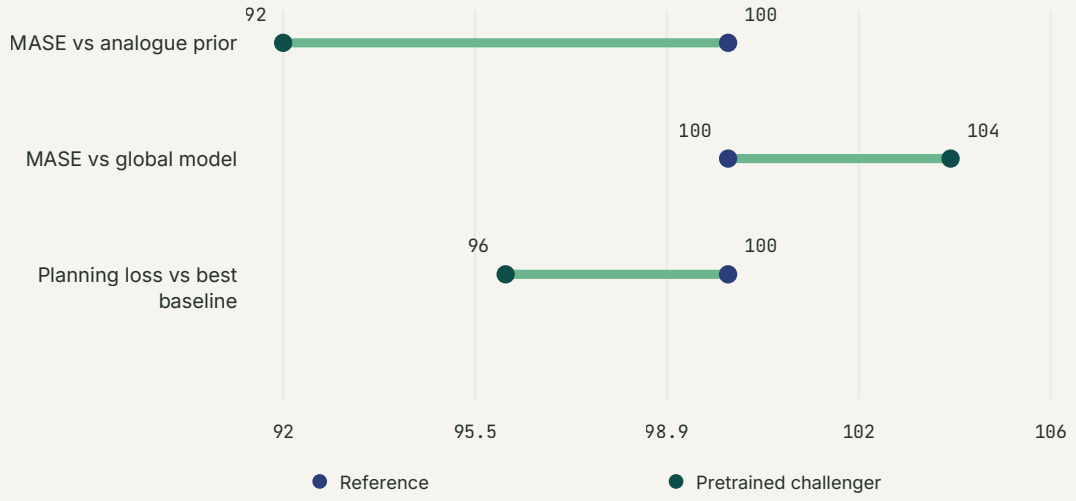
Distribution quality is assessed separately. Proper scoring rules compare probabilistic forecasts without reducing them to a favourable mean or median. Coverage and interval width remain visible beside those scores because a probabilistic architecture does not guarantee calibrated tails. The release comparison should include exactly the output form that planners would receive, including any count transformation, context correction or combination layer. [S14]

> **WORKED EXAMPLE: BEATING A SCALING REFERENCE IS NOT BEATING THE BASELINE**
>
> Assume an illustrative launch-replay cell in which the as-of naive MAE is 24 units, the pretrained challenger's MAE is 18 units and the global attribute baseline's MAE is 16 units. The challenger's MASE is 18 divided by 24, or 0.75. The baseline's MASE is 16 divided by 24, or approximately 0.67. The challenger therefore improves on the naive scaling reference but loses to the global baseline. These figures are constructed to demonstrate the gate. They are not observed apparel results, and the cell would still need distribution and planning reviews.

The illustrative readout in **Exhibit 47**, p. 165 extends that lesson to three comparisons. Indexed loss below 100 favours the challenger. The constructed candidate beats the analogue prior on central error and the best baseline on planning loss, but fails against the global model on central error. It is not released. Selecting only the favourable panels would change the governance conclusion without changing the evidence.

EXHIBIT 47 / ILLUSTRATIVE INDEXED LOSS

A pretrained challenger must clear two different benchmarks



ILLUSTRATIVE [S12] [S14] [S33] Illustrative constructed results for a release-readout template. Start values are fixed reference indices of 100. End values 92, 104 and 96 are assumed challenger results, not measurements. The example deliberately fails the global-baseline MASE comparison despite favourable analogue-prior MASE and planning-loss comparisons.

A planning-relevant improvement must then survive the decision review. Practical machine-learning guidance treats realised system performance as more important than an isolated loss improvement. For apparel, the review asks whether the forecast changes a viable lot, replenishment or inventory decision for the better after availability and override effects are included. A small statistical gain that leaves the plan unchanged cannot carry additional operating burden. [S33]

10.19 RELEASE CONDITIONS ARE FIXED BEFORE SCORES ARRIVE

The release charter names the primary metrics, eligible launch cells, minimum support, acceptable planning consequence and operating limits before results are computed. It also states what happens when evidence is thin or mixed. The default is to retain the approved baseline and gather more launches, not to search additional metrics until the challenger wins. Later model variants repeat the complete comparison rather than inheriting approval from an earlier checkpoint.

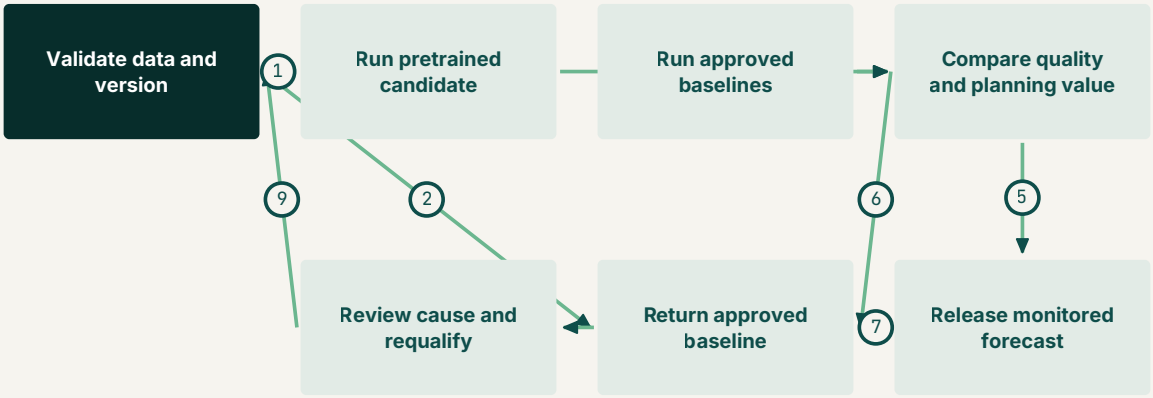
GFT NOTE

GFT recommends a challenger register beside the baseline pack. Begin with zero-shot inference on frozen reference launches, then test any adaptation as a new candidate. Promote a pretrained forecaster only when it beats the frozen analogue prior and the strongest eligible global or simple baseline on the same launches, provides acceptable uncertainty, improves the planning readout and clears the operating register. Keep the approved baseline live in shadow. Any failed input check, expired version, breached decision clock or adverse monitoring result returns planners to that baseline with the reason recorded.

Exhibit 48, p. 166 places the fallback inside the serving sequence. The quality gate protects the input and model version, while the comparison gate protects the planning output. Reversion is a normal tested path, not an admission that the forecasting service has failed. It preserves a usable decision when the challenger cannot support one.

EXHIBIT 48 / FORECAST RELEASE AND REVERSION SEQUENCE

Fallback belongs inside the pretrained-model design



1. Pass
2. Fail
5. Clear all gates
6. Mixed or insufficient
7. Monitoring breach
9. New evidence

DERIVED [S33] [S39] [S40] Derived GFT control sequence from cited system-value, technical-debt and monitoring guidance. This is a proposed operating design, not an implemented GFT client service or a measured result.

The resulting service treats pretrained forecasting as an option whose value can be measured and withdrawn. Public research supplies capable candidates. Owned launch evidence decides whether they deserve influence. The baseline remains available at every cycle, and each release remains reversible when transfer, operations or planning value deteriorates.

Pretraining can supply a challenger; only the launch evidence can award it a place.

CHAPTER SOURCE NOTES

S25 arXiv. 2024-11-04.

S26 arXiv. 2024-04-17.

S27 arXiv. 2024-02-08.

S14 Journal of the American Statistical Association, American Statistical Association. 2007-03.

S37 Cell Press, Patterns; open full text from PubMed Central. 2023-08-04.

S39 Neural Information Processing Systems Foundation, NIPS 2015. 2015.

S01 InTech, International Journal of Engineering Business Management. 2013.

S38 ACM Computing Surveys; open record from Bournemouth University. 2014-04-30.

S40 Google Cloud Documentation. 2024-08-28.

S12 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2006.

S33 Google for Developers. 2025-08-25.

11

Distributions and coherent hierarchies make forecasts usable

A forecast becomes a planning instrument when its uncertainty is calibrated and its component forecasts add up.

CHAPTER 11
DECISION TEST

A forecast becomes a planning instrument when its uncertainty is calibrated and its component forecasts add up.

A single point forecast is an operational trap for a greenfield apparel manufacturer because it conceals asymmetric loss and breaks apart across organisational boundaries. [S08] When a venture with no historical sales commits capital to yarn, greige fabric, dye lots and sewing-line capacity, the cost of ordering too much stock rarely equals the cost of running out during an initial retail test. [S06] Commercial teams plan revenue by channel and region, merchandisers allocate budgets by garment style and colourway, and cutting-room supervisors schedule markers by individual size variant. [S02] If the forecasting engine emits only isolated point predictions at each grain, the sum of the size forecasts will contradict the style total, and the sum of the channel forecasts will contradict the factory production plan. [S17] A forecast becomes a usable planning instrument only when it satisfies two structural conditions: it must express uncertainty through calibrated predictive distributions that expose both central tendency and tail risk, and it must reconcile every disaggregated variant and aggregated total into a single coherent hierarchy. [S14] [S17]

In GFT's Cameroon textile programme, where industrial policy seeks to shift domestic cotton fibre processing from an inherited baseline of about 5 percent toward a national target of 50 percent by 2030, capital discipline at launch is paramount. [R06] [R07] Policy ambition explains why local garment manufacturing matters, but it does not guarantee sell-through for any specific shirt, uniform or dress. [R06] Historical underutilisation in domestic textile plants, including July 2023 capacity utilisation figures of 33.7 percent at Garoua, 48.9 percent at CICAM 1 and 10.83 percent at CICAM 2, alongside an estimated underutilisation cost of about CFAF 2 billion in 2020, illustrates the financial penalty when production capacity, inventory commitments and market absorption drift apart. [R12] Moving from the candidate baselines and global models of Chapter 8, Chapter 9 and Chapter 10 to an operational release requires transforming raw model outputs into calibrated quantile fans and reconciled hierarchical tables before a single production order is signed. [S15] [S17]

11.1

Quantiles represent asymmetric demand risk

A point forecast forces a planner to guess downside and upside exposure, whereas a predictive distribution states those risks explicitly across the range of plausible outcomes. [S14] In early apparel launches, demand uncertainty, seasonal swings and missing history coexist, meaning that weekly demand is frequently skewed rather than symmetric. [S01] A deterministic model predicting an illustrative 150 units a week tells the production controller nothing about whether the 10th percentile sits at an illustrative 120 units or an illustrative 30 units, nor whether a wholesale placement could push the 90th percentile to an illustrative 320 units. [S14] Because manufacturing commitments carry different lead times and reversal costs, a usable cold-start forecast must publish a fixed grid of quantiles rather than a solitary expectation. [S16]

11.2 MULTI-QUANTILE GRIDS SEPARATE CENTRAL SCHEDULING FROM TAIL PROTECTION

The M5 Uncertainty competition established a practical benchmark for retail demand distributions by requiring nine quantiles (`0.005`, `0.025`, `0.165`, `0.250`, `0.500`, `0.750`, `0.835`, `0.975` and `0.995`) to describe future unit sales across hierarchical levels. [S16] Paired around the median (`0.500`), these quantiles form four nested central prediction

intervals covering 50 percent (0.250 to 0.750), 67 percent (0.165 to 0.835), 95 percent (0.025 to 0.975) and 99 percent (0.005 to 0.995) of the distribution. [S16] Adopting this nine-quantile structure inside a greenfield apparel venture prevents teams from inventing ad hoc buffers on top of an unexamined mean. [S14] [S16]

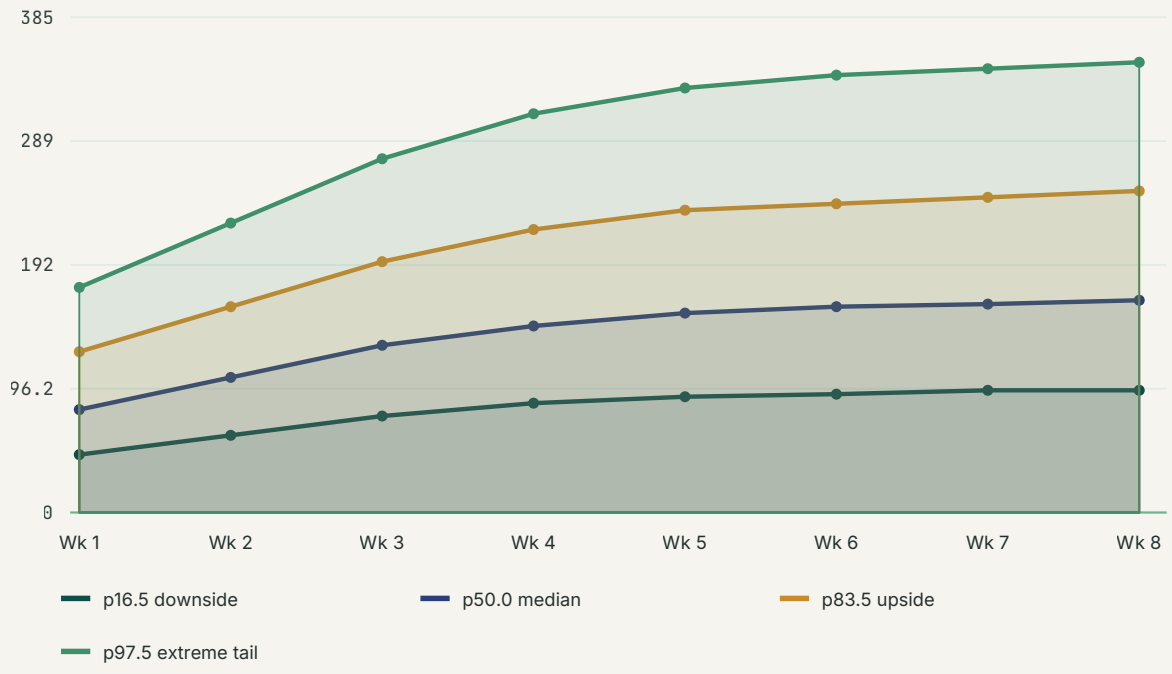
Each region of the quantile grid answers a distinct operational question along the production warp. [S08] Lower-tail quantiles (0.025 and 0.165) define the downside volume that the venture can expect to clear almost regardless of early promotional disappointments, anchoring irreversible raw-material or long-lead trim commitments in an illustrative launch plan. [S16] The central quantiles (0.250, 0.500 and 0.750) bracket the operating regime for sewing-line balancing and standard shift rosters. [S16] Upper-tail quantiles (0.835, 0.975 and 0.995) quantify the surge capacity, greige fabric reservation and conditional dyehouse slots required if early repeat orders accelerate toward the top of the analogue prior. [S16] [S28]

11.3 QUANTILE FANS MUST PRESERVE TEMPORAL EVOLUTION ACROSS THE LAUNCH RAMP

Uncertainty is not static across the life of a new style. [S03] Prior to launch, when the distribution relies on analogue launch curves, market covariates and calibrated choice research, the gap between the 16.5th and 83.5th percentiles is wide because transfer uncertainty dominates. [S28] [S31] As availability-verified pilot observations arrive during the first four to eight weeks, the posterior distribution updates and narrows over near-term horizons while remaining appropriately wide for later seasonal peaks. [S28] **Exhibit 49**, p. 171 illustrates how a multi-quantile fan represents central and tail trajectories across an eight-week launch window without disguising uncertainty as a single line. [S16]

EXHIBIT 49 / ILLUSTRATIVE WEEKLY UNITS

A launch fan exposes central and tail demand pathways across the pilot window



ILLUSTRATIVE [S16] Illustrative weekly quantile trajectories (p16.5, p50.0, p83.5 and p97.5) for a cold-start apparel style across an eight-week launch window, structured after the M5 Uncertainty quantile grid (S16). Values are illustrative planning examples, not observed GFT factory sales.

As **Exhibit 49**, p. 171 shows, presenting four illustrative quantile boundaries (p16.5, p50.0, p83.5 and p97.5) makes the skewness of launch demand legible to both commercial buyers and factory schedulers. [S16] In Week 4 of this illustrative fan, the distance from the median (145 units) down to the p16.5 boundary (85 units) is 60 units, whereas the distance up to the p83.5 boundary (220 units) is 75 units, and up to the p97.5 tail (310 units) is 165 units. What **Exhibit 49**, p. 171 cannot establish by itself is whether those nominal probability labels match reality: an uncalibrated fan is merely a graphic until its empirical coverage is verified on held-out launches. [S15]

11.4

Calibration must be checked rather than inferred from model class

Emitting a probability or quantile from a machine-learning architecture does not guarantee that the stated uncertainty is trustworthy. [S15] Whether a candidate forecast comes from DeepAR, a Temporal Fusion Transformer, Chronos or a hierarchical Bayesian regression, its raw intervals reflect internal model assumptions and training residuals rather than guaranteed cold-start coverage. [S21] [S24] [S25] [S28] Under domain shift between historical analogues and a newly launched garment, raw neural and parametric intervals frequently suffer from overconfidence, producing narrow bands that exclude realised demand far more often than their nominal level promises. [S15] [S38]

11.5 POST-HOC RECALIBRATION AND CONFORMAL METHODS REQUIRE EXPLICIT DATA CONDITIONS

Statistical learning offers two remedies for miscalibrated regression and quantile outputs, each carrying strict operational prerequisites. [S15] [S1101] First, calibrated regression procedures use a held-out calibration dataset to learn an isotonic mapping from predicted cumulative probabilities to empirical cumulative frequencies. [S15] When applied to Bayesian or probabilistic models, this mapping is guaranteed to produce calibrated uncertainty estimates given enough data, but that sufficient-data qualification is decisive in a cold start. [S15] Fitting an isotonic recalibration curve on only three or four pilot items will overfit noise and distort the tails. [S15] [S37]

Second, conformalized quantile regression combines classical quantile regression with split conformal prediction to construct prediction intervals that attain valid finite-sample coverage without distributional assumptions. [S1101] Unlike constant-width conformal methods, conformalized quantile regression adapts to heteroscedasticity by adjusting lower and upper quantile estimates with conformity scores evaluated on a calibration split, producing tighter intervals where signal is strong and wider intervals where volatility is high. [S1101] Even so, finite-sample conformal guarantees rely on exchangeability between calibration launches and the target launch. [S1101] If the target launch introduces a new retail channel, a holiday regime or an unrecorded stockout, exchangeability fails and empirical coverage must be re-verified. [S04] [S06] [S1101]

11.6 GROUP-CONDITIONAL AUDITS EXPOSE HIDDEN FAILURES INSIDE ACCEPTABLE TOTALS

A portfolio-wide calibration score can mask severe miscalibration in the specific segments where capital is most at risk. [S15] If an apparel maker over-covers core institutional basics (achieving an illustrative 92 percent empirical coverage on a nominal 80 percent interval) while under-covering seasonal fashion styles and peripheral sizes (achieving an illustrative 58 percent coverage on the same nominal 80 percent interval), the pooled average may sit near an illustrative 79 percent and appear sound. [S15] [S16] To prevent this aggregation illusion, GFT requires calibration audits to report empirical coverage and interval width separately across product families, channels, size tiers and horizons, as shown in **Exhibit 50**, p. 173. [S15]

EXHIBIT 50 / ILLUSTRATIVE EMPIRICAL COVERAGE (%)

Calibration can fail by product group or channel even when aggregate coverage looks sound



ILLUSTRATIVE [S15] [S1101] Illustrative comparison of empirical coverage (%) across nominal 50%, 67%, 80% and 95% intervals for an aggregate portfolio, core basics, seasonal fashion and sparse tail sizes, demonstrating why group-conditional calibration checks are required alongside portfolio totals (S15, S1101). Values are illustrative.

In **Exhibit 50**, p. 173, the illustrative aggregate portfolio tracks the nominal targets (50, 67, 80 and 95 percent) within one percentage point (49, 66, 79 and 94 percent), passing a naive global check. [S15] Disaggregating the audit reveals that core basics are conservatively wide (54, 71, 84 and 97 percent), whereas seasonal fashion (36, 51, 64 and 81 percent) and sparse tail sizes (32, 46, 59 and 76 percent) suffer severe under-coverage. [S15] [S1101] A planner relying on the nominal 80 percent interval for seasonal fashion would experience misses in 36 percent of weeks instead of 20 percent. Group-conditional coverage tables paired with interval width checks ensure that recalibration is stratified by demand regime. [S15] [S1101]

11.7

Proper scores compare distributions without collapsing them to one error

Evaluating a probabilistic forecast requires metrics that assess the entire predictive distribution honestly. [S14] If a model is judged only on the root mean squared error or mean absolute error of its central tendency, forecasters have no incentive to report

accurate tail probabilities. [S14] Strictly proper scoring rules solve this problem by assigning a numerical score under which a forecaster optimises the expected score if and only if the quoted distribution equals the true underlying distribution. [S14]

11.8 PINBALL LOSS AND CRPS MEASURE CALIBRATION AND SHARPNESS SIMULTANEOUSLY

Probabilistic forecasting aims to maximise sharpness (concentrating probability mass into narrow, informative intervals) subject to calibration (matching stated probabilities to empirical frequencies). [S14] For quantile forecasts, the canonical proper scoring rule is the pinball loss (quantile loss), which penalises an observation `y` falling above a predicted `tau`-quantile `q_tau` by `tau * (y - q_tau)` and penalises an observation falling below `q_tau` by `(1 - tau) * (q_tau - y)`. [S14] [S16] At the median (`tau = 0.500`), pinball loss is proportional to absolute error. [S16] At the `0.975` quantile, an under-forecast error is penalised `39` times more heavily than an over-forecast error of the same size (`0.975 / 0.025 = 39`), pushing the upper quantile upward until only `2.5` percent of realisations exceed it. [S14] [S16]

> **METHOD**
>
> **Operational protocol for multi-quantile scoring and coverage verification**
>
> 1. **Define the target grid:** Fix the nine M5 quantiles (`0.005`, `0.025`, `0.165`, `0.250`, `0.500`, `0.750`, `0.835`, `0.975` and `0.995`) for every active horizon and hierarchical node. [S16]
> 2. **Filter censored observations:** Exclude or flag stockout-censored periods before computing upper-tail scores, because an observed sale capped by zero stock cannot validate an uncensored `0.835` or `0.975` demand quantile. [S06] [S07]
> 3. **Compute pinball loss at each quantile:** For each uncensored realisation `y` and quantile forecast `q_tau`, calculate `L_tau(y, q_tau) = max(tau * (y - q_tau), (tau - 1) * (y - q_tau))`. [S14] [S16]
> 4. **Average across the grid to approximate CRPS:** Average `2 * L_tau` across all nine quantiles to approximate the Continuous Ranked Probability Score (CRPS), scaling by the as-of naive denominator where defined. [S12] [S14] [S16]
> 5. **Verify non-crossing and empirical coverage:** Check that `q(tau_1) ⩽ q(tau_2)` for all `tau_1 < tau_2`, and compute empirical coverage and mean width for each central band (50%, 67%, 95% and 99%) by product family and channel. [S15] [S16]

Averaging pinball loss across the quantile grid links quantile evaluation to the Continuous Ranked Probability Score (CRPS), which integrates the squared difference between the predictive cumulative distribution and the empirical step function of the outcome. [S14] Unlike the logarithmic score, which can degenerate when a discrete count outcome falls in a zero-density region, CRPS and pinball loss remain well-behaved across continuous and discrete count distributions and measure error in the unit of demand. [S14] [S1102]

11.9 POINT METRICS REMAIN BESIDE PROPER SCORES TO PROTECT OPERATIONAL INTERPRETABILITY

Proper distributional scores complement point accuracy metrics rather than replacing them. [S12] [S14] A candidate model could improve average CRPS by tightening its extreme 0.995 tail while introducing a persistent 12 percent illustrative upward bias at the median (0.500), inflating sewing-line schedules if planners schedule to the median. [S12] [S33] Consequently, GFT's evaluation contract places proper distributional scores, empirical calibration checks, point error (MASE and MAE) and coherence diagnostics side by side in **Exhibit 51**, p. 175. [S12] [S14] [S17]

EXHIBIT 51 / DIAGNOSTIC SPECIFICATION

Different distributional and coherence checks govern different launch failure modes

Diagnostic measure	Primary formula or check	Failure mode exposed	Release gate action
Pinball loss (quantile score)	Asymmetric piecewise linear loss at target tau	Miscalibrated tail location or poor sharpness	Compare against prior and baseline at each tau
Empirical interval coverage	Observed share inside nominal central interval	Overconfident narrow bands or excessive width	Recalibrate via isotonic or conformal adjustment
Continuous ranked probability score	Integrated squared difference of cumulative distributions	Global distributional shape and dispersion error	Reject candidate if worse than pooled analogue prior
Scaled point error and bias	MASE where defined, plus MAE and signed percentage bias	Systematic overproduction or underproduction at median	Block release if central bias breaches planning tolerance
Coherence discrepancy	Max absolute residual across hierarchy summing matrix	Component size or channel forecasts fail to sum to total	Apply MinT or sample-path projection before release

DERIVED [S12] [S14] [S15] [S16] [S17] Derived diagnostic contract synthesising strictly proper scoring rules (S14), calibrated regression checks (S15), M5 multi-quantile evaluation (S16), scaled error baselines (S12) and trace-minimising reconciliation (S17).

As **Exhibit 51**, p. 175 establishes, each diagnostic guards against a specific failure mode. [S14] [S17] Pinball loss and CRPS test distributional skill against the analogue prior and baselines, empirical coverage verifies interval honesty, MASE and signed bias protect central capacity planning, and coherence discrepancy checks ensure that component forecasts add up. [S12] [S15] [S17] What **Exhibit 51**, p. 175 cannot do automatically is trade off a small gain in tail pinball loss against a worse central bias; that decision requires human thresholds set under the governance rules of Chapter 12 and Chapter 13. [S33] [S36]

11.10

Reconciliation keeps product, size, channel and market forecasts coherent

Apparel demand spans multiple levels of aggregation. [S11] In the M5 competition, accuracy and uncertainty were evaluated across 42,840 hierarchical series spanning products, departments, categories, stores and states. [S11] [S16] Even for a greenfield manufacturer launching an initial collection of an illustrative 12 to 25 styles in Cameroon, the planning structure forms both a nested hierarchy (total demand split by region, channel and account) and a grouped structure (category crossed with style, colourway and size). [S02] [S11] [S17]

11.11 INDEPENDENT BASE FORECASTS INEVITABLY CONTRADICT EACH OTHER ACROSS GRAINS

When models produce base forecasts independently at the style level, channel level and bottom SKU-size level, those forecasts rarely add up. [S17] [S1102] A global attribute model might forecast 1,000 illustrative units for a woven cotton workshirt in Month 2. [S02] Simultaneously, channel regressions might sum to 1,080 illustrative units across wholesale and direct retail, while bottom-level intermittent size models sum to 920 illustrative units. [S04] [S18]

Published without reconciliation, these numbers cause every department to act on a different plan: procurement orders cloth for 1,000 shirts, sales allocates 1,080 shirts, and the cutting room lays markers for 920 shirts. [S17] Pure bottom-up aggregation sums noisy size series and ignores style-level attribute signals, whereas pure top-down disaggregation relies on static historical proportions that do not exist in a cold start and cannot capture how size mix differs between institutional wholesale and online retail. [S17] [S1102]

11.12 MINIMUM TRACE RECONCILIATION COMBINES INFORMATION ACROSS EVERY LEVEL

Optimal forecast reconciliation generates base forecasts at all levels and projects them onto the coherent linear subspace where aggregation constraints hold exactly. [S17] [S1102] Let `y_t` denote the vector of all `n` series at time `t`, and `b_t` the `m`-dimensional vector of bottom-level series, linked by an `n x m` summing matrix `S` so that `y_t = S * b_t`. [S17] [S1102] Base forecasts `y_hat` map into coherent forecasts `y_tilde = S * G * y_hat` via a matrix `G` satisfying `S * G * S = S`, preserving unbiasedness. [S17]

The Minimum Trace (MinT) estimator chooses `G` using the `h`-step-ahead base forecast error covariance matrix `W_h`, minimising the trace of the coherent error covariance matrix (and thus total mean squared error under unbiasedness). [S17] Because estimating an unconstrained matrix `W_h` is infeasible with short launch histories, cold-start implementations use structural scaling (weighting each node by the count of bottom series it sums) or shrinkage toward a diagonal variance matrix. [S17]

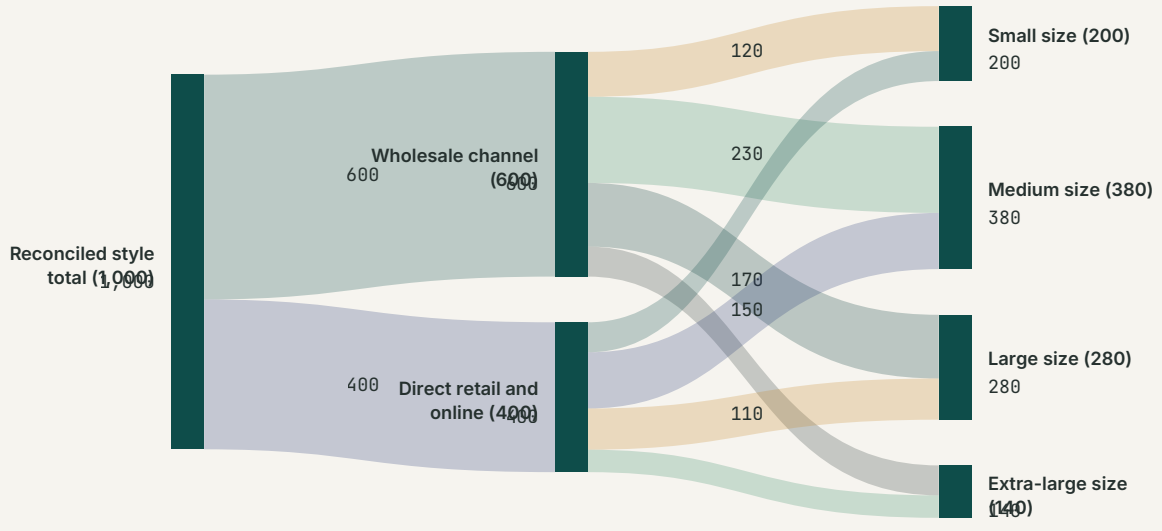
11.13 PROBABILISTIC RECONCILIATION MUST PRESERVE COHERENT SAMPLE PATHS AND QUANTILES

Point coherence does not guarantee quantile coherence because quantiles are not additive: the 90th percentile of a style total does not equal the sum of the 90th percentiles of its individual sizes unless those sizes are perfectly correlated. [S1102] [S1103] Summing the 95th percentile of every size variant to size a fabric buffer will substantially overstate the 95th percentile of style-level fabric demand. [S1102]

To reconcile full distributions, GFT uses a two-step sample-path procedure. [S1102] [S1103] First, the engine draws B joint future sample paths (for example, an illustrative 1,000 trajectories) from the calibrated base distributions at all nodes, using empirical copula reordering or Bayesian posterior draws to retain rank correlations across sizes and channels. [S28] [S1102] Second, the engine applies the MinT projection S * G to each sample vector, yielding B coherent sample paths in which every trajectory adds up from bottom size-channel cells to style, channel and market totals. [S1102] [S1103] Under Gaussian errors, MinT projection also minimises the multivariate logarithmic score and beats ordinary least squares projection on marginal logarithmic scores. [S1102] Empirical studies confirm that probabilistic reconciliation improves both point and distributional accuracy across hierarchies. [S1103] **Exhibit 52**, p. 177 shows how a reconciled 1,000-unit style forecast conserves volume across channels and sizes. [S17]

EXHIBIT 52 / ILLUSTRATIVE RECONCILED UNITS

Coherent reconciliation preserves volume conservation from style total to channels and sizes



ILLUSTRATIVE [S11] [S17] Illustrative flow conservation across a grouped apparel hierarchy after MinT reconciliation (S17), mapping a 1,000-unit monthly style forecast through two sales channels (wholesale 600, direct retail 400) into four size groups (Small 200, Medium 380, Large 280, Extra-Large 140). Values are illustrative.

As **Exhibit 52**, p. 177 demonstrates, the 1,000 illustrative units at the style node equal the sum of the wholesale (600 units) and direct (400 units) channels, and each channel's size allocations sum to the cross-channel size totals (200 Small, 380 Medium,

280 Large and 140 Extra-large). [S17] What **Exhibit 52**, p. 177 cannot correct is systematic bias in the base inputs: because MinT assumes unbiased base forecasts, stockout censorship must be corrected before projection. [S06] [S17]

WORKED EXAMPLE: RECONCILING AN INCOHERENT LAUNCH STYLE ACROSS CHANNELS AND SIZES

Consider an illustrative cold-start launch of a cotton uniform shirt in Month 2, where three modelling grains produce conflicting base forecasts:

1. **Incoherent base inputs (illustrative):**
2. **Top level (Style total):** A pooled analogue and attribute regression predicts an illustrative mean of 1,000 units (central 80 percent interval 760 to 1,260 units). [S02]
3. **Middle level (Channels):** Separate channel priors predict an illustrative 650 units for wholesale and 430 units for direct retail, summing to 1,080 units (+80 units above the style total). [S04]
4. **Bottom level (Eight channel-by-size cells):** Individual size models across Small (S), Medium (M), Large (L) and Extra-large (XL) predict illustrative volumes of [110, 210, 155, 75] in wholesale (sum 550) and [70, 130, 95, 55] in direct retail (sum 350), totalling 900 units (-100 units below the style total). [S18]

1. **Structural WLS / MinT reconciliation arithmetic (illustrative):**
2. Under structural variance scaling (where base error variance is proportional to the number of bottom cells summed: 8 for the style total, 4 per channel, and 1 per bottom cell), linear projection balances the three levels inversely to their scaled error variances. [S17]
3. The reconciled total settles at 1,000 illustrative units, channel totals adjust from [650, 430] to [600, 400] units, and bottom cells adjust to [120, 230, 170, 80] in wholesale and [80, 150, 110, 60] in direct retail, matching **Exhibit 52**, p. 177. [S17]

1. **Probabilistic quantile aggregation check (illustrative):**
2. Summing the separate illustrative 90th percentiles of the eight bottom cells ([165, 305, 230, 115] in wholesale and [115, 205, 155, 88] in direct retail) would yield 1,378 units. [S1102]
3. Projecting 1,000 correlated Monte Carlo sample paths through the reconciliation matrix yields a coherent style-level 90th percentile of 1,240 illustrative units (138 units below the naive sum), with coherent size 90th percentiles of 255 (S), 475 (M), 355 (L) and 185 (XL) units. [S1102]
4. The fabric buyer reserves greige cloth for 1,240 illustrative units rather than 1,378, avoiding a 138-unit over-order. This calculation is an illustrative method demonstration, not an observed GFT result. [S17]

11.14

Sparse count demand resists false continuous precision

While aggregated style and channel totals behave smoothly once weekly volume exceeds several dozen units, bottom-level apparel series (such as a specific colour-way in size Double-Extra-Large) are almost always sparse and intermittent. [S18] Intermittent demand occurs sporadically, with many periods recording zero demand interspersed with occasional small purchases. [S18] Applying continuous Gaussian intervals or unadjusted linear projection to sparse series creates immediate distortions: lower intervals turn negative, central quantiles take fractional values such as `0.37` shirts, and standard estimators overstate average demand. [S18] [S19]

11.15 OCCURRENCE AND SIZE MUST BE SEPARATED AND CORRECTED FOR BIAS

Croston's method recognised that intermittent demand requires modelling two components: the probability of nonzero demand occurrence (or inter-demand interval) and the conditional demand size when an order occurs. [S18] As Syntetos and Boylan proved, combining separate exponentially smoothed estimates of demand size and inter-demand interval by simple division introduces a positive inversion bias that systematically over-forecasts demand per period. [S19] The Syntetos-Boylan Approximation (SBA) corrects this bias by multiplying the ratio by `(1 - alpha / 2)`, where `alpha` is the interval smoothing parameter. [S19]

In a probabilistic hierarchy, this separation governs bottom-level quantile generation. [S18] [S21] Rather than fitting continuous Gaussian likelihoods to zero-heavy variants, global probabilistic models such as DeepAR or hierarchical Bayesian count models use discrete likelihoods (negative binomial, Poisson or zero-inflated hurdle distributions) that place positive mass at zero and restrict quantiles to non-negative integers. [S21] [S28] When sample paths from discrete bottom distributions are projected onto the coherent subspace, negative values are resolved through non-negative constrained reconciliation or integer rounding so planners never see negative size allocations. [S17] [S1102]

11.16 AGGREGATION ACROSS TIME AND PRODUCTS CHANGES THE INTERMITTENT SIGNAL

Sparsity also depends on the temporal grain of the decision. [S1103] A peripheral size variant recording zeros on four out of five days looks smoother when aggregated into four-week production buckets. [S18] Temporal hierarchies reconcile forecasts across time frequencies (weekly replenishment, four-week sewing cycles and quarterly fabric windows) using the same projection logic applied to cross-sectional product hierarchies. [S1103] Whenever sparse variants feed a style or channel total, GFT's release reports both the Syntetos-Boylan bias check at the bottom grain and the sensitivity of the reconciled total to pooling the sparsest cells. [S18] [S19] [S1103]

11.17

Scenario overlays remain distinct from statistical intervals

Even a calibrated, coherent predictive distribution describes uncertainty only within its conditioning assumptions and reference class. [S08] Statistical prediction intervals state what range of weekly demand to expect at a chosen probability level given the analogue pool, pilot transactions and current offer structure. [S08] [S14] They do not state what follows if a commercial or policy condition changes outside that regime. [S08]

11.18 STRUCTURAL REGIME CHANGES BELONG IN NAMED SCENARIOS, NOT INFLATED INTERVALS

Greenfield apparel makers regularly confront discrete contingencies: a ministry or corporate buyer may award an institutional uniform tender, a port delay may hold up imported trims for six weeks, or management may test a 15 percent illustrative promotional discount. [S01] [S08] Treating these discrete events as random variance inside a 95 percent statistical interval corrupts both the model and the planning review. [S08] [S36]

Artificially widening a prediction interval to cover a hypothetical uniform contract destroys its empirical calibration meaning and prevents honest scoring with pinball loss or CRPS. [S14] [S15] Conversely, relying solely on the statistical fan leaves the venture exposed to regime shifts. [S08] Following forecasting theory and the NIST AI Risk Management Framework, GFT separates model-based statistical distributions from human-authored scenario overlays, assigning every scenario a named owner, a conditioning trigger and an explicit review date. [S08] [S36]

GFT NOTE

Keep statistical quantiles and commercial scenarios in separate layers across GFT planning tools

In GFT's Cameroon textile programme and its companion planning tools (`cold-start-forecast`, `ramp-planner` and `safety-stock`), planners must never overwrite a calibrated quantile fan to reflect a commercial target or policy ambition. [S08] [S36] Specifically: 1. **Base predictive distribution:** Generated by the approved prior or model, calibrated on held-out launch replays, reconciled across the hierarchy and scored automatically against realised uncensored demand. [S15] [S17] 2. **Conditional scenario overlay:** Stored as a separate delta layer with an explicit trigger condition (for example, an illustrative institutional uniform tender of `3,000` pieces in Quarter 3 or an illustrative `10` percent price cut), a named commercial owner and an expiry date. [S08] [S36] 3. **Policy boundary:** National industrialisation benchmarks, such as Cameroon's 50 percent cotton processing target by 2030 or announced third-party capacity such as the Camtext SA project (`12,000` tonnes a year processed, announced 12 November 2025), are macro supply context and must never be entered as demand scenarios or sales priors for a garment line. [R07] [R09] [R10]

Detailed working-capital, safety-stock and reorder-point optimisation using these calibrated distributions is addressed in `wp12`, while this chapter fixes the distributional contract that feeds those calculations. [S08]

11.19

Every distribution carries provenance, version and expiry

A probability distribution is reproducible and auditable only when the exact state of the data, priors, calibration mappings and reconciliation matrices that produced it travels with the numbers. [S37] [S40] When forecasts circulate as detached spreadsheets, nobody can reconstruct two weeks later whether the intervals used pre-pilot analogue weights, whether stockout days were masked, or which calibration window governed interval widths. [S06] [S37]

11.20 AS-OF METADATA PREVENTS HINDSIGHT LEAKAGE AND SILENT DRIFT

Data leakage is a spurious relationship between inputs and the target arising from collection, sampling or preprocessing artefacts, and it is as hazardous in calibration and reconciliation as in feature engineering. [S37] Estimating a reconciliation covariance matrix `W_h` or a conformal quantile adjustment using records from Weeks 5 to 8 and applying it to a launch replay at Week 2 contaminates coverage and coherence metrics with future information. [S15] [S17] [S37]

To prevent retrospective leakage and operational confusion, GFT wraps every published forecast distribution in an immutable provenance envelope, shown in **Exhibit 53**, p. 182. [S37] [S40] Each release package records four metadata layers: the as-of provenance layer (training cut-off timestamp, feature snapshot hash and stockout censorship mask flags), the probabilistic base layer (model ID, analogue prior shrinkage weight and discrete versus continuous family), the coherence and calibration layer (calibration method, sample size, group-conditional coverage status and reconciliation matrix hash), and the planning layer (decision horizon, separate scenario IDs and a hard expiry timestamp). [S15] [S17] [S28] [S37] [S40]

EXHIBIT 53 / SYSTEM GOVERNANCE LAYERS

A governed forecast distribution wraps raw quantiles in calibration, reconciliation and provenance layers



DERIVED [S15] [S17] [S37] [S40] Derived release architecture for cold-start probabilistic forecasts, combining post-hoc calibration checks (S15), trace-minimising hierarchical reconciliation (S17), as-of leakage controls (S37) and MLOps monitoring and rollback metadata (S40).

As **Exhibit 53**, p. 182 shows, a planner inspecting a forecast in `cold-start-forecast` or `ramp-planner` never receives uncontextualised model outputs. [S40] Every figure traces through the reconciliation and calibration layer to the base model and its as-of cut-off. [S15] [S17] [S37] When calibration sample support falls below the minimum independent launch count required by Chapter 12, the envelope explicitly flags the distribution as prior-dominated rather than implying certified empirical coverage. [S15] [S39]

WARNING

Unmasked stockouts and stale calibration windows silently corrupt tail quantiles

Two operational failure modes routinely invalidate probabilistic apparel forecasts during early launch months: 1. **Censored upper tails:** When a popular size or colour-way stocks out mid-week, recorded sales freeze while latent demand continues. [S06] [S07] If unmasked stockout weeks enter the quantile loss or conformal calibration split, the model learns that demand plateaus at the inventory cap, compressing the 0.835, 0.975 and 0.995 quantiles and causing repeated under-ordering. [S06] [S07] Resolution requires applying the availability masks from Chapter 7 before computing conformity scores or reconciliation residuals. [S07] 2. **Stale calibration under seasonal or channel drift:** A calibration mapping fitted during a quiet pilot month will under-cover demand when a holiday promotion begins or distribution expands to regional wholesale accounts. [S04] [S38] Resolution requires automated monitoring of empirical coverage by week and channel, with automatic reversion to wider analogue prior intervals whenever rolling coverage breaches the approved floor. [S38] [S40]

11.21 A DISTRIBUTION RELEASE GATE PREPARES THE HAND-OFF TO BACKTESTING AND SERVICE OPERATIONS

Structuring every cold-start forecast as a calibrated nine-quantile distribution reconciled across sizes, styles and channels eliminates the two most common excuses for ignoring quantitative evidence at launch: that a point forecast was inevitably wrong, or that the model's totals contradicted the merchandiser's size breakdown and the commercial director's channel plan. [S16] [S17] The statistical distribution states what current evidence supports and how wide residual doubt remains; the reconciliation layer ensures every operational table adds up; the scenario layer records what human leaders wish to test beyond the historical regime; and the provenance envelope preserves the exact information set for retrospective scoring. [S08] [S14] [S17] [S37] Chapter 12 now specifies how these probabilistic, hierarchical forecasts are tested in chronological launch replays without hindsight leakage, and Chapter 13 shows how planners govern, challenge and roll back the resulting forecast service inside a monthly operating cadence. [S12] [S33] [S40]

A single number hides the bet; a calibrated, coherent distribution shows every planner the exact risk on the table.

CHAPTER SOURCE NOTES

S08 International Journal of Forecasting, Elsevier; full text through arXiv. 2022.

S06 Lancaster University Management School, Munich Personal RePEc Archive. 2018-11-19.

S02 arXiv. 2019-06-27.

S17 Journal of the American Statistical Association, Taylor & Francis; abstract through RePEc IDEAS. 2019.

S14 Journal of the American Statistical Association, American Statistical Association. 2007.

R06 Cameroon Trade Hub. Publication date unrecorded; inherited source-text check.

R07 Ecofin Agency. Publication date unrecorded; inherited source-text check.

R12 Business in Cameroon. Publication date unrecorded; inherited source-text check.
S15 Proceedings of the 35th International Conference on Machine Learning, PMLR. 2018-07-03.
S01 InTech, International Journal of Engineering Business Management. 2013.
S16 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2022.
S28 Journal of Industrial Ecology, Wiley. 2022-10-17.
S03 arXiv. 2020-07-10.
S31 INFORMS, Marketing Science; open manuscript from MIT. 2020.
S21 arXiv; journal version in International Journal of Forecasting, Elsevier. 2019-02-22.
S24 arXiv. 2020-09-27.
S25 arXiv. 2024-11-04.
S38 ACM Computing Surveys; open record from Bournemouth University. 2014-04-30.
S1101 arXiv; Neural Information Processing Systems (NeurIPS 2019). 2019-05-08.
S37 Cell Press, Patterns; open full text from PubMed Central. 2023-08-04.
S04 MDPI, Processes. 2021-09-03.
S07 arXiv. 2026.
S12 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2006.
S1102 arXiv; Journal of Business & Economic Statistics. 2021-03-20.
S33 Google for Developers. 2025-08-25.
S36 National Institute of Standards and Technology, U.S. Department of Commerce. 2023.
S11 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2022.
S18 International Journal of Forecasting, Elsevier; abstract through Lancaster EPrints. 2005.
S1103 arXiv; European Journal of Operational Research, Elsevier. 2018-08-08.
S19 International Journal of Production Economics, Elsevier; abstract through Lancaster EPrints. 2001.
R09 Ecofin Agency. Publication date unrecorded; inherited source-text check.
R10 Ecofin Agency. Publication date unrecorded; inherited source-text check.
S40 Google Cloud Documentation. 2024-08-28.
S39 Neural Information Processing Systems Foundation, NIPS 2015. 2015.

GREEN FUTURE TECH

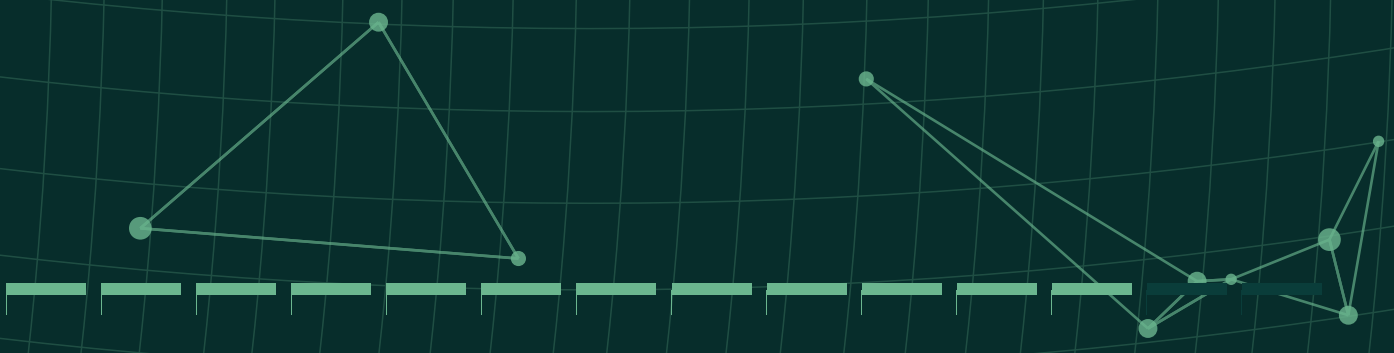
LIMITED / HONG KONG SAR

12

PART IV: PROVE VALUE AND GRADUATE WITH EVIDENCE / VALIDATION DESIGN

Cold-start backtests must replay the information available at launch

The only fair test is one that denies the model every piece of information it would not have had at the decision moment.



CHAPTER 12
DECISION TEST

The only fair test is one that denies the model every piece of information it would not have had at the decision moment.

A cold-start backtest is a rehearsal of ignorance. It must remove the target launch from training, reconstruct the evidence that existed at each decision origin and score only outcomes that were still unknown at that origin. This rule applies to product attributes, sales, stock records, external series, analogue selection, demand correction, preprocessing, calibration and forecast combination. A later row split is not enough when the claim concerns an unseen product. The valid test holds out whole launches, repeats several launch origins and compares every candidate with the same frozen baselines. Release follows only when accuracy, distribution quality, planning consequences and operating fitness survive that rehearsal with adequate launch support.

This design measures the capability the venture actually intends to use. Before launch, a model may know approved product attributes and earlier evidence, but no target sales. After the first selling interval, it may know only the transactions and availability records that had arrived by then. An early-repeat forecast receives another bounded information set. Each condition needs its own replay and its own artefact versions. In Cameroon, where the inherited sector baseline records about 95 percent of cotton fibre exported raw against about 5 percent processed locally and national policy restates a 50 percent local-processing target by 2030, policy ambition explains why new apparel capacity matters, not how many garments a launch will sell. [R06] [R07] The resulting backtest score is not a timeless property of an algorithm or investment, legal or tax advice. It is evidence about a specified forecasting service, evaluated on a stated launch population, under stated information constraints.

12.1

Rolling origins preserve chronology

A forecast origin is the instant at which the information set closes and the forecast becomes fixed. Everything used to produce that forecast must have been available by then. Everything scored must belong to its future. Expanding-window evaluation formalises this direction by training on successive historical supersets and testing later observations. Scikit-learn's time-series documentation contrasts these successive training sets with ordinary cross-validation and warns that time-dependent samples need a time-aware scheme. [S13] Empirical comparison of time-series performance estimation methods likewise indicates that when non-stationary variation is present in real-world series, out-of-sample methods that preserve the temporal order of observations produce the most accurate loss estimates. [S1201]

The cold-start harness applies that principle at two levels. Across the reference corpus, later origins may add completed launches that had become available. Within each target launch, later origins may add only the target observations that had arrived by the new cut-off. The two expansions are recorded separately. Otherwise, a backtest can appear chronological within one sales series while silently using a reference launch that did not yet exist when the original decision was made.

Several origins are necessary because one cut-off can be unusually kind to a method. A pre-launch origin tests transfer from attributes and analogues. An origin immediately after the first selling interval tests the first update from owned evidence. A later origin tests whether the model uses a short ramp without allowing that ramp to

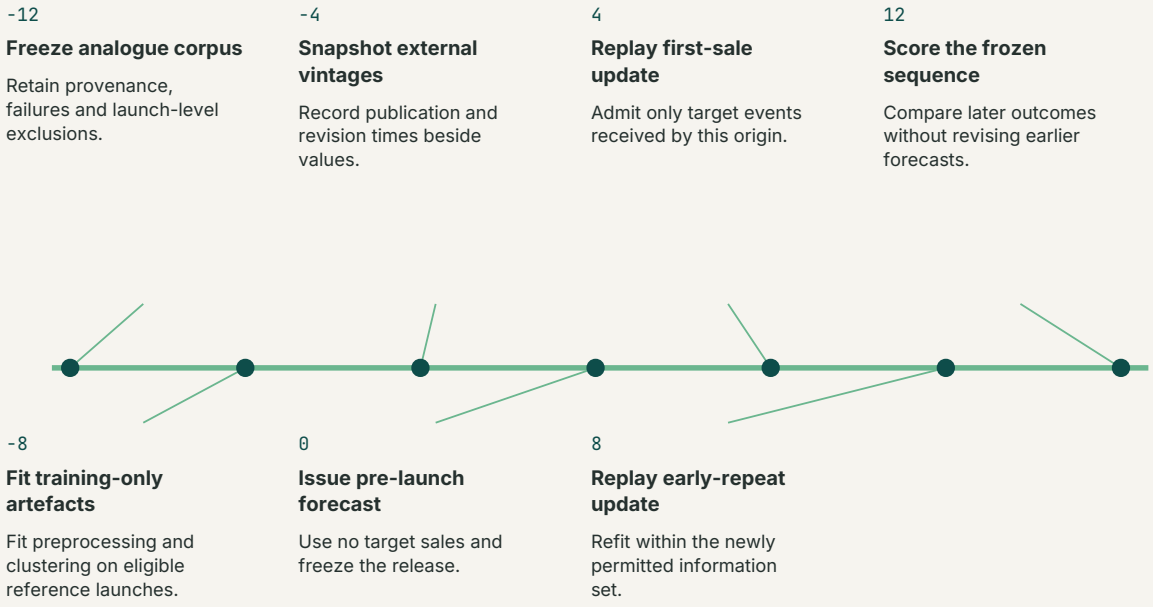
dominate its earlier prior. These are different forecasting tasks, not repeated measurements of one task. Their scores remain separate before any aggregate is considered.

The origin register therefore stores the decision deadline, data cut-off, forecast horizon, eligible reference launches, target-history length and expected publication lag for every input. Known future fields receive special treatment. A planned price or approved campaign date may be legitimate when genuinely known at the origin. The realised sales response to that campaign is not. A revised market statistic is represented by the vintage available at the cut-off, not by the cleanest version downloaded after the launch.

Exhibit 54, p. 187 gives an illustrative sequence for one wholly held-out launch. Its numeric positions are weeks relative to launch, not a recommended planning calendar. The important feature is the repeated closure and freezing of evidence. The target launch never enters fitting for that replay, while each later forecast is rebuilt only from the information then permitted.

EXHIBIT 54 / ILLUSTRATIVE WEEKS RELATIVE TO TARGET LAUNCH

Launch replay separates every training and decision cut-off



ILLUSTRATIVE [S02] [S13] [S1201] Illustrative chronology derived from unseen-item and expanding-window principles. The target launch is excluded from every fitting stage in this replay. Positions are planning assumptions, not observed dates or recommended intervals. Actual origins must follow the venture's decision and data-arrival clocks.

The sequence also prevents retrospective model shopping. A team that tries several origins and reports only the best has created another selection bias. The origin set belongs in the evaluation protocol before candidates are compared. Any additional

origin is appended under a documented rule, with all prior results retained. Where an origin lacks the evidence required for a legitimate operational forecast, the result is an ineligible case with a reason, not a quiet deletion.

> **METHOD**
>
> Create one origin manifest per launch and decision. Freeze its cut-off, horizon, target-history allowance, reference-launch allowance and external-data vintages before fitting. Run the complete candidate and baseline pipeline at every manifest. Store forecasts immediately, then reveal outcomes only for scoring. A changed cut-off creates a new manifest; it never edits the old one.

12.2

Held-out products test the unseen-item claim

An unseen-item forecast must be tested on an item whose identity and outcomes were absent from training. Attribute-based fashion research builds generalised models from other products' attributes and merchandising factors, so its relevant generalisation question is whether those relationships transfer to a new item. [S02] A random row split cannot answer it. Later rows for a product already represented in training test temporal continuation for a known item, which is a different and usually easier deployment condition.

Product exclusion must cover every route through which target demand could return. Obvious routes include sales rows, returns and target-derived aggregates. Less obvious routes include normalisation statistics, selected features, demand corrections, analogue weights and hyperparameter choices. Even a target identifier can allow a flexible model to memorise historical behaviour rather than learn transferable attribute relationships. The split therefore occurs before any target-dependent transformation, not after a convenient feature table has been built for the whole corpus.

Visual transfer needs an additional exclusion boundary. In the opened cold-start fashion architecture, new products are matched to clusters and receive corrected demand signals from similar historical products. [S05] In a faithful replay, the target product cannot influence those corrected signals, cluster centroids, neighbour rankings or the choice of cluster count using its own outcomes. Its image and approved static attributes may enter at inference. Its demand history may not shape the reference structure that supposedly predicts it.

There is also a difference between leaving out a product and leaving out an independent launch group. Colour and size variants can share photography, a campaign, inventory and customers. Several launches can share one institutional order or channel event. Treating all related variants as independent holdouts can make transfer look stronger than the deployment evidence warrants. The group definition should follow these dependencies and remain fixed before modelling. Where independence is doubtful, the conservative choice is to hold out the wider group and report the reduced support.

Chronology still applies inside the held-out group. A product launched after the forecast origin cannot become an analogue merely because its full history exists in the evaluation file. The eligible training pool is the intersection of two rules: the product group is permitted, and its information was available by the origin. This joint rule blocks both entity leakage and temporal leakage without reducing the test to a single convenient split.

Finally, product holdouts need a role distinction. Validation launches may guide model and hyperparameter selection under a frozen procedure. Final test launches remain untouched until the selected pipeline is run once for the release readout. Reusing the final products to choose the winner turns them into validation evidence and removes the independent check. If the corpus is too small for separate roles, the honest result is preliminary evidence with wider uncertainty, not a claim of confirmed generalisation.

12.3

Held-out launches test the entire ramp

A product holdout becomes a launch replay only when the whole launch history is removed together. The test then reproduces the pre-launch prior, the first-sale update and the early-repeat update in sequence. It does not select a convenient later window from an otherwise known product. This extends unseen-product evaluation and chronological splitting into a single deployment rehearsal. [S02] [S05] [S13]

The evaluation corpus is frozen before candidates are compared. Its register identifies every launch, its source and access right, extraction time, product and channel attributes, complete observation window, availability rule, correction version and exclusion reason. Failed, withdrawn and slow launches remain visible when they belong to the deployment population. Removing them because their outcomes are awkward would change the population after seeing the answer and would overstate the quality of launch transfer.

Pilot structure is part of that provenance. Where a pilot assigned customers, locations or time cells to different offers, those cells remain intact in the replay. The evaluation cannot move a favourable treatment cell into training while leaving its control counterpart in the test set and still claim to assess the original experiment. Controlled-experiment guidance supports randomised designs for causal interpretation; preserving assignment is necessary to retain that meaning in a later forecast evaluation. [S32]

The launch register also records what was decided rather than only what was sold. Release timing, initial lot, price, promotion, listing intervals, stockouts and replenishment can all alter the observed ramp. These fields let the reviewer distinguish a forecasting failure from a selling condition that the model could not have known. They also prevent a later, tidier reconstruction of the launch from replacing the conditions under which evidence was actually earned.

Support is reported by independent launch group, not only by rows. Many weekly observations from one launch provide limited independent proof that a method transfers to another launch. Related variants can reduce effective support further. Every product, channel and horizon cell therefore carries its launch count, observation

count, dependency grouping and eligibility count for each metric. No universal minimum is asserted here. The release owner sets a defensible requirement before seeing candidate results, using the decision risk and the structure of the available corpus.

Sparse cells remain preliminary. They are not pooled into a broader cell merely because the broader result looks more settled, and they are not assigned a precise confidence statement that their support cannot sustain. The readout can show their observed errors with a clear limitation, then specify what additional launches would change the state. If independent support is too weak for the intended release, the approved outcome is to retain the prior or baseline and collect more evidence.

> **WARNING**
>
> A frozen corpus does not remain frozen if the modelling team can edit exclusions after inspecting scores. Store the corpus manifest and exclusion reasons with the evaluation run. Any correction creates a new corpus version and reruns every candidate. Never describe two differently filtered corpora as one head-to-head comparison.

12.4

As-of audits block feature and preprocessing leakage

Leakage is not limited to placing future sales in a feature column. Kapoor and Narayanan define it as a spurious relationship between predictors and target created by collection, sampling or preprocessing, and their taxonomy includes contaminated train-test separation, illegitimate features and mismatched evaluation distributions. [S37] A cold-start audit must therefore inspect the entire evidence path, from event capture to the scored forecast.

The first check concerns the target itself. If observed sales are corrected for stockouts, the correction must use only information permitted at the relevant stage. A correction fitted on the full launch can encode later demand, substitutions or availability into the training target. The same applies to return allocation and cancellation rules. Target derivation is versioned and fitted inside the training fold, then applied under a frozen rule. Where no defensible correction is available, the affected observations are marked censored or excluded under the protocol rather than filled with invented demand.

The second check concerns every learnt artefact around the model. Imputation values, scaling parameters, feature selection, outlier rules, visual embeddings adapted to the reference catalogue, analogue clusters, combination weights and calibration maps are all fitted on training evidence only. A pipeline is not clean merely because the final estimator receives no target rows. Earlier stages can already have transferred the answer.

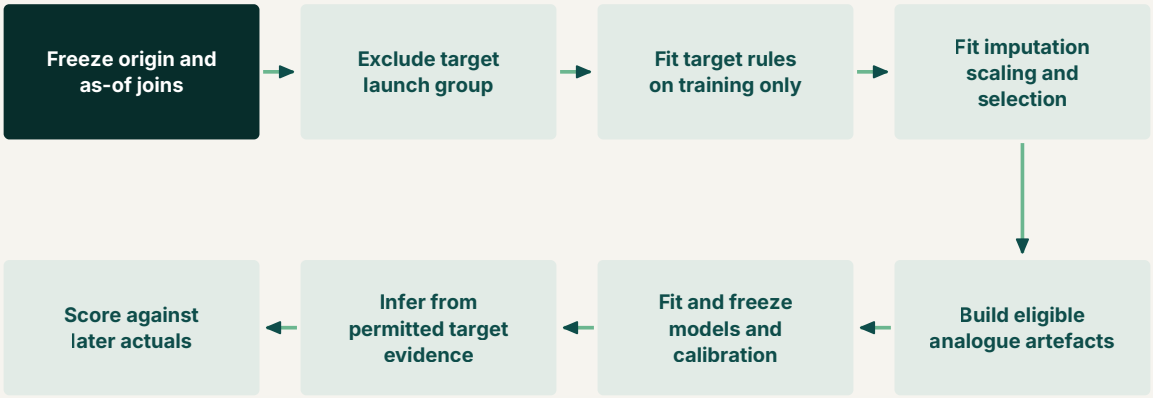
The third check concerns time. Each feature carries an event time, an available time and, where relevant, a revision or ingestion time. Event time alone is insufficient for externally published data and late-arriving operational records. A trade value for an earlier month can still be unavailable at the forecast origin if publication occurred

later. An updated attribute can replace the description customers actually saw. The as-of join selects the version whose available time precedes the cut-off, not the record whose nominal period merely looks earlier.

Exhibit 55, p. 191 places these controls in one derived GFT sequence. It is a proposed validation architecture, not an implemented client system. The central rule is that fitting finishes before the held-out launch is transformed or scored. Later actuals enter only at the final scoring node.

EXHIBIT 55 / VALIDATION CONTROL SEQUENCE

Every learnt transformation belongs inside the replay fold



DERIVED [S13] [S37] Derived GFT audit sequence from chronological evaluation and leakage taxonomy. The flow covers target derivation, preprocessing, analogue construction, calibration and combination. It is a control design, not evidence of an implemented pipeline or measured performance.

Two practical probes make the audit testable. First, a deliberate canary feature stamped after the origin should cause the information gate to reject the fold. Second, rerunning a frozen fold should reproduce its forecasts and transformations from stored artefacts rather than from the current database. These probes test the harness, not forecast skill. They are still necessary because an elegant score produced by an unverifiable pipeline cannot support release.

> **GFT NOTE**
>
> GFT recommends an as-of ledger beside every model result. For each input, record source event time, available time, revision, transformation and training-fold owner. Fit every learnt transformation inside the fold and test the gate with a deliberately future-stamped canary. If the ledger or canary test fails, stop the comparison and repair the harness before reading model rankings.

12.5

Scaled errors make heterogeneous series comparable

Launch forecasts cover products with different volumes and error units. MASE addresses comparability by scaling absolute forecast error with a naive benchmark, and Hyndman and Koehler propose it for accuracy comparisons across multiple time series. [S12] In launch replay, the scaling denominator must be computed only from information available at the forecast origin. It must also retain the same seasonal and availability definitions as the forecast contract.

A pre-launch target often has no valid naive denominator. A very short history can make it unstable, and an all-zero eligible history can make it undefined. The harness must not replace an absent denominator with one, drop the inconvenient launches or calculate the denominator using future target sales. MASE is reported only for eligible cases, with the eligible-launch count and denominator diagnostics beside the result.

Where MASE is unavailable, the readout retains MAE and signed bias in demand units. It also reports aggregate relative absolute error against the frozen analogue baseline: sum candidate absolute error divided by sum baseline absolute error over the same eligible forecast cases. This is GFT arithmetic on paired outputs, not a replacement definition of MASE. A zero baseline denominator is disclosed and prevents the ratio. Unscaled errors remain visible so that a relative gain cannot hide operationally large misses.

Exhibit 56, p. 192 supplies an illustrative paired readout for three pre-launch forecast cases. Actuals and forecasts are constructed garments in one forecast period, not observations or a Cameroon apparel result. Signed error is actual minus forecast, so a negative value denotes overforecasting. The example shows why absolute error, relative error and bias must coexist.

EXHIBIT 56 / ILLUSTRATIVE GARMENTS PER FORECAST PERIOD

Paired launch errors expose skill and overforecasting together

Launch case	Actual	Analogue forecast	Candidate forecast	Baseline signed error	Candidate signed error
A	120	100	110	20	10
B	80	110	95	-30	-15
C	60	50	75	10	-15

ILLUSTRATIVE [S12] Illustrative constructed inputs for the worked calculation, not observed launches. Signed error equals actual minus forecast. Pre-launch MASE is undefined because no eligible target-series naive denominator exists. The analogue baseline is frozen before candidate comparison.

WORKED EXAMPLE: RELATIVE ERROR DOES NOT ERASE BIAS

For the illustrative cases in **Exhibit 56**, p. 192, baseline absolute errors are 20, 30 and 10 garments, giving MAE of 60 divided by 3, or 20.00 garments. Candidate absolute errors are 10, 15 and 15, giving MAE of 40 divided by 3, or 13.33 garments. Aggregate relative absolute error is 40 divided by 60, or 0.667, equivalent to an illustrative reduction of 33.3 percent. These are constructed arithmetic results, not expected model performance.

The illustrative baseline signed errors sum to zero, while candidate signed errors sum to minus 20 garments, an average of minus 6.67. The candidate has lower absolute error but systematic overforecasting in this tiny example. Pre-launch MASE remains undefined because no target-history denominator is eligible. A release review would need more independent launches, horizon and channel cuts, and decision consequences before treating the relative result as useful evidence.

Aggregation needs the same care as the formula. An average of per-launch ratios gives each eligible launch equal weight; a ratio of summed errors gives greater influence to high-volume cases and large misses. Neither is universally correct. The protocol names the estimand before the run and shows both where they answer different planning questions. Error distributions and signed bias by horizon then reveal whether a central summary conceals repeated underproduction at the commitment horizon or overproduction during the early ramp.

12.6

Distribution scores sit beside point errors

A useful central forecast can still carry a damaging distribution. It may place too little probability on a large order, issue intervals that rarely contain realised demand or produce broad ranges that cover outcomes without helping a planner choose. Proper scoring rules assess the predictive distribution against the event that materialises and encourage honest probabilistic assessments. [S14] The release harness therefore evaluates the complete published distribution, not only its median or mean.

Quantile loss is retained for every published quantile and horizon. A distributional score such as the continuous ranked probability score can summarise the full cumulative forecast where the output contract supports it. Empirical coverage records how often realised demand falls inside a stated central interval, while interval width records the price of that coverage in demand units. These measures answer different questions and cannot be substituted for one another.

Retail uncertainty evaluation supports this multi-quantile view. The M5 uncertainty competition required nine quantiles to describe complete future-sales distributions, including central and tail outcomes. [S16] The cold-start harness need not copy those exact quantiles. It should evaluate the quantiles the planning service actually publishes, with tail definitions fixed before the replay rather than chosen after the errors are visible.

Coverage is reported by product family, channel, horizon and launch age. Aggregate coverage can conceal a service that is conservative for institutional orders and overconfident for size-level consumer demand. Width belongs beside every coverage

figure, and observed support belongs beside both. A narrow interval that misses often is not precise skill. A very wide interval that covers almost everything may still be unusable for a material commitment.

Calibration itself requires validation. Kuleshov and colleagues provide a regression-calibration procedure and explicitly qualify its guarantee by the availability of enough data. [S15] The calibration map is fitted on training or dedicated calibration launches, never on the final test outcomes. Where the launch corpus is small, the readout shows empirical coverage descriptively and avoids claiming that a calibrated population-level frequency has been established.

The audit also checks quantile coherence and output validity. Retail sales research on count data notes that point forecasts alone are insufficient for supply-chain and inventory decisions and that full discrete predictive distributions should be evaluated with proper scoring rules rather than relying only on continuous approximations or point metrics. [S1202] Quantiles must not cross, demand support must match the contract and any count transformation must be identical in replay and operation. Negative demand or impossible order quantities cannot be silently clipped after scoring. If a repair is required, it belongs in the candidate specification and is evaluated as part of that candidate. This keeps distribution quality attached to the forecast planners would actually receive.

12.7

Decision-weighted readouts preserve operational meaning

The same numerical error can have different consequences across launch conditions. An overforecast before a material commitment can create excess output, while an underforecast during a constrained replenishment interval can leave demand unobserved. Sales-only estimates can be negatively biased when stockouts censor the focal product, so availability must shape both eligibility and interpretation of forecast error. [S06] The readout therefore separates fully observable, partly censored and unobservable periods under rules fixed before comparison.

Results are cut by product family, channel, horizon and availability state. The purpose is not to multiply tables until a favourable cell appears. Each segment corresponds to a decision the service will actually support, and the complete set is reported together. A candidate that improves the style total while worsening the size-level commitment has not passed an undifferentiated release gate. A candidate that performs well only when stockout-affected cases are removed has evidence for a narrower deployment condition.

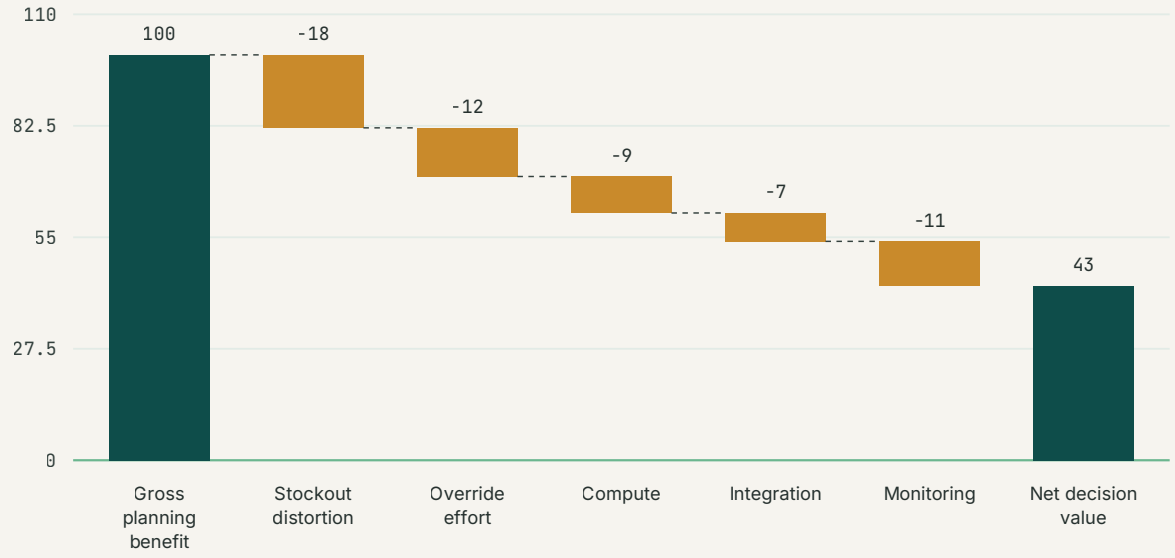
The planning replay then maps each frozen forecast distribution into the decision rule under test. That rule may select a first lot, reserve capacity or release a replenishment quantity. Demand quantiles remain separate from capacity, material, cash and commercial constraints. The protocol records the rule and its cost or loss weights before revealing actuals. It then compares the outcome of using the candidate with the outcome of using each baseline on the same information and constraints.

This step is necessary because a lower statistical loss does not guarantee a better system result. Google's engineering guidance explicitly rejects a change that improves log loss while degrading system performance. [S33] In an apparel launch, a small central-error gain can be irrelevant if it increases severe overproduction, misses a decision deadline or creates forecasts that planners must repeatedly override. The final readout asks whether the model improved the approved planning result after those consequences are included.

Exhibit 57, p. 195 illustrates that reconciliation with constructed decision-value points. The opening gross planning benefit is 100 points. Deductions are 18 for stockout distortion, 12 for forecast-override effort, 9 for compute, 7 for integration and 11 for monitoring, leaving 43 points because 100 minus 18 minus 12 minus 9 minus 7 minus 11 equals 43. These illustrative values show the required arithmetic only. They are not money, a return estimate or a benchmark for a real venture.

EXHIBIT 57 / ILLUSTRATIVE DECISION-VALUE POINTS

Model value should survive each operational deduction



ILLUSTRATIVE [S06] [S33] [S39] Illustrative constructed case, not observed value or a return forecast. Inputs are gross planning benefit 100, stockout distortion 18, override effort 12, compute 9, integration 7 and monitoring 11. Net equals 100 - 18 - 12 - 9 - 7 - 11 = 43. Actual inputs must be measured for the pilot and stated decision.

Offline decision replay has a limit. If a larger recommended lot would have changed availability, sell-through or the later training record, the observed counterfactual is not present in the file. The report must not pretend that the complete outcome of a different production history is known. It can score the fixed decision rule under documented assumptions, mark the feedback limitation and require a controlled pilot for claims that depend on changed operations.

METHOD

Freeze a decision scorecard beside the statistical scorecard. For each intended use, record the planning rule, eligible availability states, constraint set, asymmetric error consequences, operating effort and unresolved counterfactuals. Compare candidate and baselines on identical launch cases. Report every segment, not only the aggregate, and carry unmeasurable decision effects forward as limitations rather than assigning them a favourable value.

12.8

Release requires a baseline win, calibration and operating fitness

Release is a comparative decision, not a demonstration that a candidate can produce a plausible forecast. The comparison set includes the frozen analogue prior, eligible naive methods, an attribute baseline and a small combination baseline. M4 results caution against assuming that pure machine-learning methods automatically beat naive and combination benchmarks. [S10] A broad review of forecast combinations likewise finds that equally weighted averages can perform well against more elaborate weighting schemes. [S20] These rivals stay in the harness throughout the model ladder.

The first gate is evidence integrity. Every origin manifest, launch exclusion, as-of join and fold artefact must pass. The second is support: independent launch counts and effective cell sizes must meet the pre-registered release condition, or the result remains preliminary. The third is paired point performance, including MAE, eligible MASE, bias and the agreed baseline-relative measure. The fourth is distribution quality across the quantiles and intervals intended for planners.

The fifth gate is decision performance under the frozen planning rule, including availability-sensitive segments and operating consequences. The sixth is operating fitness: runtime, failure behaviour, data dependencies, monitoring effort, reproducibility and the tested baseline fallback. Technical-debt research warns that a small accuracy benefit purchased through a large increase in system complexity is rarely wise. [S39] A candidate therefore earns deployment as a complete service component, not as an isolated score.

Exhibit 58, p. 197 summarises why no single metric can carry that decision. In this illustrative GFT control coding, a value of 2 means the cut is mandatory for the release readout, 1 means it is a required diagnostic or sensitivity cut, and 0 would mean not applicable, although no zero is needed in this design. [S06] [S12] [S14] [S33] The matrix is a derived GFT control requirement, not an empirical score or a claim that one candidate has passed.

EXHIBIT 58 / GFT REQUIREMENT CODE: 2 MANDATORY, 1 DIAGNOSTIC

No single metric can carry the release decision

	Product	Channel	Horizon	Availability
MASE where eligible	2	2	2	1
MAE and signed bias	2	2	2	2
Quantile loss	2	2	2	1
Coverage and width	2	2	2	1
Planning outcome	2	2	2	2
Runtime and failure behaviour	1	1	2	1
Independent launch support	2	2	2	2

Cell scale: 1 to 2

DERIVED [S06] [S12] [S14] [S33] Derived release-readout matrix from availability, scaled-error, proper-scoring and system-value principles. Codes are control requirements, not measured performance. MASE appears only where its denominator is eligible. Every cell also carries independent launch support and, where relevant, interval width.

The release protocol names its owners before the final readout. The evidence owner certifies corpus and as-of controls. The modelling owner certifies reproducible fitting and scoring. The planning owner confirms that metrics and thresholds represent the decision. NIST guidance calls for human judgement in choosing trustworthiness metrics and their precise thresholds, which supports explicit approval rather than an automatic promotion rule. [S36] The release and reversal conditions are signed before the preferred candidate's result is seen.

Four outcomes should remain available. A candidate can graduate for a defined product, channel and horizon. It can remain a shadow challenger while more launches accumulate. It can be rejected because the evidence or operating burden is unfavourable. Or the whole comparison can be declared insufficient, leaving the approved prior or baseline in place. The last outcome is a successful control decision when support is weak, not a failed modelling exercise.

Re-entry follows the same discipline. A candidate that failed calibration, arrived too late or depended on unstable features does not return after a cosmetic repair. Its changed pipeline receives a new version and repeats the relevant launch replay on evidence permitted for that decision. Any final holdout consumed by an earlier release decision is no longer described as untouched. Fresh independent launches provide the cleanest confirmation.

The handover pack contains the corpus and origin manifests, fold artefacts, paired forecasts, metric eligibility, segment support, distribution diagnostics, planning outcomes, runtime evidence, approvals and fallback test. It tells the service owner exactly which claims were tested and which remain open. Later monitoring can then

compare live behaviour with the validated conditions rather than treating release as permanent proof. The backtest's product is not a winning algorithm. It is a bounded, reversible decision with an inspectable evidence trail.

A backtest earns trust by forgetting everything the forecaster could not have known.

CHAPTER SOURCE NOTES

R06 Cameroon Trade Hub. Publication date unrecorded; inherited source-text check.

R07 Ecofin Agency. Publication date unrecorded; inherited source-text check.

S13 scikit-learn. 2026-10-30.

S1201 arXiv; journal version in Machine Learning, Springer. 2019-05-28.

S02 arXiv. 2019-06-27.

S05 Springer Nature, International Journal of Data Science and Analytics. 2026-01-08.

S32 Springer Nature, Data Mining and Knowledge Discovery. 2008-07-30.

S37 Cell Press, Patterns; open full text from PubMed Central. 2023-08-04.

S12 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2006.

S14 Journal of the American Statistical Association, American Statistical Association. 2007.

S16 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2022.

S15 Proceedings of the 35th International Conference on Machine Learning, PMLR. 2018-07-03.

S1202 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2016.

S06 Lancaster University Management School, Munich Personal RePEc Archive. 2018-11-19.

S33 Google for Developers. 2025-08-25.

S39 Neural Information Processing Systems Foundation, NIPS 2015. 2015.

S10 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2018.

S20 arXiv; journal version in International Journal of Forecasting, Elsevier. 2022-09-23.

S36 National Institute of Standards and Technology, U.S. Department of Commerce. 2023.

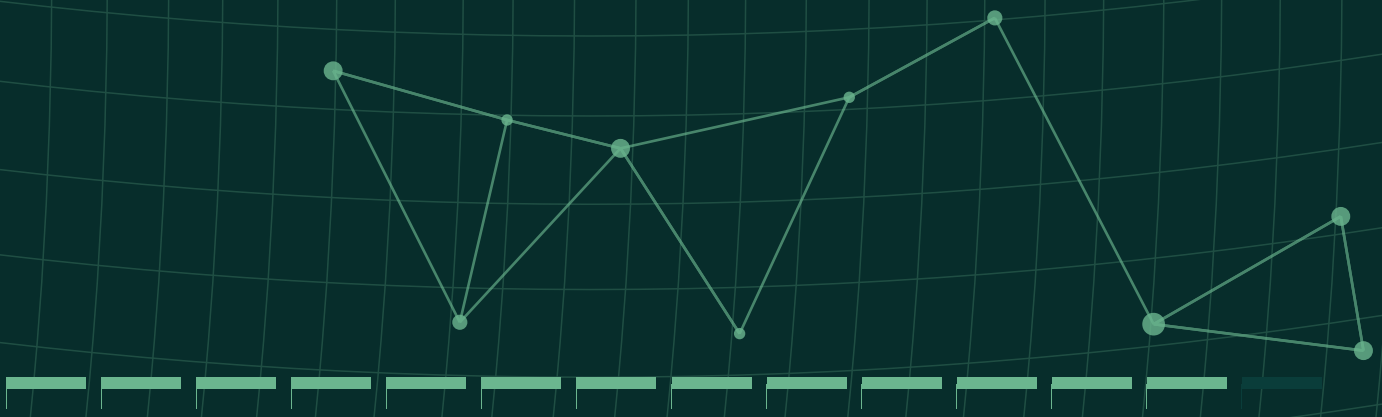
GREEN FUTURE TECH

LIMITED / HONG KONG SAR

13

Planners need a forecast service, not a model artefact

Forecasting creates value only when a named planning rhythm can understand, challenge and reverse the model's output.



CHAPTER 13
DECISION TEST

Forecasting creates value only when a named planning rhythm can understand, challenge and reverse the model's output.

A forecast service is the governed path from versioned evidence to a planning decision. A model artefact is only one component of that path. For a new apparel maker, the service must publish one authoritative forecast release, show its range and provenance, let planners test production scenarios, record every override and return to an approved baseline when evidence or operations deteriorate. The planning team, not the modelling team alone, owns the resulting commitment. This arrangement makes forecasting inspectable at the moment when money, materials and capacity are still reversible.

The distinction matters because an accurate file can still produce an unmanageable decision. Commercial staff may read a different extract from production staff. A stock update may arrive after the forecast was cut. A planner may change a number without retaining the reason. A candidate model may fail while no approved fallback is ready. These are service failures, and a better algorithm does not repair them. The operating design must therefore treat publication, challenge, approval, monitoring and reversal as parts of the forecasting method itself.

13.1

One governed forecast version reaches the planning table

The planning table needs a single release identity, not a claim that only one model was considered. Candidate, baseline and shadow forecasts can coexist in evaluation. The service then publishes the approved result for a stated decision cycle. That release binds the demand distribution to its training cut-off, feature snapshot, availability correction, calibration artefact, model configuration and forecast origin. Monitoring guidance treats data statistics, online performance, notification and roll-back as connected operational concerns, which is why the release must remain reproducible after publication. [S40]

A release register gives that identity an operational meaning. It records who approved publication, which planning decisions the release may inform, when it expires and which baseline remains available. Commercial and production users query the same release rather than exporting separate copies. Later actuals arrive as new evidence, never as silent edits to the forecast that was previously approved. If a correction is necessary, the service issues a new version with a reason and preserves the superseded one for scoring.

This discipline also separates authority from authorship. The modelling role can propose a candidate and supply its validation pack. The data role can certify that required inputs are complete and current. The planning owner can accept, challenge or reject the release for the decision at hand. None of these roles should be able to change the published forecast unilaterally. Human oversight is most useful when responsibility for metrics and intervention thresholds is explicit rather than distributed vaguely across a project team. [S36]

13.2 THE MANAGEMENT CADENCE CONSUMES THE RELEASE, NOT ANOTHER REPORT

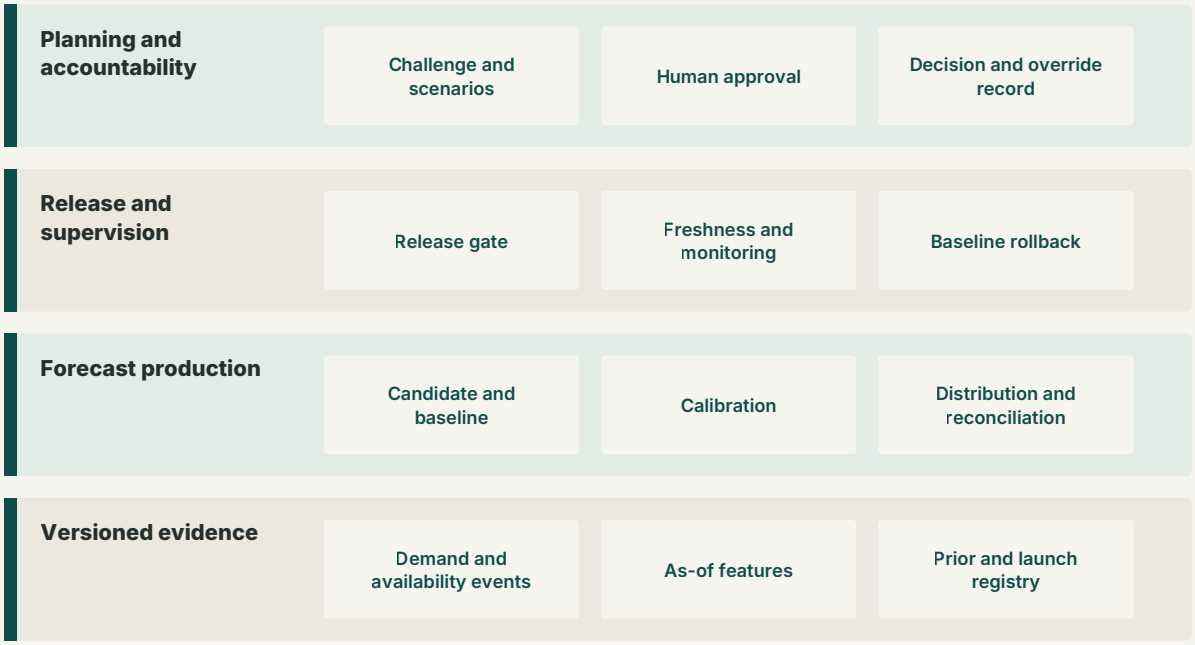
The forecast belongs inside the existing sales and operations rhythm. Before the meeting, the service freezes the information set and publishes any exceptions. During the meeting, planners review material changes, demand ranges and decisions that cross approved thresholds. Afterwards, the service retains the approved scenario, production decision and dissent. A weekly exception review can address urgent availability or order events without creating a second, contradictory forecast process.

A dashboard should consequently open with decisions requiring attention, not with a catalogue of model diagnostics. Statistical detail remains available for inspection, but it does not displace the planning question. The first screen identifies the product, channel and horizon, the current release, the decision deadline and any failed freshness checks. The reviewer can then move from the total to the contributing variants and from a changed forecast to the evidence that caused it.

Exhibit 59, p. 201 shows the required division of responsibility. Evidence, modelling and release controls remain separate layers so that a candidate cannot certify its own fitness. The planning layer receives an approved distribution and returns decisions, overrides and outcomes to the evidence layer. The architecture is a GFT operating proposal derived from the cited controls, not an implemented client system or a measured result.

EXHIBIT 59 / OPERATING RESPONSIBILITIES

A forecast service joins evidence, models and planning controls



DERIVED [S33] [S36] [S37] [S40] Derived GFT architecture from cited system-value, human-oversight, leakage and monitoring guidance. The layers define control responsibilities and feedback paths. They do not represent an implemented GFT client service or a measured outcome.

The feedback path is as important as the forward path. A forecast release that never receives the approved lot, actual availability, planner override and realised outcome cannot be evaluated as a planning instrument. It can be scored only as a numerical prediction. Closing that loop allows later review to distinguish model error, stale evidence, constrained availability and a deliberate human decision.

> **METHOD**
>
> Create a release manifest before creating a serving endpoint. Bind the forecast to its target definition, product and channel grain, horizon, forecast origin, data cut-off, availability rule, model and baseline versions, calibration window, approval owner and expiry. Store the manifest with the output. Reject publication if any required field is absent. Preserve superseded manifests so that a later reviewer can reconstruct exactly what planners knew and approved.

13.3

The planner sees range, assumptions and freshness

A useful forecast view answers three questions before it invites a decision. What does the model expect, how much variation does it admit, and is the underlying evidence current enough for this decision? A central estimate without the other two answers invites false confidence. Retail uncertainty evaluation uses multiple quantiles to describe future sales, including adverse and favourable tails, because different parts of the distribution support different planning concerns. [S16]

The planner should see the central forecast and selected tail quantiles in demand units, with the target event and horizon beside them. A prediction interval must be labelled as model-based uncertainty under stated assumptions. A price, promotion or supply scenario is a different object: it asks what follows when a named condition changes. Keeping those objects separate prevents a planner from treating a conditional management choice as a statistical probability.

Freshness is equally visible. The interface states when sales, stock, returns, product attributes and external features were last updated for the release. It also distinguishes an absent event from an absent record. Data leakage research shows that collection, sampling and preprocessing choices can create spurious relationships, so a clean-looking forecast cannot compensate for an information set whose timing or construction is unclear. [S37]

13.4 MISSING EVIDENCE PRODUCES A STATE, NOT SILENT REPAIR

Every required input needs a freshness rule and a failure state. A late sales feed may leave a previous release usable for a long-horizon assortment discussion but unfit for immediate replenishment. Missing availability may block interpretation of apparently weak sales. An outdated attribute taxonomy may make a new product invisible to an analogue model. The response should follow the decision risk, not a universal technical timeout.

The service therefore publishes a status with the forecast: current, usable with stated caution, or blocked. These are operational labels whose thresholds are approved during validation. They are not traffic-light decoration. A caution names the affected evidence and the decisions exposed to it. A block identifies the fallback and the role responsible for restoration. No interpolation occurs merely to keep the display complete.

Exhibit 60, p. 203 sets out the minimum reading contract. The rows are illustrative fields for a service design, not measurements from an existing deployment. Together they let a reviewer establish what is being forecast, which evidence shaped it, how uncertain it is and what would make it unsafe to use.

EXHIBIT 60 / REQUIRED PLANNER-FACING FIELDS

A readable release carries decision context beside the forecast

Field group	What the planner sees	Decision protected
Target	Event, grain, channel and horizon	Do not substitute one demand definition for another
Distribution	Central case, selected quantiles and units	Make upside and downside visible
Assumptions	Prior weight, scenarios and availability rule	Separate model conditions from management choices
Freshness	Input updates, forecast origin and expiry	Do not act on stale or incomplete evidence
Provenance	Release, model, calibration and feature versions	Reproduce and challenge the published output
Control	Owner, status, fallback and next review	Keep intervention and reversal actionable

ILLUSTRATIVE [S16] [S36] [S37] [S40] Illustrative GFT planner contract derived from cited quantile, oversight, leakage and monitoring principles. No implementation status or forecast performance is claimed. Thresholds and expiry rules must be approved for the actual decision and pilot evidence.

This reading contract changes the planner's question from whether a number looks plausible to whether the release is fit for a named commitment. That is the correct standard for a cold-start venture, where uncertainty and missing history are expected conditions rather than embarrassing exceptions.

13.5

Production scenarios translate quantiles without automating commitment

A predictive distribution describes demand. It does not specify how many garments to make. The planning service must translate selected quantiles into comparable production cases while keeping supply, capacity, cash and commercial constraints visible as separate inputs. A low case can expose idle-capacity and inventory risk. A

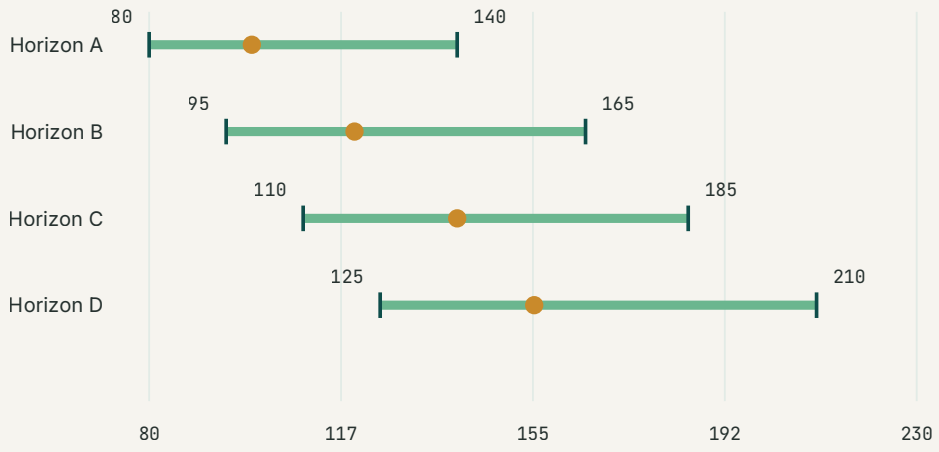
central case can support the working plan. A high case can test whether material, line time and channel inventory would allow the venture to respond. None of these cases is automatically the recommended lot.

The scenario sheet therefore carries two blocks that never merge. The demand block contains the released quantiles, their forecast origin and the availability assumptions. The commitment block contains minimum lots, changeover consequences, material availability, capacity reservations, delivery promises and any management risk allowance. The final quantity is a human decision made across both blocks. NIST's risk-management guidance specifically places human judgement around the choice of metrics and thresholds, which supports explicit approval rather than an unexamined conversion from quantile to production. [S36]

Exhibit 61, p. 204 provides a constructed illustration of the demand block only. Its values are not observations, recommended service levels or a Cameroon apparel forecast. The widening ranges demonstrate how a planner might receive central and tail cases across successive decision horizons. A real release would use calibrated output from the approved model and would show sample support beside any calibration claim.

EXHIBIT 61 / ILLUSTRATIVE GARMENTS PER WEEK

Production cases retain demand uncertainty across the planning window



ILLUSTRATIVE [S16] [S36] Illustrative constructed demand ranges, not observed sales or recommended production. Low, mid and high values are assumed planning quantiles: low [80,95,110,125], mid [100,120,140,155], high [140,165,185,210] for four successive horizons. They demonstrate the interface only. Capacity, supply and commercial constraints are deliberately excluded.

The exhibit should lead to a capacity conversation, not to automatic selection of the upper bound. If the high case cannot be served with existing materials and line time, the planner can define a conditional response: reserve an option, split the lot, delay a variant or accept unmet demand. Each response needs an owner and a reversal point. The demand forecast remains unchanged while management compares those choices.

This separation also makes post-launch learning possible. If actual demand falls near the high case but production follows the central case, the outcome does not prove that the forecast was wrong. If the forecast was high and the planner lowered it because a material constraint made additional output impossible, that constraint belongs in the decision record. Forecast skill and planning choice can then be reviewed without confounding one with the other.

> **GFT NOTE**
>
> GFT recommends a scenario sheet with two signatures. The forecast owner approves the demand distribution and its evidence status. The planning owner approves the production commitment and records the constraints applied. The service may reject an expired or blocked forecast, but it must not convert a quantile into a lot without that second approval. Keep the unmodified model release, the selected scenario and the final commitment as three linked artefacts.

13.6

Overrides need reasons and subsequent scoring

A planner override is new evidence entering the forecast through a human route. It may reflect a confirmed institutional order, a channel change, local knowledge or a concern that the analogue set no longer applies. It may also reflect optimism, anchoring or pressure to fill capacity. The service should neither prohibit overrides nor accept them invisibly. It should make their assumptions inspectable and their consequences measurable.

Structured elicitation provides a better starting point than an edited spreadsheet cell. The Delphi method uses anonymous responses, summarised group feedback and opportunities to revise, allowing expert judgement to be challenged without forcing immediate consensus. [S29] A planning override need not become a full Delphi exercise, but it can adopt the same discipline: state the reason independently, expose the evidence, allow revision and preserve dissent.

The override record contains the original release, adjusted horizon and product cells, magnitude, reason code, supporting evidence, author role and timestamp. A small reason taxonomy is sufficient: confirmed demand signal, offer change, availability correction, channel event, analogue concern, capacity constraint or other explained judgement. Capacity constraints should normally alter the commitment rather than the demand forecast. Keeping that distinction prevents a production limitation from being learned later as weaker customer demand.

13.7 OVERRIDE REVIEW COMPARES LIKE WITH LIKE

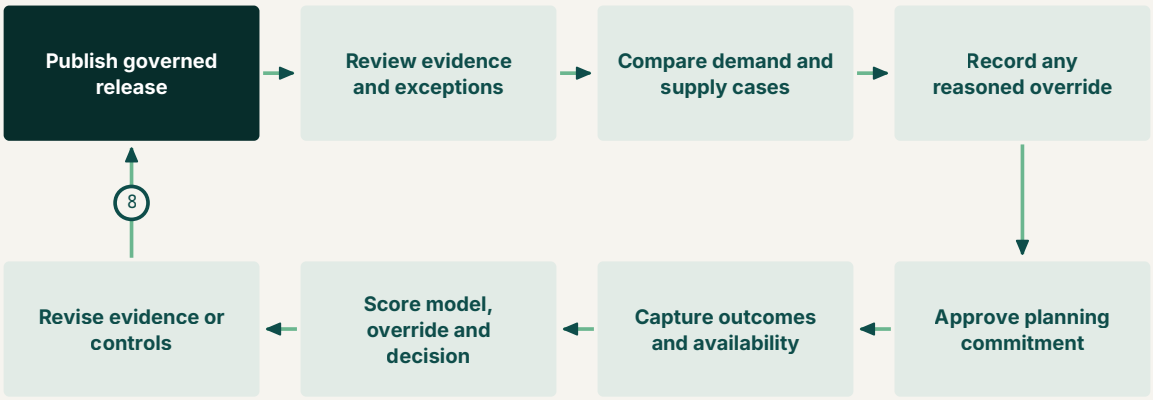
Every override remains paired with the unmodified model output and the eventual actual under the same target definition. The review reports both errors and bias, but it also checks whether the compared period was observable. A high override followed by a stockout cannot be declared correct merely because recorded sales rose to the available inventory. Sales-only demand estimates can be negatively biased when availability constrains the focal product, so stockout-contaminated outcomes need separate treatment. [S06]

Review should also distinguish individual cases from systematic behaviour. One successful override does not establish a general planning advantage, and one poor override does not prove that local knowledge is worthless. Reason codes and launch groups reveal whether certain kinds of intervention repeatedly help or harm. Changes to override rules follow that accumulated evidence rather than the most memorable meeting.

Exhibit 62, p. 206 places overrides within the complete planning loop. Challenge occurs before approval, while actuals and scoring occur after the decision. The model cannot erase its earlier output, and the planner cannot erase the override. Both remain available for fair comparison.

EXHIBIT 62 / FORECAST-TO-OUTCOME SEQUENCE

The planning loop closes with measured forecast use



8. Next governed cycle

DERIVED [S06] [S29] [S33] [S36] Derived GFT operating sequence from cited availability, structured-elicitation, system-value and oversight principles. This is a proposed control design, not evidence of an implemented monthly process or its results.

The loop also creates a useful boundary between learning and governance. Analysts can investigate whether overrides contain repeatable signal. Planners can change the reason taxonomy or approval threshold. Model owners can propose a new feature. None should rewrite history to make the previous cycle look cleaner. The preserved sequence is the evidence from which better rules can be built.

13.8

Monitoring watches inputs, errors, calibration and availability

Monitoring must explain whether the service remains fit to publish, not merely whether a server is responding. Four views are required. The input view checks completeness, freshness, valid identifiers and changes in product or channel populations. The output view checks forecast levels, dispersion and quantile consistency. The per-

formance view compares realised demand with the release using the agreed point and distribution measures. The use view tracks overrides, planner adoption, decision outcomes and fallback events. Each view can fail while the others appear normal.

Concept drift concerns a change in the relationship between inputs and the target over time. That is a deeper problem than a movement in average sales. A model may continue receiving valid fields while the meaning of those fields changes because the assortment, channel or buying process has moved away from the training population. Drift-adaptation research supplies a taxonomy for detecting and responding to such change, but it does not supply an apparel-specific alarm threshold. [S38]

The monitoring design therefore begins with expected behaviour recorded at release. Input summaries cover attributes, price, availability, channel mix and valid history length. Output summaries cover the central forecast, interval width and the share of forecasts using priors or fallbacks. Performance summaries retain errors by product family, channel, horizon and availability condition. These segments show where transfer is weakening instead of hiding the change inside a total.

13.9 ALERTS NEED PRESCRIBED RESPONSES AND HUMAN REVIEW

An alert is useful only when it names the affected releases, the decision exposure and the next action. A stale feed may block publication. A shift in product attributes may require renewed analogue review. Deteriorating interval coverage may lead to wider uncertainty or reversion to a calibrated baseline. Repeated overrides for one reason may indicate a missing feature or a broken planning assumption. None of these signals should trigger silent retraining simply because automation is available.

Operational guidance connects data summaries and realised model performance with notification and rollback when values deviate from expectations. [S40] The service should implement that chain explicitly: detect, classify, notify, contain, investigate and requalify. Containment can mean retaining the release with a caution, restricting its permitted decisions or returning to the baseline. The choice depends on evidence and decision risk, and the responsible role is named in advance.

Availability remains a monitoring field in its own right. A period with many stockouts can make forecast error look artificially small if both the forecast and recorded sales are constrained. It can also make an underforecast appear acceptable because shelves emptied. The performance readout therefore marks censored periods and retains them in a separate sensitivity view rather than allowing them to certify model skill. [S06]

Thresholds need the same evidentiary discipline as forecasts. They should come from launch replay, pilot operations and the cost of a missed or false intervention. A threshold copied from another retailer may be unusable for a small launch corpus. Where support is thin, the service can require review without claiming that a statistically precise drift boundary has been estimated. The honest state is enhanced caution with an explicit evidence-collection plan.

Monitoring also includes the absence of decisions. If planners repeatedly ignore a release, export it into another workbook or use only its central estimate while discarding the range, the service is not performing its intended role. Adoption and compre-

hension are operational signals. They may point to poor forecast skill, unclear presentation, an inconvenient cadence or lack of trust. The response should follow investigation rather than assuming that more model complexity will restore use.

13.10

Rollback to a baseline is a normal control

A forecasting service needs a usable answer when its preferred candidate cannot be trusted. That answer may be the frozen analogue prior before launch, a naive or attribute baseline after observations arrive, or a previous approved release whose evidence conditions remain valid. The fallback is selected during release design, not improvised after an incident. It runs in the same evaluation and uses the same demand contract so that reversion does not introduce a new target or grain.

Rollback conditions cover evidence, model and decision health. Failed identity checks, unavailable critical features, expired calibration, invalid quantiles, excessive latency and breached performance expectations can all make a candidate ineligible. The service records which condition fired, which baseline took over and which planning decisions were affected. Operational monitoring guidance explicitly includes rollback among the responses to unexpected data or model performance. [S40]

A fallback must be tested while conditions are normal. The team rehearses the technical switch, checks that planners can recognise the active baseline and confirms that downstream scenario sheets still reconcile. It also verifies that the baseline's own assumptions remain valid. Returning to an obsolete prior is not a safe outcome merely because the prior is simple. The fallback register therefore carries freshness, eligibility and expiry rules as well.

13.11 RE-ENTRY REQUIRES NEW EVIDENCE, NOT A CLEARED ALARM

A candidate does not return automatically when a technical fault disappears. Re-entry requires the cause to be classified, the correction to be validated and the release gate to be repeated on the appropriate launches. If drift caused the incident, the review asks whether the reference population, features or target definition changed. If calibration failed, it tests the replacement calibration on evidence that was not used to fit it. If the operating burden caused missed deadlines, it reassesses whether the candidate belongs in production at all.

This discipline limits technical debt. Research on machine-learning systems warns that a small accuracy benefit can be unwise when it requires a large increase in system complexity. [S39] A fallback that depends on another fragile chain of corrections may preserve mathematical sophistication while weakening operational control. The preferred baseline is the simplest approved forecast that supports the decision reliably under the current evidence.

Rollback is consequently not an admission that the forecasting programme has failed. It is evidence that the programme can contain failure. A service that never needs reversion may be operating in stable conditions, or it may be unable to recognise deterioration. Recording and reviewing fallback events makes that distinction testable and gives planners a reason to trust the release mechanism rather than any individual model.

13.12

Value is measured in decisions rather than model loss alone

The value review begins after the forecast has influenced a real or faithfully replayed decision. It asks whether the release improved the planning outcome after accounting for availability, overrides, operating effort and the decisions that remained feasible. A lower forecasting loss is useful evidence, but it is not the final test. Practical machine-learning guidance rejects a change that improves log loss while degrading system performance, and the same principle applies to a production plan. [S33]

A balanced scorecard keeps statistical skill, evidence quality, planner use and operating health beside one another. MASE can support comparison across heterogeneous series when its naive denominator is available and nonzero, but a cold-start service must also retain unscaled error and bias where that denominator does not exist. [S12] Distribution review considers calibration and sharpness rather than treating any probabilistic output as automatically reliable. Proper scoring rules provide a principled basis for comparing complete predictive distributions. [S14]

Exhibit 63, p. 210 defines the minimum review set. It deliberately supplies no universal target values. Thresholds must come from the venture's decision costs, pilot evidence and operating capacity. A numerical target imported from another context would conceal precisely the cold-start uncertainty that the service is meant to manage.

EXHIBIT 63 / ILLUSTRATIVE MEASUREMENT FAMILIES

The service scorecard balances skill, use and operating health

MAE, eligible MASE

Point skill

Include denominator support and launch counts

Signed bias

Directional error

Segment by horizon and channel; mark stockout-affected outcomes

Score and coverage

Distribution quality

Retain interval width and assess the release planners received

Freshness, fallback

Release health

Link every alert and fallback cause to a decision exposure

Adoption, overrides

Human use

Retain reason codes and compare model and human against the same actuals

Planning outcome

Decision result

Record approval, constraints and feasible counterfactuals

ILLUSTRATIVE [S06] [S12] [S14] [S33] [S40] Illustrative GFT scorecard with measure names rather than claimed values. No accuracy, adoption, freshness or return target is asserted. Each measure requires a local formula, owner, segmentation rule, review cadence and action threshold before use.

Decision outcomes need precise definitions. For a first lot, the review may record unsold units, unmet observed orders, emergency replenishment and the production options that were genuinely available. For a later cycle, it may examine whether a release changed allocation or avoided an unnecessary commitment. These outcomes are not automatically caused by the forecast. Price, promotion, capacity and availability may have changed as well. The review should state the attribution limit and, where feasible, compare phased or otherwise comparable decisions rather than claiming a universal return.

WORKED EXAMPLE: LOWER AVERAGE ERROR CAN STILL PRODUCE A WORSE PLAN

Assume an illustrative four-period planning cell with actual demand of 80, 100, 120 and 100 units. A baseline forecasts 100 units in every period. Its absolute errors are 20, 0, 20 and 0, giving MAE of 10 units and zero mean signed bias. A candidate forecasts 90, 95, 105 and 95. Its absolute errors are 10, 5, 15 and 5, giving MAE of 8.75 units. Now assign an illustrative planning loss of 1 point per overforecast unit and 4 points per underforecast unit. The baseline loses 20 plus 80, or 100 points. The candidate loses 10 plus 20 plus 60 plus 20, or 110 points. The candidate has lower MAE but higher planning loss. Every value and cost weight is constructed to demonstrate the review method, not observed apparel performance or a recommended cost ratio.

The example shows why bias direction and decision asymmetry belong beside aggregate accuracy. It does not prove that underforecasting is always more costly, nor that the baseline should win after one cell. A genuine review would examine enough independent launches, check uncertainty and availability, and include the service's operating burden. The purpose is to prevent a favourable statistical summary from ending the discussion before its planning consequence is known.

Value review also covers negative value. An unused forecast consumes integration and review effort. A forecast that encourages excessive overriding may add noise. A complex candidate that misses the decision clock can be worse than a timely baseline. Retiring a model, feature or report is a valid outcome when its measured contribution does not justify the work it creates.

13.13

Build-against-buy review protects data ownership and exit

The build-against-buy decision concerns components, not sovereignty over the whole forecasting problem. Commodity storage, orchestration or managed inference may be purchased when they reduce operating work. The demand definitions, event history, availability record, release register and evaluation evidence must remain portable and inspectable. A vendor can supply a model endpoint without owning the meaning of the venture's demand or the authority to approve its production plan.

Vendor assessment starts with the exact service boundary. The register lists fields sent, transformations applied, hosting and retention arrangements, model and configuration identifiers, monitoring access and the outputs returned. It also records licence and support terms, revision policy and the ability to export features, forecasts and logs. Legal and contractual terms require the appropriate professional review; the technical register makes that review concrete rather than implying legal clearance.

Reproducibility is the decisive exit test. The team should be able to take a historical release manifest and its exported evidence, run the approved baseline elsewhere and reproduce a usable planning package. It should also test what happens when the external service is unavailable, a model version changes or a required export is incomplete. Monitoring and rollback controls are meaningful only if the organisation can inspect their signals and act on them through an alternative route. [S40]

13.14 TOTAL OPERATING BURDEN BELONGS IN THE PROCUREMENT COMPARISON

A low headline inference price can coexist with expensive integration, correction and supervision. The comparison therefore includes data preparation, availability capture, calibration, launch replay, planner training, incident handling, version review and exit work. It also includes the opportunity cost of maintaining features that do not improve decisions. Technical-debt research supports rejecting a tiny accuracy benefit when its system-complexity cost is disproportionate. [S39]

The review is repeated when the service or the venture changes. A new channel may alter the data contract. A vendor revision may change output under the same product name. A growing launch corpus may make an owned global model practical, or it may confirm that a simpler purchased component remains sufficient. Neither building nor buying is a permanent virtue. The approved configuration is the one that preserves evidence ownership, planning value and a tested exit under current conditions.

GFT's practical default is consequently proportionate: buy commodity capability where it is demonstrably useful, keep the demand record and release controls under the venture's authority, and require every candidate to leave a reproducible trail. The service is complete when planners can understand a release, challenge it, act within their constraints and continue safely without the preferred model. Anything less is a model artefact with a planning interface attached.

A forecast earns its place through governed decisions, not through the sophistication of the artefact that produced it.

CHAPTER SOURCE NOTES

S40 Google Cloud Documentation. 2024-08-28.

S36 National Institute of Standards and Technology, U.S. Department of Commerce. 2023.

S33 Google for Developers. 2025-08-25.

S37 Cell Press, Patterns; open full text from PubMed Central. 2023-08-04.

S16 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2022.

S29 RAND Corporation. 2023-12-29.

S06 Lancaster University Management School, Munich Personal RePEc Archive. 2018-11-19.

S38 ACM Computing Surveys; open record from Bournemouth University. 2014-04-30.

S39 Neural Information Processing Systems Foundation, NIPS 2015. 2015.

S12 International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2006.

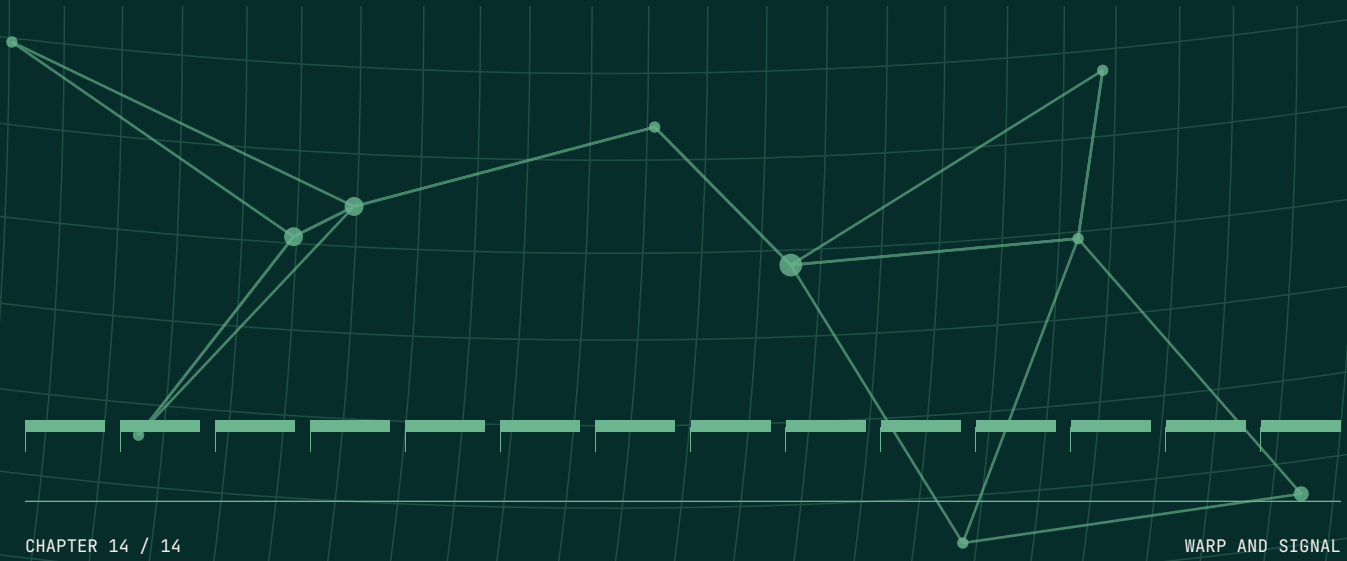
S14 Journal of the American Statistical Association, American Statistical Association. 2007.

14

PART IV: PROVE VALUE AND GRADUATE WITH EVIDENCE / RECOMMENDATIONS AND ROADMAP

The eighteen-month roadmap replaces priors with owned evidence

A new apparel maker should buy evidence in stages and let each forecasting capability earn its place through measured launch performance.



CHAPTER 14
DECISION TEST

A new apparel maker should buy evidence in stages and let each forecasting capability earn its place through measured launch performance.

A new apparel maker should use the illustrative eighteen-month roadmap to acquire evidence in stages, with each capability earning authority at a documented gate. A new apparel maker begins with borrowed evidence, but it should not remain dependent on that evidence after customers have supplied usable observations. Each phase therefore has a defined artefact, a decision gate and a reversal path. Spending moves from contracts and measurement to pilots and controlled model comparison. The venture can stop at any gate without losing the value of the work already completed.

Graduation does not abolish priors. Analogues, market context and expert judgement remain necessary for a new style, channel or market without target history. Owned launch evidence increasingly tests and revises their assumptions. A prior also remains the fallback when new data are sparse, stock-censored or unlike the planned decision. The roadmap is complete when the organisation can learn and reverse safely, not when it has installed the most elaborate forecaster.

14.1

Recommendation one funds the evidence plan before the model platform

The first expenditure should produce a forecast contract, an analogue register, a calibrated choice instrument, a pilot charter and a clean demand record. These artefacts determine which model questions are worth asking. They also remain useful if the venture later changes software supplier, model architecture or production plan. A platform selected before those artefacts exist must guess at the target event, product grain, forecast horizon and availability rule. That guess becomes an expensive constraint on every later decision.

The forecast contract connects prediction to a commitment. It names the decision owner, target event, product and channel grain, horizon, information cut-off, availability treatment, evaluation measure and escalation rule. The contract prevents a convenient sales extract from silently becoming the target merely because a modelling tool accepts it.

The evidence plan then separates three kinds of pre-launch work. Analogue research establishes an outside view from comparable products and launch conditions. Primary choice research estimates preferences for attributes that the venture can actually deliver, with an explicit calibration step between stated choices and real purchases. Research on preference learning supports that calibration problem rather than a direct conversion from survey enthusiasm to sales. [S1402] Pilot design specifies how real offers, availability and purchases will test the resulting range. Controlled-experiment guidance supports deliberate assignment and cautious causal interpretation where the operating design permits it. [S32]

Clean capture belongs in the same first work package. Every transaction needs product, channel, price and time context. Availability intervals and stockout timestamps must sit beside sales, because sales-only demand estimates can be negatively biased when the focal product is unavailable. [S06] As-of timestamps preserve what the team knew at each decision point. Leakage research shows why col-

lection, sampling and preprocessing must all be audited, not just the final train and test split. [S37] These fields are cheaper to define before the pilot than to reconstruct after its results have influenced a production commitment.

This order also limits technical debt. Machine-learning systems research warns against accepting a small accuracy benefit at the cost of a large increase in system complexity. [S39] For a small industrial team, the practical response is a complexity budget. It covers data preparation, model operation, monitoring, specialist effort, vendor dependence and recovery. A candidate that consumes most of that budget before the demand contract is stable has displaced the learning programme rather than supported it.

> **GFT NOTE**
>
> Fund the first release of the evidence register before funding a model platform. GFT would require these inspectable artefacts at that release: the signed forecast contract, the analogue and assumption register, the calibrated research protocol, the reversible pilot charter and the demand-event dictionary with availability and as-of fields. Procurement can then test commodity tools against known requirements. The venture owns the evidence and decision rules even when it buys computation or software.

The funding rule is therefore conditional rather than anti-technology. Buy a commodity component when it performs a defined task, exports its data and can be replaced without losing the launch record. Defer infrastructure whose main justification is future scale that the evidence has not yet supported. Each release should answer a named uncertainty and leave a reusable record. That discipline makes an early stop a valid programme outcome rather than a failed technology project.

14.2

Phase 0 defines the contract, analogues and pre-launch prior

The illustrative Phase 0 ends with an inspectable predictive range and a pilot that can change it. It does not end with a single promised sales number. The team assembles the demand dictionary, analogue scorecard, external covariate registry, choice-research results and structured expert judgements into one prior register. Every component carries provenance, observation date, intended use and known limitations. Dependencies are recorded so that an expert who has already seen the survey is not counted as an independent confirmation of it.

The analogue scorecard makes comparison auditable. Attribute-based fashion research uses clothing, footwear and merchandising factors to infer demand for new items from historical products, which requires both a relevant reference catalogue and consistent descriptors. [S02] Category, fabric, silhouette, colour, price band, occasion, channel and launch timing therefore become explicit fields. Failed, withdrawn and slow-selling analogues remain in the reference class where access rights permit. A pool composed only of memorable winners would encode survival rather than launch uncertainty.

Partial pooling offers a disciplined way to retain shared and local evidence. Hierarchical Bayesian modelling allows information to be shared across heterogeneous sites while preserving local parameters, although the cited empirical application is material demand rather than apparel. [S28] In this roadmap, the method is a candidate for organising analogue variation, not a source of automatic credibility. The prior register stores the whole launch distribution, the reasons for inclusion or exclusion and sensitivity to alternative pooling choices. A narrow range produced mainly by an assumption must be identified as such.

Choice research adds a different evidence form. The opened preference-learning manuscript distinguishes inexpensive stated preferences from actual purchase evidence and discusses choice-based conjoint analysis for attribute trade-offs. [S1402] GFT would vary only deliverable attributes and record the proposed calibration to purchases. The pilot must test that translation at real prices and availability. Attribute utility alone does not establish market share.

The pre-launch gate has six required outputs. The demand owner approves the target dictionary. The commercial owner approves the analogue and choice-evidence boundaries. The planning owner records which production or assortment commitment the range will inform. The data owner demonstrates the event schema and recovery procedure. The pilot owner freezes assignment, availability, stop and adaptation rules. The sponsor receives the prior range, its sensitivity and the conditions under which the pilot will stop, continue or expand.

Reversibility is part of the gate, not an appendix. The pilot plan must preserve the option to return to the prior, widen it or postpone a commitment when capture fails. A forecast version records its information cut-off and the weight assigned to borrowed evidence. Later revisions never overwrite that version. The illustrative Phase 0 is complete only when another qualified reviewer can reconstruct what was believed, why it was believed and which observations would change the plan.

14.3

Phase 1 uses the first twelve weeks to validate evidence and capture

The illustrative first twelve-week pilot is a measurement test before it is a model test. Its purpose is to expose defects in the offer, target and record while production exposure remains limited. The charter states which assortment, price, channel or scale decision the readout will inform. It also fixes the launch cells, release timing, availability checks, returns treatment and stop rules before results arrive. A favourable early sales total cannot rewrite those rules merely because the venture wants to expand.

Offer comparison should preserve the conditions needed for interpretation. Where random assignment is operationally feasible, comparable customers, locations or time cells receive defined offers so that differences can support a causal reading. Controlled experiments provide the strongest design for establishing such relationships in the settings covered by the cited guide, but an apparel pilot still needs its own

feasibility and contamination checks. [S32] Where assignment is not feasible, the readout describes the selection mechanism and avoids causal language. A distributor choosing its own order timing is not equivalent to a randomised offer cell.

Availability receives the same design attention as the offer. Stock counts, listing uptime, stockout timestamps, substitutes and returns determine whether low sales mean weak demand or constrained observation. The stockout-annotation literature makes the general problem explicit: observed sales can systematically under-represent latent demand when a product is unavailable, although its fresh-retail evidence is not direct apparel proof. [S07] The pilot therefore retains both the raw sales record and the availability version used in every analysis. A corrected target, if one is justified, is a new derived field rather than an overwrite.

The first comparison is between the calibrated prior and observed purchase evidence under the actual pilot conditions. The team records central error, dispersion, channel differences and any systematic shortage or surplus. It also records where the forecast cannot be interpreted because stock or listing data failed. More observations do not automatically justify more confidence. A revision may widen the range when prices change mid-pilot, channels diverge or an institutional order dominates a consumer series.

> **WORKED EXAMPLE: COMPARABLE EXPOSURE EARNS INFLUENCE**
>
> An illustrative pilot records 180 fulfilled units over six weeks, with no stock available throughout one week. The calendar-week rate is 180 / 6 = 30 units. The available-week rate is 180 / 5 = 36 units. Neither calculation recovers demand during the unavailable week. Assume comparable offers and availability throughout the other weeks, solely to demonstrate the following update.
>
> For this illustrative working rule, give an assumed prior rate of 30 units per available week the weight of two comparable weeks. Its pseudo-count is 30 × 2 = 60 units. Combine that pseudo-count with the 180 observed units over five available weeks: (60 + 180) / (2 + 5) = 34.29 units per available week, rounded. Borrowed influence is 2 / 7 = 28.57 percent and owned influence is 5 / 7 = 71.43 percent. This is a deliberately chosen weighted-rate rule, not a fitted posterior or an uncertainty interval. Its assumed comparability and prior strength require separate testing; the unavailable week contributes neither an observed zero nor an imputed sale.

The data gate runs beside the commercial readout. Validation checks event sequences, product identifiers, paid prices, return links and late-arriving corrections. As-of audits confirm that each replay uses only values available at its forecast origin. Leakage can enter through collection, sampling or preprocessing, so scaling, imputation and feature selection belong inside the training portion of every later backtest. [S37] No complex model graduates while the target definition or these records fail their completeness checks.

The illustrative Phase 1 has three legitimate outcomes. **Pass** means the evidence record and pilot design support controlled comparison. **Extend** means the pilot remains usable but needs more observations or a longer returns window. **Redesign**

means the target, assignment or availability record cannot support the intended inference. Early sales alone cannot decide among these outcomes. The gate protects the venture from scaling a misleading measurement process.

14.4

Phase 2 uses months four to eight for controlled model comparison

The illustrative months four to eight window introduces competition among models only after the evidence gate clears. Every candidate receives the same frozen launch corpus, forecast origins, target definitions and decision segments. Whole launches are held out when the claim concerns unseen products. Related size and colour launches remain grouped where they share a campaign, stock pool or order, preventing many dependent rows from masquerading as independent proof. Sparse cells stay visible and are labelled preliminary.

The comparison begins with eligible baselines. Attribute regression tests whether product descriptors explain demand across historical items. Intermittent-demand methods become candidates where zero-heavy series reflect genuine sporadic demand rather than missing listings or unavailable stock; the defining research describes demand that occurs sporadically with some periods showing no demand. [S18] Naive, smoothing and equally weighted combination forecasts remain in the harness where their required histories exist. A new method must beat the appropriate simple rival, not merely another elaborate candidate.

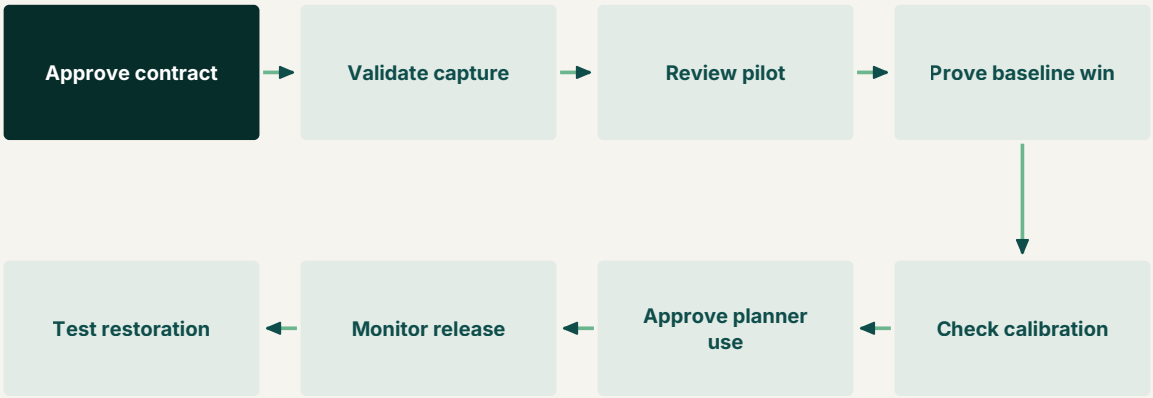
Cross-series models enter as a second family. A global approach fits one function across many series, allowing statistical evidence to be shared rather than fitting every product independently. [S22] DeepAR supplies one probabilistic candidate trained across related series. [S21] Partial pooling supplies a different route, balancing local and shared parameters with prior-sensitivity and posterior-predictive checks. [S28] These approaches borrow evidence in different ways. The evaluation should show which borrowed population helps, which local differences remain and whether performance survives a held-out launch.

Pretrained challengers enter through the same gate. Chronos, for example, tokenises scaled series values and trains language-model architectures for probabilistic forecasting. [S25] That capability justifies a controlled trial, not an exemption from target-domain validation. The candidate uses the same availability-corrected target, launch split, horizons and planning measures as every other model. Any calibration, adaptation or context layer is fitted only on information available inside the relevant training fold.

Exhibit 64, p. 219 makes the release sequence visible. These are proposed GFT requirements informed by the cited methods. They confer no authority through an invented readiness score. A failed requirement holds the last approved state while its owner repairs the evidence.

EXHIBIT 64 / ILLUSTRATIVE RELEASE SEQUENCE

Every release requires evidence and a reversal path



ILLUSTRATIVE [S32] [S33] [S40] Illustrative GFT process design, informed by controlled experiments, decision value and monitoring guidance. Nodes are requirements, not measured readiness scores. Failure at any gate holds the current approved prior or baseline; the recorded issue must be repaired before re-entry.

The gates cannot be averaged into a passing result. A strong backtest cannot compensate for an invalid contract. A calibrated model cannot compensate for missing rollback. Each owner supplies evidence and a reversal action. The demand owner can stop target misuse, the data owner can halt a contaminated release, the model owner can withdraw a failed candidate and the planning owner can reject an unusable output. Monitoring and rollback retain authority after release rather than serving as a closing ceremony.

Release requires planning-relevant improvement. Engineering guidance gives a direct warning: a change that improves a statistical loss while degrading system performance should not be accepted as progress. [S33] The model review therefore compares forecast error, bias, distribution quality, runtime and decision outcome by product, channel and availability condition. It also tests whether the approved baseline can be restored without a new implementation project.

Insufficient evidence produces a hold decision. If the frozen corpus lacks enough independent launches for a defensible replay, the venture remains with the approved prior or baseline and invests in further evidence. It does not force an illustrative Phase 2 release to keep the roadmap on schedule. The timeline is a planning allowance; the evidence gate is the actual dependency.

14.5

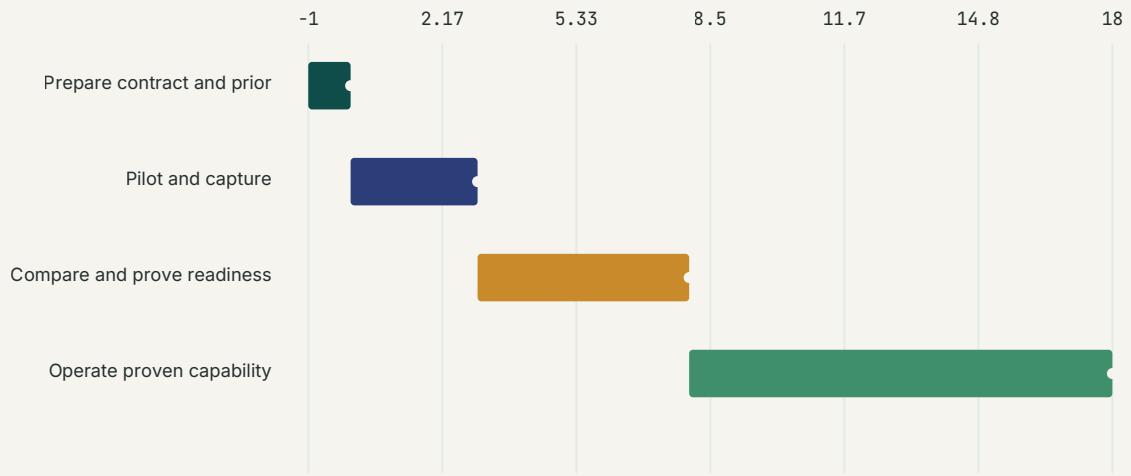
Phase 3 uses months nine to eighteen to scale only proven capability

The illustrative months nine to eighteen window turns an accepted candidate into an operating capability. Scale here means reliable use in defined planning decisions, not a universal roll-out across every product and channel. A model that won on consumer sell-through does not automatically graduate to institutional orders. A calibrated style total does not establish calibrated size-level demand. Each extension receives its own evidence pack and may remain outside the release boundary.

Exhibit 65, p. 220 uses pilot launch as the common origin. The illustrative preparation allowance precedes that origin. Twelve weeks are approximated as three months for drawing; the subsequent phases cover months four to eight and nine to eighteen after launch. Readiness testing belongs inside controlled comparison. Gate failures can extend any phase without creating permission to release.

EXHIBIT 65 / ILLUSTRATIVE MONTHS RELATIVE TO PILOT LAUNCH

Preparation precedes the eighteen-month learning window



ILLUSTRATIVE Illustrative planning allowances, not observed durations or GFT commitments. Pilot launch is month 0. Preparation occupies -1 to 0; the first twelve weeks are approximated by months 0 to 3. Months four to eight occupy 3 to 8; months nine to eighteen occupy 8 to 18. Release readiness is part of controlled comparison. Dates never override failed gates; this is not a resourced critical path.

Later operation adds structure only where the decision needs it. Hierarchical reconciliation becomes relevant when independent size, style, channel and regional forecasts contradict one another. MinT reconciliation seeks coherent forecasts across a collection of series under an unbiasedness assumption, so its use requires the assumption and resulting coherence to be inspected rather than treated as a formatting step. [S1405] The hierarchy follows the planning structure. It does not conceal weak variant evidence inside an apparently sound total.

Drift response becomes a standing control. Concept-drift research concerns changes in the relationship between inputs and targets and reviews adaptation and evaluation strategies. [S38] For this service, drift includes an assortment shift, a new channel mix, a changed price policy and a break in availability capture. Monitoring therefore

watches input meaning and population as well as forecast errors. A response can be renewed evaluation, a wider range, selective retraining or reversion to the approved baseline.

Retraining is selective because a new data cut is not automatically a better training set. The release package identifies the training cut-off, feature snapshot, model version, calibration window and intended decision horizon. Operational guidance calls for summary statistics, realised performance monitoring, notifications and rollback when values deviate from expectations. [S40] The venture tests the replacement on the same launch-replay logic before release. If the new candidate fails, the prior approved version or baseline remains available.

Retirement is equally important. Weak models, unused features and duplicated reports leave the service when their evidence no longer justifies their burden. The register records why each component was retained or removed and whether any downstream decision still depends on it. This keeps the final system proportionate to the planning problem. The illustrative Phase 3 succeeds when the organisation can reproduce, challenge and reverse every retained release without relying on an undocumented individual or supplier console.

14.6

GFT would measure skill, calibration, contamination, adoption and planning outcome

The final scorecard answers distinct questions. Is the forecast accurate enough for the decision? Are its probabilities honest? How much of the learning record is contaminated by unavailable stock or broken capture? Do planners understand and use the release? Did the resulting decision improve under a comparison that the evidence can support? A single accuracy number cannot carry these answers.

Skill reporting starts with the forecast contract and the available denominator. MASE scales absolute error by a naive benchmark and supports comparison across heterogeneous series. [S1404] At a cold start, however, that benchmark may be absent or zero. The report then marks MASE undefined rather than coercing it into a number. MAE, signed bias and an aggregate comparison with the frozen analogue baseline remain available where their definitions fit the target. Eligible-launch counts and denominator diagnostics appear beside every result.

Distribution reporting sits beside point error. Proper scoring rules encourage honest probabilistic assessments and provide a principled basis for comparing forecast distributions. [S1403] The release review retains quantile loss, empirical coverage and interval width by product group, channel and horizon. Coverage without width can reward an unhelpfully broad range. Width without coverage can reward confident error. Sample support stays visible because a small group of launches cannot sustain the same conclusion as a broad, independent evaluation corpus.

Contamination is a measure of evidence quality, not a nuisance variable to remove from the dashboard. The stockout-affected share divides unavailable eligible product-channel hours by observed eligible hours; unknown availability is reported separately. Availability matters because sales-only estimates can be negatively

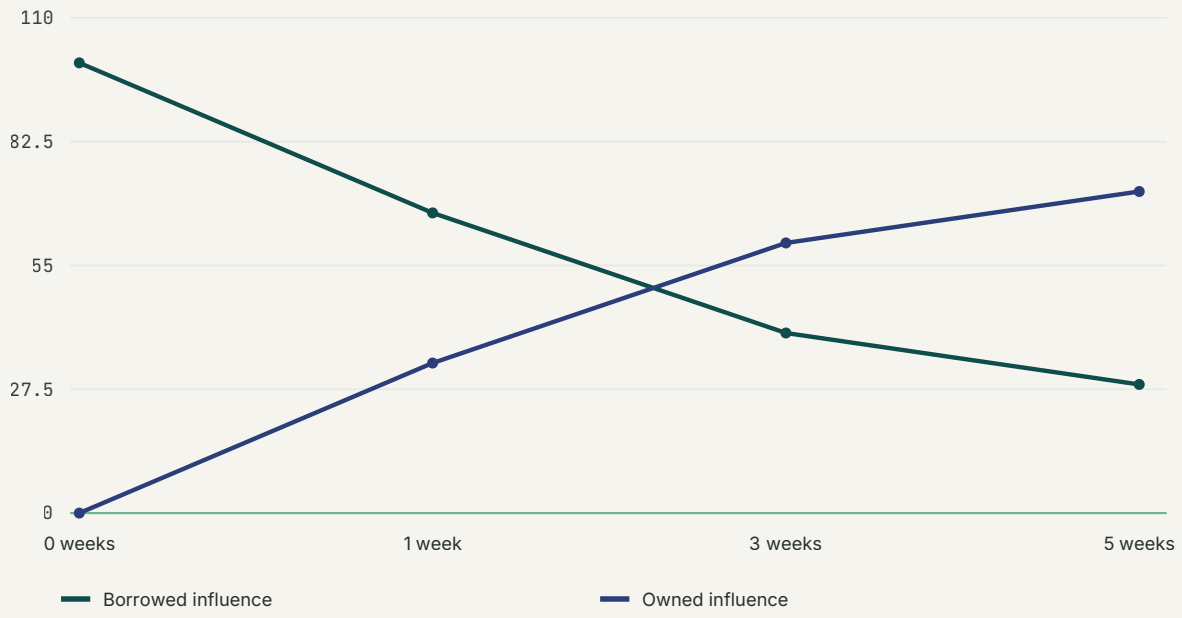
biased under stockouts. [S06] The scorecard also tracks missing product attributes, late events, unknown listing states and corrections whose demand basis is unresolved. A model should not receive a clean-looking accuracy score because difficult intervals disappeared without a frozen exclusion rule.

Prior weight makes the graduation mechanism visible. It records how much of the published forecast still depends on analogue, expert or other borrowed evidence under the approved updating rule. There is no universal decay schedule. New target observations can be weak, inconsistent or generated under changed conditions, in which case the range may widen and the borrowed component may retain influence. What matters is that the weight and its reason can be inspected.

Exhibit 66, p. 222 extends the illustrative worked example. Influence changes with comparable available exposure under the stated weighting rule. It does not change merely because time passes. This distinction prevents a planned phase date from masquerading as statistical confidence.

EXHIBIT 66 / ILLUSTRATIVE PERCENT OF WEIGHTED-RATE INFLUENCE

Comparable exposure changes weight under an explicit rule



ILLUSTRATIVE Illustrative rule from the worked example: prior strength is two comparable available weeks. At observed exposure n = 0, 1, 3 and 5 weeks, prior weight is 100 × 2 / (2+n); owned weight is 100 minus prior weight. Values rounded to two decimals. This is neither a posterior uncertainty calculation nor a prescribed learning trajectory. Changed offers or availability invalidate comparability.

Adoption and operating health complete the picture. Override rate counts changed eligible forecast cells relative to all eligible published cells, with reasons and later outcomes retained. Freshness measures the age of the oldest required valid input at the forecast cut-off; an absent required input is missing, not fresh. Rollback events remain in the log with their trigger and consequence; zero rollbacks is not a target when reversion is a normal safety control. Planner adoption is assessed through actual use in the agreed planning rhythm, not through access to a dashboard.

Decision outcome requires the most careful wording. The team specifies in advance which planning result the forecast is meant to improve and what comparison is credible. A before-and-after change may describe association without proving that the model caused it. Where a controlled pilot or valid replay supports a stronger comparison, the report states that design. Engineering guidance explicitly rejects a model change whose statistical gain degrades system performance, so the planning result has a vote in release. [S33]

Every measure receives an owner, formula, segment, review cadence and action threshold. Human oversight guidance leaves the choice of trustworthiness metrics and their precise thresholds to human judgement within the relevant context. [S36] The roadmap therefore imports no generic improvement percentage as a venture target. Thresholds are set from pilot variation, decision consequences and operating capacity, then approved before the next release.

Exhibit 67, p. 223 converts those requirements into the management artefact. This illustrative GFT dictionary combines established measurement ideas with proposed roles and cadences. It is a management design, not an externally observed scorecard.

EXHIBIT 67 / WORKING DEFINITIONS, OWNERS AND REVIEW RHYTHM

The final measurement dictionary turns recommendations into management

Measure	Working rule	Owner and cadence
MASE	Mean absolute error divided by defined nonzero naive scale	Model owner, by release
MAE and bias	Absolute error, plus mean forecast minus actual with sign retained	Demand owner, by planning cycle
Quantile loss	Proper quantile score by horizon and decision segment	Model owner, by release
Coverage and width	Actuals inside interval / eligible outcomes; mean upper minus lower bound	Model owner, by release
Stockout-affected share	Unavailable eligible cell-hours divided by observed eligible cell-hours	Data owner, by planning cycle
Prior weight	Borrowed-evidence weight under the approved update rule	Demand owner, at every revision
Override rate	Changed eligible forecast cells divided by all eligible cells; reasons retained	Planning owner, by planning cycle
Freshness	Cut-off minus oldest required valid input timestamp; missing stays missing	Data owner, at publication

ILLUSTRATIVE [S06] [S1404] [S1403] [S33] [S36] [S40] Illustrative GFT management design. MASE is undefined without a positive naive scale. Bias is forecast minus actual. Coverage needs width and eligible sample support. Stockout share describes observed conditions, not recovered demand. Prior weight follows the approved update rule. Proposed owners and cadences are role assignments, not GFT staffing claims. Decision outcomes need a stated comparison design.

EXHIBIT 68 / WORKING DEFINITIONS, OWNERS AND REVIEW RHYTHM

The final measurement dictionary turns recommendations into management, continued

Measure	Working rule	Owner and cadence
Decision outcome	Predefined planning result under a stated comparison design	Sponsor and planner, at gate review
Rollback	Event, trigger, restoration and consequence retained in release log	Service owner, at every event

ILLUSTRATIVE [S06] [S1404] [S1403] [S33] [S36] [S40] Illustrative GFT management design. MASE is undefined without a positive naive scale. Bias is forecast minus actual. Coverage needs width and eligible sample support. Stockout share describes observed conditions, not recovered demand. Prior weight follows the approved update rule. Proposed owners and cadences are role assignments, not GFT staffing claims. Decision outcomes need a stated comparison design.

The dictionary is a control surface rather than a reporting inventory. When a threshold is breached, the named owner chooses the pre-agreed action: hold the release, widen the range, repair capture, repeat evaluation or restore the baseline. The action and its later result enter the register. The cold-start-forecast tool should export the approved range and provenance. A ramp-planner can then test production scenarios; a safety-stock calculation still needs its own lead-time and service assumptions. The kpi-cascade should retain measure owners and responses. These are proposed handoffs, not a claim that a tool has been deployed or validated.

14.7

Decision gates keep uncertainty explicit for sponsors and planners

A gate is a decision meeting supported by a reproducible evidence pack. It is not a demonstration of the preferred model. The pack contains the forecast contract, prior register and sensitivity analysis; the pilot charter and observed conditions; the frozen corpus and launch-replay design; candidate and baseline results; calibration and residual uncertainty; operating readiness; and the proposed reversal conditions. Reviewers can then distinguish what was observed, what was derived and what remains an assumption.

Prior sensitivity belongs at every gate, not only before launch. Post-hoc calibration procedures can support calibrated uncertainty when enough suitable data are available, which makes the sufficient-data qualification part of the review. [S15] If the interval changes materially under a defensible alternative prior, analogue pool or calibration window, the sponsor should see that variation. A single selected specification would hide an important source of doubt.

The pilot section of the pack preserves the difference between designed and observed conditions. It records assignment, actual exposure, availability, returns, deviations and any human interventions. Controlled design supports causal interpret-

ation only within its valid assignment and analysis conditions. [S32] Where those conditions failed, the gate can still approve learning or another pilot, but it should not approve a causal claim or a production scale-up that the design cannot support.

The backtest section shows who and what was held out. Independent launch counts appear beside row counts. As-of feature snapshots, preprocessing fits and calibration windows follow the forecast origin. Results are segmented by horizon, product family, channel and availability state. The sponsor should be able to see whether a candidate won on the deployment population or only on a convenient aggregate. Any excluded launch carries a reason fixed before the final comparison.

Operational readiness receives the same scrutiny as statistical performance. The pack identifies the service owner, data owner, planning owner and release approver. It includes freshness checks, runtime, failure behaviour, monitoring, fallback and a tested restoration path. It also states which supplier components are used, what data leave the venture's control and how the record can be exported. These are release conditions rather than procurement details to be resolved after planners become dependent on the output.

Each gate records an explicit decision. **Proceed** authorises the next bounded phase. **Hold** retains the current prior, baseline or service while a named evidence gap is repaired. **Narrow** limits the release to the products, channels or horizons that passed. **Reverse** restores the last approved state and records the trigger. The decision log names the owner, evidence version, conditions and review point. Silence or elapsed time never counts as approval.

Sponsors also receive the unresolved uncertainties in decision language. The pack states which commitments would change under the low, central and high cases, which assumptions drive the difference and what evidence could resolve it. It does not convert a predictive interval into a promise of sales, profit or investment return. This chapter offers a planning and measurement method, not investment, legal or tax advice. A specific venture commitment requires its own commercial, financial, legal and operational review.

14.8

The research agenda builds Cameroon apparel evidence rather than importing effect sizes

The durable research asset is a relevant, well-governed launch corpus. Fashion forecasting research describes uncertainty, missing history and seasonality as recurring conditions, but that general framing does not supply effect sizes for a particular Cameroon apparel venture. [S1401] Visual-similarity research likewise offers a candidate transfer mechanism based on corrected demand from similar historical products, not proof that a new local manufacturer will inherit the reported performance. [S05] Local evidence must test both the mechanism and its limits.

Future work should assemble anonymised launch distributions with clear product, price, channel and timing descriptors. Stockout annotations and listing intervals should accompany sales so that later researchers can distinguish observed purchases from latent-demand corrections. Attribute taxonomies should be stable enough for

comparison while retaining version history when the offer changes. Channel-specific validation should separate consumer sell-through, distributor orders and institutional commitments rather than pooling them into one demand label.

The corpus should preserve disappointing and constrained launches. Exclusion rules may remove unusable records from an analysis, but the register retains their existence and exclusion reason. Otherwise published evidence will overstate launch predictability.

Sharing requires rights and confidentiality to be designed before aggregation. An anonymised research product should remove direct identifiers, assess re-identification risk from rare orders or small channels and document the transformation applied to every released field. Commercially sensitive price, customer and supplier details need not be published to make the method inspectable. Aggregated distributions, data dictionaries and validation designs can contribute evidence while keeping the underlying demand record under its owner's control.

Policy context must remain outside the demand label. The inherited sector register records about 95 percent of Cameroon's cotton fibre exported raw and about 5 percent processed locally, with the original publication date not recorded; it separately records a national target of 50 percent local processing by 2030. [R06] [R07] These figures explain an industrial ambition, not the sales distribution of a garment, a channel or a new firm. They are inherited sector observations and targets, not current GFT operating results.

The research question is whether local-processing ambition can be matched by repeatable demand learning. Evidence must come from actual offers, availability and buying behaviour. Publish where models fail, where stock censoring blocks inference and where simple baselines remain sufficient.

For GFT's Cameroon textile programme, the recommended close is therefore practical. Build the contract and prior, run the reversible pilot, protect the launch record, compare candidates on whole launches and operate only what proves useful. Publish or share evidence only where rights and confidentiality permit. Keep the national ambition visible as context while letting owned observations decide the next forecast revision.

A forecast graduates when owned evidence can challenge it, and the plan can still reverse it.

CHAPTER SOURCE NOTES

S1402 MIT Sloan School of Management, author manuscript. 2020-04-29.

S32 Springer Nature, Data Mining and Knowledge Discovery. 2008-07-30.

S06 Lancaster University Management School, Munich Personal RePEc Archive. 2018-11-19.

S37 Cell Press, Patterns; open full text from PubMed Central. 2023-08-04.

S39 Neural Information Processing Systems Foundation, NIPS 2015. 2015.

S02 arXiv. 2019-06-27.

S28 Journal of Industrial Ecology, Wiley. 2022-10-17.

S07 arXiv. 2026-06-09.

S18 International Journal of Forecasting, Elsevier; abstract through Lancaster EPrints. 2005.

S22 arXiv. 2021-03-26.

S21 arXiv; journal version in International Journal of Forecasting, Elsevier. 2019-02-22.

S25 arXiv. 2024-11-04.

S33 Google for Developers. 2025-08-25.

S40 Google Cloud Documentation. 2024-08-28.

S1405 Monash University, authors' research record for Journal of the American Statistical Association. 2019.

S38 ACM Computing Surveys; open record from Bournemouth University. 2014.

S1404 Rob J Hyndman and Anne B Koehler, author manuscript. 2005-11-02.

S1403 American Statistical Association, author copy hosted by University of Washington. 2007-03.

S36 National Institute of Standards and Technology, U.S. Department of Commerce. 2023.

S15 Proceedings of the 35th International Conference on Machine Learning, PMLR. 2018-07.

S1401 International Journal of Engineering Business Management, original paper hosted by HRČAK. 2013.

S05 Springer Nature, International Journal of Data Science and Analytics. 2026-01-08.

R06 Cameroon Trade Hub. Publication date unrecorded; inherited source-text check.

R07 Ecofin Agency. Publication date unrecorded; inherited source-text check.

Provenance travels with the claim.

Available inherited records and all chapter-local source records are graded below. Identical source identities are consolidated without losing citation aliases. Missing annex references are excluded. Unrecorded dates are disclosed rather than invented.

R01 Inherited evidence R01 GRADE C

Business in Cameroon. Publication date unrecorded. businessincameroon.com

Cameroon seed-cotton production 2022-2024: 330,000 t (2022), 314,455 t (2023), about 340,000 t (2024).

Inherited source-text check; original source not independently re-opened for this edition.

R02 Inherited evidence R02 GRADE C

Fibre2Fashion; Business in Cameroon. Publication date unrecorded. fibre2fashion.com; businessincameroon.com

2025 seed-cotton projection above 350,000 t.

Inherited source-text check; original source not independently re-opened for this edition.

R03 Inherited evidence R03 GRADE C

SODECOTON. Publication date unrecorded. sodecoton.cm

Company profile: more than 220,000 cotton producers supported in northern Cameroon; nine ginning plants; two oil mills.

Inherited source-text check; original source not independently re-opened for this edition.

R04 Inherited evidence R04 GRADE C

Financial Fortune Media. Publication date unrecorded. financialfortunemedia.com

SODECOTON injects roughly US$100 million a year into rural communities.

Inherited source-text check; original source not independently re-opened for this edition.

R05 Inherited evidence R05 GRADE C

Ecofin Agency. Publication date unrecorded. ecofinagency.com

CFA-zone cotton output 2.3 million t in 2024/25, down 11.5 percent.

Inherited source-text check; original source not independently re-opened for this edition.

R06 Inherited evidence R06 GRADE C

Cameroon Trade Hub. Publication date unrecorded. cameroontradehub.cm

About 95 percent of cotton fibre exported raw against 5 percent processed locally; industrialisation target of 50 percent local processing.

Inherited source-text check; original source not independently re-opened for this edition.

R07 Inherited evidence R07 GRADE C

Ecofin Agency. Publication date unrecorded. ecofinagency.com

SND30 restates the 50 percent local-processing target by 2030.

Inherited source-text check; original source not independently re-opened for this edition.

R08 Inherited evidence R08 GRADE C

Business in Cameroon. Publication date unrecorded. businessincameroon.com

Announced SODECOTON and CICAM merger to create a cotton-textile industrial pole targeted by 2030.

Inherited source-text check; original source not independently re-opened for this edition.

R09 Inherited evidence R09 GRADE C

Ecofin Agency. Publication date unrecorded. ecofinagency.com

Camtext SA integrated cotton-to-textile project, US$300 million (180 billion XAF); Panafritex and Arise with SODECOTON, CNPS and Marlo; 12 November 2025.

Inherited source-text check; original source not independently re-opened for this edition.

R10 Inherited evidence R10 GRADE C

Ecofin Agency. Publication date unrecorded. ecofinagency.com

Camtext SA: 12,000 t a year processed; about 12,000 jobs at the Dibamba hub and 3,000 at Garoua; return on investment near 11 years; optimal capacity around 2033.

Inherited source-text check; original source not independently re-opened for this edition.

R11 Inherited evidence R11 GRADE C

Business in Cameroon; MMI News. Publication date unrecorded. businessincameroon.com; mimimefoinfos.com

CICAM market share from about 85 percent three decades ago to about 5 percent today.

Inherited source-text check; original source not independently re-opened for this edition.

R12 Inherited evidence R12 GRADE C

Business in Cameroon. Publication date unrecorded. businessincameroon.com

CICAM capacity utilisation, July 2023 report: Garoua 33.7 percent, CICAM 1 48.9 percent, CICAM 2 10.83 percent; underutilisation cost about CFAF 2 billion in 2020.

Inherited source-text check; original source not independently re-opened for this edition.

R13 Inherited evidence R13 GRADE C

Business in Cameroon; MMI News. Publication date unrecorded. businessincameroon.com; mimimefoinfos.com

CICAM restructuring cost estimate CFAF 30.7 to 48.2 billion; one relaunch plan cited at 70.2 billion.

Inherited source-text check; original source not independently re-opened for this edition.

R14 Inherited evidence R14 GRADE C

WageIndicator. Publication date unrecorded. wageindicator.org

Cameroon minimum wage (SMIG) XAF 41,875 a month in 2023, raised to XAF 43,969 a month from 23 February 2024, about US$76, a rise of 5 percent.

Inherited source-text check; original source not independently re-opened for this edition.

R15 Inherited evidence R15 GRADE C

wage.is. Publication date unrecorded. wage.is/cameroon

Average salary across sectors in Cameroon about US$359 a month.

Inherited source-text check; original source not independently re-opened for this edition.

R16 Inherited evidence R16 GRADE C

WageIndicator. Publication date unrecorded. wageindicator.org

Informal share of employment about 90 percent; standard 40-hour week, 48-hour maximum, overtime at 1.2 times.

Inherited source-text check; original source not independently re-opened for this edition.

R17 Inherited evidence R17 GRADE C

just-style. Publication date unrecorded. just-style.com

Source metadata retained for traceability. This entry is not cited in the chapter text.

Inherited source-text check; original source not independently re-opened for this edition.

R18 Inherited evidence R18 GRADE B

Congressional Research Service IF10149; AGOA (Wikipedia). Publication date unrecorded. congress.gov; en.wikipedia.org

AGOA: Cameroon terminated effective 1 January 2020 for gross human-rights violations; remained ineligible; listed ineligible for calendar year 2025.

Inherited source-text check; original source not independently re-opened for this edition.

R23 Inherited evidence R23 GRADE C

MacMap; AfCFTA. Publication date unrecorded. macmap.org

Textiles and clothing among unresolved AfCFTA tariff-line categories; 92.3 percent of lines agreed excluding vehicles, textiles and clothing.

Inherited source-text check; original source not independently re-opened for this edition.

R24 Inherited evidence R24 GRADE C

Mordor Intelligence. Publication date unrecorded. mordorintelligence.com

Africa textile market about US$37.3 billion in 2025, projected US$49.41 billion by 2030 at 4.73 percent a year.

Inherited source-text check; original source not independently re-opened for this edition.

R25 Inherited evidence R25 GRADE C

Market.us. Publication date unrecorded. market.us

Africa textile market to about US$72.7 billion by 2035 at 6.9 percent a year; fashion and apparel 48.1 percent of the market in 2024.

Inherited source-text check; original source not independently re-opened for this edition.

R26 Inherited evidence R26 GRADE C

Statista. Publication date unrecorded. statista.com

Cameroon clothing and footwear spend about US$94.70 a person in 2024 and about 8.0 pieces a person; apparel eCommerce revenue US$20.32 million (2023) to US$33.61 million (2027) at 13.41 percent a year; user penetration 18.0 percent (2023) to 22.7 percent (2027).

Inherited source-text check; original source not independently re-opened for this edition.

R27 Inherited evidence R27 GRADE B

McKinsey, The State of Fashion 2025. Publication date unrecorded. mckinsey.com

AI at scale in fashion: inventory cost down 35 to 40 percent; forecast accuracy up 20 to 30 percent; revenue per customer up 15 to 25 percent; forecast errors cut 20 to 50 percent; lost sales down up to 65 percent; 75 percent of executives prioritising; sector operating-profit potential US$150 to 275 billion over three to five years.

Inherited source-text check; original source not independently re-opened for this edition.

R28 Inherited evidence R28 GRADE C

DataIntelo. Publication date unrecorded. dataintelo.com

Source metadata retained for traceability. This entry is not cited in the chapter text.

Inherited source-text check; original source not independently re-opened for this edition.

R29 Inherited evidence R29 GRADE C

MDPI; industry inspection literature. Publication date unrecorded. mdpi.com

Manual defect inspection 60 to 70 percent accurate, missing 20 to 30 percent of defects; computer vision above 90 percent across more than 40 defect types; line-scan inspection up to 60 metres a minute.

Inherited source-text check; original source not independently re-opened for this edition.

R30 Inherited evidence R30 GRADE C

Design Journal (Taylor and Francis). Publication date unrecorded. tandfonline.com

Source metadata retained for traceability. This entry is not cited in the chapter text.

Inherited source-text check; original source not independently re-opened for this edition.

R31 Inherited evidence R31 GRADE C

arXiv 2509.04705. Publication date unrecorded. arxiv.org/abs/2509.04705

Source metadata retained for traceability. This entry is not cited in the chapter text.

Inherited source-text check; original source not independently re-opened for this edition.

R32 Inherited evidence R32 GRADE C

Wise; XE. Publication date unrecorded. wise.com; xe.com

CFA franc to US dollar reference rates; franc pegged to the euro at 655.957. The GFT study's method note places the franc near 590 to 610 per US dollar across the 2023 study window and near 563 to 577 more recently, and treats currency conversions as estimates.

Inherited source-text check; original source not independently re-opened for this edition.

c01-S01 Demand Forecasting in the Fashion Industry: A Review GRADE B

InTech, International Journal of Engineering Business Management. 2013. https://iris.luiss.it/retrieve/handle/11385/76460/1076/InTech-Demand_forecasting_in_the_fashion_industry_a_review.pdf

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S02 Fashion Retail: Forecasting Demand for New Items GRADE B

arXiv. 2019-06-27. https://arxiv.org/abs/1907.01960

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S03 Product age based demand forecast model for fashion retail GRADE B

arXiv. 2020-07-10. https://arxiv.org/abs/2007.05278

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S04 Demand Forecasting for Multichannel Fashion Retailers by Integrating Clustering and Machine Learning Algorithms GRADE B

MDPI, Processes. 2021-09-03. https://www.mdpi.com/2227-9717/9/9/1578

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S05 Deep learning for new fashion product demand prediction: integrating visual similarity and demand correction in cold-start scenarios GRADE B

Springer Nature, International Journal of Data Science and Analytics. 2026-01-08. https://link.springer.com/article/10.1007/s41060-025-00888-8

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S06 Retail forecasting: research and practice GRADE B

Lancaster University Management School, Munich Personal RePEc Archive. 2018-11-19. https://mpra.ub.uni-muenchen.de/89356/1/MPRA_paper_89356.pdf

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S07 FreshRetailNet-LT: A Stockout-Annotated Censored Demand Dataset for Latent Demand Recovery and Forecasting in Fresh Retail GRADE B

arXiv. 2026. https://arxiv.org/html/2505.16319v3

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S08 Forecasting: theory and practice GRADE B

International Journal of Forecasting, Elsevier; full text through arXiv. 2022. https://arxiv.org/html/2012.03854

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S09 Statistical and Machine Learning forecasting methods: Concerns and ways forward GRADE B

PLOS ONE, Public Library of Science. 2018-03-27. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0194889

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S10 The M4 Competition: Results, findings, conclusion and way forward GRADE B

International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2018. https://ideas.repec.org/a/eee/intfor/v34y2018i4p802-808.html

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S11 **M5 accuracy competition: Results, findings, and conclusions** GRADE B

International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2022. https://ideas.repec.org/a/eee/intfor/v38y2022i4p1346-1364.html

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S12 **Another look at measures of forecast accuracy** GRADE B

International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2006. https://ideas.repec.org/a/eee/intfor/v22y2006i4p679-688.html

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S13 **3.1. Cross-validation: evaluating estimator performance** GRADE B

scikit-learn. 2026-10-30. https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation-of-time-series-data

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S14 **Strictly Proper Scoring Rules, Prediction, and Estimation** GRADE B

Journal of the American Statistical Association, American Statistical Association. 2007. https://www.stat.washington.edu/raftery/Research/PDF/Gneiting2007jasa.pdf

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S15 **Accurate Uncertainties for Deep Learning Using Calibrated Regression** GRADE B

Proceedings of the 35th International Conference on Machine Learning, PMLR. 2018-07-03. https://proceedings.mlr.press/v80/kuleshov18a.html

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S16 **The M5 uncertainty competition: Results, findings and conclusions** GRADE B

International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2022. https://ideas.repec.org/a/eee/intfor/v38y2022i4p1365-1385.html

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S17 **Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization** GRADE B

Journal of the American Statistical Association, Taylor & Francis; abstract through RePEc IDEAS. 2019. https://ideas.repec.org/a/taf/jnlasa/v114y2019i526p804-819.html

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S18 **The accuracy of intermittent demand estimates** GRADE B

International Journal of Forecasting, Elsevier; abstract through Lancaster EPrints. 2005. https://eprints.lancs.ac.uk/id/eprint/73097/

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S20 **Forecast combinations: an over 50-year review** GRADE B

arXiv; journal version in International Journal of Forecasting, Elsevier. 2022-09-23. https://arxiv.org/html/2205.04216

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S21 **DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks** GRADE B

arXiv; journal version in International Journal of Forecasting, Elsevier. 2019-02-22. https://arxiv.org/abs/1704.04110

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S22 **Principles and Algorithms for Forecasting Groups of Time Series: Locality and Globality** GRADE B

arXiv. 2021-03-26. https://arxiv.org/abs/2008.00444

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S25 **Chronos: Learning the Language of Time Series** GRADE B

arXiv. 2024-11-04. https://arxiv.org/abs/2403.07815

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S26 **A decoder-only foundation model for time-series forecasting** GRADE B

arXiv. 2024-04-17. https://arxiv.org/abs/2310.10688

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S28 **A hierarchical Bayesian regression model that reduces uncertainty in material demand predictions** GRADE B

Journal of Industrial Ecology, Wiley. 2022-10-17. https://onlinelibrary.wiley.com/doi/10.1111/jiec.13339

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S29 **Delphi Method** GRADE B

RAND Corporation. 2023-12-29. https://www.rand.org/topics/delphi-method.html

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S30 **A New Product Growth for Model Consumer Durables** GRADE B

Management Science, INFORMS; abstract through EconPapers. 1969. https://econpapers.repec.org/RePEc:inm:ormnsc:v:15:y:1969:i:5:p:215-227

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S31 **Preference Learning and Demand Forecast** GRADE B

INFORMS, Marketing Science; open manuscript from MIT. 2020. https://hdl.handle.net/1721.1/130372

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S32 **Controlled experiments on the web: survey and practical guide** GRADE B

Springer Nature, Data Mining and Knowledge Discovery. 2008-07-30. https://link.springer.com/article/10.1007/s10618-008-0114-1

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S33 **Rules of Machine Learning: Best Practices for ML Engineering** GRADE C

Google for Developers. 2025-08-25. https://developers.google.com/machine-learning/guides/rules-of-ml

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S36 **Artificial Intelligence Risk Management Framework (AI RMF 1.0)** GRADE A

National Institute of Standards and Technology, U.S. Department of Commerce. 2023. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S37 **Leakage and the reproducibility crisis in machine-learning-based science** GRADE B

Cell Press, Patterns; open full text from PubMed Central. 2023-08-04. https://pmc.ncbi.nlm.nih.gov/articles/PMC10499856/

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S38 **A Survey on Concept Drift Adaptation** GRADE B

ACM Computing Surveys; open record from Bournemouth University. 2014-04-30. https://eprints.bournemouth.ac.uk/22491/

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S39 **Hidden Technical Debt in Machine Learning Systems** GRADE B

Neural Information Processing Systems Foundation, NIPS 2015. 2015. https://papers.nips.cc/paper_files/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S40 **MLOps: Continuous delivery and automation pipelines in machine learning** GRADE C

Google Cloud Documentation. 2024-08-28. https://docs.cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S0101 **On the use of learning-based forecasting methods for ameliorating fashion business processes: A position paper** GRADE B

arXiv. 2022-11-09. https://arxiv.org/abs/2211.04798

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c01-S0102 **Forecasting: Principles and Practice (3rd ed), Chapter 6: Judgmental forecasts** GRADE B

OTexts. 2021-05-31. https://otexts.com/fpp3/judgmental.html

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c02-S01 Same source record as c01-S01.

c02-S02 Same source record as c01-S02.

c02-S03 Same source record as c01-S03.

c02-S04 Same source record as c01-S04.

c02-S05 Same source record as c01-S05.

c02-S06 Same source record as c01-S06.

c02-S07 Same source record as c01-S07.

c02-S08 Same source record as c01-S08.

c02-S11 Same source record as c01-S11.

c02-S12 Same source record as c01-S12.

c02-S13 Same source record as c01-S13.

c02-S14 Same source record as c01-S14.

c02-S15 Same source record as c01-S15.

c02-S16 Same source record as c01-S16.

c02-S17 Same source record as c01-S17.

c02-S18 Same source record as c01-S18.

c02-S19 On the bias of intermittent demand estimates GRADE B

International Journal of Production Economics, Elsevier; abstract through Lancaster EPrints. 2001. https://eprints.lancs.ac.uk/id/eprint/73104/

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c02-S22 Same source record as c01-S22.

c02-S24 Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting GRADE B

arXiv. 2020-09-27. https://arxiv.org/abs/1912.09363

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c02-S28 Same source record as c01-S28.

c02-S33 Same source record as c01-S33.

c02-S36 Same source record as c01-S36.

c02-S37 Same source record as c01-S37.

c02-S39 Same source record as c01-S39.

c02-S40 Same source record as c01-S40.

c02-S0201 Information Distortion in a Supply Chain: The Bullwhip Effect GRADE B

Management Science, INFORMS; abstract through RePEc IDEAS. 1997-04. https://ideas.repec.org/a/inm/ormnsc/v43y1997i4p546-558.html

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c02-S0202 Forecasting: Principles and Practice (3rd ed), Chapter 1: Getting started GRADE B

OTexts. 2021-05-31. https://otexts.com/fpp3/perspective.html

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c02-S0203 **Forecasting: Principles and Practice (3rd ed), Chapter 13: Some practical forecasting issues** GRADE B

OTexts. 2021-05-31. https://otexts.com/fpp3/counts.html

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

---

c03-S01 Same source record as c01-S01.

---

c03-S02 Same source record as c01-S02.

---

c03-S03 **Product age based demand forecast model for fashion retail** GRADE B

arXiv. 2020-07-10. https://arxiv.org/pdf/2007.05278

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

---

c03-S04 Same source record as c01-S04.

---

c03-S05 Same source record as c01-S05.

---

c03-S08 Same source record as c01-S08.

---

c03-S14 **Strictly Proper Scoring Rules, Prediction, and Estimation** GRADE B

Journal of the American Statistical Association, American Statistical Association. 2007-03. https://www.stat.washington.edu/raftery/Research/PDF/Gneiting2007jasa.pdf

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

---

c03-S22 Same source record as c01-S22.

---

c03-S28 **A hierarchical Bayesian regression model that reduces uncertainty in material demand predictions** GRADE B

Journal of Industrial Ecology, Wiley; abstract through RePEc IDEAS. 2023-02. https://ideas.repec.org/a/bla/inecol/v27y2023i1p43-55.html

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

---

c03-S30 Same source record as c01-S30.

---

c03-S44 **Population, total** GRADE A

World Bank, World Development Indicators. 2026-10-30. https://databank.worldbank.org/metadataglossary/world-development-indicators/series/SP.POP.TOTL

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c03-S45 GNI per capita (constant 2015 US$) GRADE A

World Bank, World Development Indicators. 2026-10-30. https://databank.worldbank.org/metadataglossary/world-development-indicators/series/NY.GNP.PCAP.KD

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c03-S46 Urban population (% of total population) GRADE A

World Bank, World Development Indicators. 2026-10-30. https://databank.worldbank.org/metadataglossary/world-development-indicators/series/SP.URB.TOTL.IN.ZS

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c04-S01 Same source record as c01-S01.

c04-S02 Same source record as c01-S02.

c04-S03 Same source record as c01-S03.

c04-S04 Same source record as c01-S04.

c04-S06 Same source record as c01-S06.

c04-S08 Same source record as c01-S08.

c04-S09 Same source record as c01-S09.

c04-S10 Same source record as c01-S10.

c04-S12 Same source record as c01-S12.

c04-S13 Same source record as c01-S13.

c04-S24 Same source record as c02-S24.

c04-S28 Same source record as c01-S28.

c04-S33 Same source record as c01-S33.

c04-S34 Predicting consumer behavior with Web search GRADE B

Proceedings of the National Academy of Sciences. 2010. https://pmc.ncbi.nlm.nih.gov/articles/PMC2955127/

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c04-S35 The Effects of Weather on Retail Sales GRADE A

Board of Governors of the Federal Reserve System, Finance and Economics Discussion Series. 2000. https://www.federalreserve.gov/econres/feds/the-effects-of-weather-on-retail-sales.htm

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

---

c04-S36 Same source record as c01-S36.

---

c04-S37 Same source record as c01-S37.

---

c04-S38 Same source record as c01-S38.

---

c04-S39 Same source record as c01-S39.

---

c04-S40 Same source record as c01-S40.

---

c04-S41 What are the data sources for UN Comtrade? GRADE A

United Nations Statistics Division, UN Comtrade Help Center. 2026-10-30. https://uncomtrade.org/docs/what-are-the-data-sources-for-un-comtrade/

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

---

c04-S42 Bilateral asymmetries GRADE A

United Nations Statistics Division, UN Comtrade Help Center. 2026-10-30. https://uncomtrade.org/docs/bilateral-asymmetries/

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

---

c04-S43 Trade Map User Guide GRADE A

International Trade Centre. 2014. https://legacy.trademap.org/Docs/TradeMap-Userguide-EN.pdf

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

---

c04-S44 Same source record as c03-S44.

---

c04-S45 Same source record as c03-S45.

---

c04-S46 Same source record as c03-S46.

---

c04-S47 Agricultural production statistics 2010-2024 GRADE A

Food and Agriculture Organization of the United Nations. 2025. https://www.fao.org/statistics/highlights-archive/highlights-detail/agricultural-production-statistics-2010-2024/en

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c04-S0401 **Forecasting: Principles and Practice (3rd ed), Section 7.6: Forecasting with regression** GRADE B

OTexts. 2021-05-31. https://otexts.com/fpp3/regression.html

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

---

c04-S0402 Same source record as c01-S0101.

---

c04-S0403 **Forecasting: Principles and Practice (3rd ed), Chapter 10: Dynamic regression models** GRADE B

OTexts. 2021-05-31. https://otexts.com/fpp3/forecasting.html

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

---

c05-S01 Same source record as c01-S01.

---

c05-S02 Same source record as c01-S02.

---

c05-S03 Same source record as c01-S03.

---

c05-S06 Same source record as c01-S06.

---

c05-S07 Same source record as c01-S07.

---

c05-S08 Same source record as c01-S08.

---

c05-S14 Same source record as c01-S14.

---

c05-S16 Same source record as c01-S16.

---

c05-S20 Same source record as c01-S20.

---

c05-S28 Same source record as c01-S28.

---

c05-S29 Same source record as c01-S29.

---

c05-S30 Same source record as c01-S30.

---

c05-S31 Same source record as c01-S31.

---

c05-S32 Same source record as c01-S32.

---

c05-S33 Same source record as c01-S33.

c05-S36 Same source record as c01-S36.

c05-S37 Same source record as c01-S37.

c05-S39 Same source record as c01-S39.

c05-S40 Same source record as c01-S40.

c05-S44 Same source record as c03-S44.

c05-S46 Same source record as c03-S46.

c05-S48 Forecasting future demand for large-screen television sets using conjoint analysis with diffusion model

GRADE B

Elsevier, Technological Forecasting and Social Change. 2006. https://www.sciencedirect.com/science/article/abs/pii/S0040162504001660

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c05-S0501 Forecasting: Principles and Practice (3rd ed.), Chapter 6: Judgmental forecasts GRADE B

OTexts, Monash University. 2026-10-30. https://otexts.com/fpp3/judgmental.html

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c05-S0502 Hypothetical Bias GRADE B

Centre for Evidence-Based Medicine, University of Oxford (Catalogue of Bias Collaboration). 2020. https://catalogofbias.org/biases/hypothetical-bias/

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c05-S0503 The preregistration revolution GRADE A

Proceedings of the National Academy of Sciences of the United States of America, PubMed Central. 2018-03-13. https://pmc.ncbi.nlm.nih.gov/articles/PMC5856500/

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c06-S04 Same source record as c01-S04.

c06-S06 Same source record as c01-S06.

c06-S07 Same source record as c01-S07.

c06-S11 Same source record as c01-S11.

c06-S28 Same source record as c01-S28.

c06-S32 Same source record as c01-S32.

c06-S33 Same source record as c01-S33.

c06-S36 Same source record as c01-S36.

c07-S02 Same source record as c01-S02.

c07-S03 Same source record as c01-S03.

c07-S04 Same source record as c01-S04.

c07-S05 Same source record as c01-S05.

c07-S06 Same source record as c01-S06.

c07-S07 Same source record as c01-S07.

c07-S18 Same source record as c01-S18.

c07-S19 Same source record as c02-S19.

c07-S34 Same source record as c04-S34.

c07-S35 Same source record as c04-S35.

c07-S36 Same source record as c01-S36.

c07-S37 Same source record as c01-S37.

c07-S39 Same source record as c01-S39.

c07-S40 Same source record as c01-S40.

c07-S41 Same source record as c04-S41.

c07-S42 Same source record as c04-S42.

c07-S0701 **Datasheets for Datasets** GRADE B

arXiv. 2021-12-01. https://arxiv.org/abs/1803.09010

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c07-S0702 **Infrastructure for Usable Machine Learning: The Stanford DAWN Project** GRADE B

arXiv. 2017-05-22. https://arxiv.org/abs/1705.07538

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c08-S02 Same source record as c01-S02.

c08-S03 Same source record as c01-S03.

c08-S08 Same source record as c01-S08.

c08-S09 Same source record as c01-S09.

c08-S10 Same source record as c01-S10.

c08-S12 Same source record as c01-S12.

c08-S18 Same source record as c01-S18.

c08-S19 Same source record as c02-S19.

c08-S20 Same source record as c01-S20.

c08-S30 Same source record as c01-S30.

c08-S33 Same source record as c01-S33.

c08-S39 Same source record as c01-S39.

c09-S02 Same source record as c01-S02.

c09-S05 Same source record as c01-S05.

c09-S21 Same source record as c01-S21.

c09-S22 Same source record as c01-S22.

c09-S23 **N-BEATS: Neural basis expansion analysis for interpretable time series forecasting** GRADE B

arXiv. 2020-02-20. https://arxiv.org/abs/1905.10437

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c09-S24 Same source record as c02-S24.

c09-S28 Same source record as c01-S28.

c09-S33 Same source record as c01-S33.

c09-S37 Same source record as c01-S37.

c09-S38 Same source record as c01-S38.

c09-S39 Same source record as c01-S39.

c10-S01 Same source record as c01-S01.

c10-S12 Same source record as c01-S12.

c10-S14 Same source record as c03-S14.

c10-S25 Same source record as c01-S25.

c10-S26 Same source record as c01-S26.

c10-S27 **Lag-Llama: Towards Foundation Models for Probabilistic Time Series Forecasting** GRADE B

arXiv. 2024-02-08. https://arxiv.org/abs/2310.08278

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c10-S33 Same source record as c01-S33.

c10-S37 Same source record as c01-S37.

c10-S38 Same source record as c01-S38.

c10-S39 Same source record as c01-S39.

c10-S40 Same source record as c01-S40.

c11-S01 Same source record as c01-S01.

c11-S02 Same source record as c01-S02.

c11-S03 Same source record as c01-S03.

c11-S04 Same source record as c01-S04.

c11-S06 Same source record as c01-S06.

c11-S07 Same source record as c01-S07.

c11-S08 Same source record as c01-S08.

c11-S11 Same source record as c01-S11.

c11-S12 Same source record as c01-S12.

c11-S14 Same source record as c01-S14.

c11-S15 Same source record as c01-S15.

c11-S16 Same source record as c01-S16.

c11-S17 Same source record as c01-S17.

c11-S18 Same source record as c01-S18.

c11-S19 Same source record as c02-S19.

c11-S21 Same source record as c01-S21.

c11-S24 Same source record as c02-S24.

c11-S25 Same source record as c01-S25.

c11-S28 Same source record as c01-S28.

c11-S31 Same source record as c01-S31.

c11-S33 Same source record as c01-S33.

c11-S36 Same source record as c01-S36.

c11-S37 Same source record as c01-S37.

c11-S38 Same source record as c01-S38.

c11-S39 Same source record as c01-S39.

c11-S40 Same source record as c01-S40.

c11-S1101 Conformalized Quantile Regression GRADE B

arXiv; Neural Information Processing Systems (NeurIPS 2019). 2019-05-08. https://arxiv.org/abs/1905.03222

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c11-S1102 Probabilistic forecast reconciliation under the Gaussian framework GRADE B

arXiv; Journal of Business & Economic Statistics. 2021-03-20. https://arxiv.org/abs/2103.11128

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c11-S1103 Reconciliation of probabilistic forecasts with an application to wind power GRADE B

arXiv; European Journal of Operational Research, Elsevier. 2018-08-08. https://arxiv.org/abs/1808.02635

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c12-S02 Same source record as c01-S02.

c12-S05 Same source record as c01-S05.

c12-S06 Same source record as c01-S06.

c12-S10 Same source record as c01-S10.

c12-S12 Same source record as c01-S12.

c12-S13 Same source record as c01-S13.

c12-S14 Same source record as c01-S14.

c12-S15 Same source record as c01-S15.

c12-S16 Same source record as c01-S16.

c12-S20 Same source record as c01-S20.

c12-S32 Same source record as c01-S32.

c12-S33 Same source record as c01-S33.

c12-S36 Same source record as c01-S36.

c12-S37 Same source record as c01-S37.

c12-S39 Same source record as c01-S39.

c12-S1201 Evaluating time series forecasting models: An empirical study on performance estimation methods

GRADE B

arXiv; journal version in Machine Learning, Springer. 2019-05-28. https://arxiv.org/abs/1905.11744

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c12-S1202 Evaluating predictive count data distributions in retail sales forecasting

GRADE B

International Journal of Forecasting, Elsevier; abstract through RePEc IDEAS. 2016. https://ideas.repec.org/a/eee/intfor/v32y2016i3p788-803.html

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c13-S06 Same source record as c01-S06.

c13-S12 Same source record as c01-S12.

c13-S14 Same source record as c01-S14.

c13-S16 Same source record as c01-S16.

c13-S29 Same source record as c01-S29.

c13-S33 Same source record as c01-S33.

c13-S36 Same source record as c01-S36.

c13-S37 Same source record as c01-S37.

c13-S38 Same source record as c01-S38.

c13-S39 Same source record as c01-S39.

c13-S40 Same source record as c01-S40.

c14-S02 Same source record as c01-S02.

c14-S05 Same source record as c01-S05.

c14-S06 Same source record as c01-S06.

c14-S07 FreshRetailNet-LT: A Stockout-Annotated Censored Demand Dataset for Latent Demand Recovery and Forecasting in Fresh Retail GRADE B

arXiv. 2026-06-09. https://arxiv.org/html/2505.16319v3

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c14-S15 Accurate Uncertainties for Deep Learning Using Calibrated Regression GRADE B

Proceedings of the 35th International Conference on Machine Learning, PMLR. 2018-07. https://proceedings.mlr.press/v80/kuleshov18a.html

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

c14-S18 Same source record as c01-S18.

c14-S21 Same source record as c01-S21.

c14-S22 Same source record as c01-S22.

c14-S25 Same source record as c01-S25.

c14-S28 Same source record as c01-S28.

c14-S32 Same source record as c01-S32.

c14-S33 Same source record as c01-S33.

c14-S36 Same source record as c01-S36.

c14-S37 Same source record as c01-S37.

c14-S38 A Survey on Concept Drift Adaptation GRADE B

ACM Computing Surveys; open record from Bournemouth University. 2014. https://eprints.bournemouth.ac.uk/22491/

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

---

c14-S39 Same source record as c01-S39.

---

c14-S40 Same source record as c01-S40.

---

c14-S1401 Demand Forecasting in the Fashion Industry: A Review GRADE B

International Journal of Engineering Business Management, original paper hosted by HRČAK. 2013. https://hrcak.srce.hr/en/160979

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

---

c14-S1402 Preference Learning and Demand Forecast GRADE B

MIT Sloan School of Management, author manuscript. 2020-04-29. https://mitsloan.mit.edu/shared/ods/documents?PublicationDocumentID=6366

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

---

c14-S1403 Strictly Proper Scoring Rules, Prediction, and Estimation GRADE B

American Statistical Association, author copy hosted by University of Washington. 2007-03. https://sites.stat.washington.edu/people/raftery/Research/PDF/Gneiting2007jasa.pdf

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

---

c14-S1404 Another look at measures of forecast accuracy GRADE B

Rob J Hyndman and Anne B Koehler, author manuscript. 2005-11-02. https://robjhyndman.com/papers/mase.pdf

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

---

c14-S1405 Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization GRADE B

Monash University, authors' research record for Journal of the American Statistical Association. 2019. https://research.monash.edu/en/publications/optimal-forecast-reconciliation-for-hierarchical-and-grouped-time/

Opened source record.

Accessed 2026-10-30. Quote retained in the machine-readable source record.

Stable definitions make comparisons useful.

Analogue market

A comparison market selected for documented economic, channel and consumer similarities, used to inform a prior rather than to assert identical demand.

Analogue product

A previously offered product whose attributes, launch conditions and outcomes provide transferable evidence for a new item.

As-of timestamp

The time at which a data value became available, used to prevent later information or revisions entering a historical forecast replay.

Backtest

A retrospective evaluation that recreates the information and decision conditions under which forecasts would have been produced.

Baseline forecast

A simple, explicit forecast against which every more elaborate candidate must demonstrate an improvement.

Bass diffusion model

A new-product adoption model that represents initial purchases through innovation and imitation parameters under specific market assumptions.

Bayesian hierarchical model

A model that estimates shared and local parameters together, allowing partial pooling across products, sites or markets.

Bias

A systematic tendency for forecasts to sit above or below realised values rather than merely to vary around them.

Calibration

The agreement between stated predictive probabilities or interval levels and the frequency with which outcomes occur.

Censored demand

Demand whose observation is limited by availability or another constraint, so recorded sales do not reveal the full amount wanted.

Cold start

A forecasting setting in which the target item, firm, channel or market lacks the history normally used to fit a model.

Concept drift

A change in the relationship between model inputs and the target that can reduce the validity of a previously fitted model.

Conjoint analysis

A research method that infers preferences for product attributes from structured trade-offs or choices.

Covariate

An input variable used alongside or instead of target history to inform a forecast.

Croston method

A family of intermittent-demand approaches that models nonzero demand sizes and the intervals between demand occurrences.

Demand event

A defined occurrence such as an order, fulfilment, sale, return or cancellation that may serve as a measurement target.

Domain shift

A mismatch between the setting in which a model was trained or benchmarked and the setting in which it must forecast.

Expanding-window validation

Chronological evaluation in which each successive training set adds later history and tests a subsequent period.

Forecast contract

The agreed specification of decision, target, grain, horizon, information cut-off, availability rule, metric and ownership.

Forecast distribution

The full set of possible demand outcomes with associated probabilities or quantiles.

Forecast horizon

The distance between the forecast origin and the future period whose demand is being predicted.

Forecast origin

The decision time at which the available information set ends and prediction begins.

Forecast value

The improvement in realised decisions attributable to a forecast after accounting for its operating costs and constraints.

Global model

A single forecasting function fitted across multiple series rather than independently for each series.

Hierarchical reconciliation

Adjustment of forecasts across aggregation levels so that component and total forecasts become coherent.

Intermittent demand

Demand that occurs sporadically, with some periods recording no demand.

Judgemental forecast

A forecast formed through structured human expertise, either alone or in combination with statistical evidence.

Latent demand

The unobserved amount customers would have wanted in the absence of availability or other observation constraints.

Leakage

Information that would be unavailable at the forecast origin entering training, features, preprocessing or evaluation and inflating apparent skill.

Local model

A forecasting function fitted independently to one product, location or other individual series.

MASE

Mean absolute scaled error, an accuracy measure that scales absolute forecast error by an available naive benchmark and is undefined when that denominator is absent or zero.

Override

A documented human change to a model forecast, retained with its reason and later outcome for evaluation.

Partial pooling

Sharing statistical strength across groups while allowing each group to retain its own parameters and uncertainty.

Prediction interval

A range intended to contain the realised value with a stated probability under the forecast model and its assumptions.

Prior

The initial evidence-based distribution or belief about demand before target observations are incorporated.

Probabilistic forecast

A forecast that represents uncertainty through probabilities, quantiles, intervals or a full distribution.

Product attributes

Stable descriptors such as category, fabric, colour, silhouette, price band and occasion used to compare or model items.

Quantile

A demand value below which a stated proportion of the predictive distribution falls.

Reference class

A documented set of comparable past cases whose outcome distribution informs an outside-view forecast.

Rolling-origin evaluation

Repeated backtesting from successive forecast origins to assess performance across time and information conditions.

Scenario

A conditional view of demand under specified changes to price, policy, promotion, supply or another assumption.

Sell-through

Sales to the final customer through a retail or distribution channel, distinct from shipments into that channel.

Seasonal-naive forecast

A baseline that repeats the value from the corresponding previous seasonal period.

SKU

Stock keeping unit, the identified product variant at which availability and sales are recorded.

Stockout

A period or event in which a product cannot be bought because available inventory is exhausted.

Transfer learning

Using knowledge learned from source products, markets or datasets to improve prediction in a different target setting.

Zero-shot forecast

A forecast for a target dataset without fitting on that target's observations, usually after pretraining on other data.

Follow the terms back to their context.

Page references point to occurrences in chapter prose. When a term does not occur in a chapter, its glossary definition is indexed.

Analogue market 54 69

Analogue product 251

As-of timestamp 63 147

Backtest 29 35 67 76 107 119 142 186 198 217 219 225

Baseline forecast 251

Bass diffusion model 251

Bayesian hierarchical model 251

Bias 18 22 32 45 48 74 76 85 92 94 114 129 134 135 150 151 164 175 178 179 187 192 193 196 205 209 211 219 221

Calibration 21 23 27 28 45 59 75 76 81 85 86 87 88 91 93 94 95 108 125 133 135 145 148 151 157 159 161 172 174 175 180 181 182 183 186 190 194 197 200 202 204 208 209 212 214 216 218 221 224 225

Censored demand 251

Cold start 17 19 21 57 172 176 221

Concept drift 207

Conjoint analysis 83 216

Covariate 69 76 77 78 121 148 215

Croston method 252

Demand event 45 125

Domain shift 161 171

Expanding-window validation 252

Forecast contract 19 23 32 34 36 39 40 44 45 87 105 111 118 126 158 192 214 215 221 224

Forecast distribution 24 181 194

Forecast horizon 40 187 214

Forecast origin 24 28 41 44 74 76 111 118 120 121 123 126 129 134 143 147 158 164 186 189 190 192 200 202 204 217 225

Forecast value 252

Global model 140 142 144 151 159 163 164 212

Hierarchical reconciliation 39 44 220

Intermittent demand 41 179

Judgemental forecast 252

Latent demand 22 32 33 34 36 37 38 42 44 45 78 89 115 183 217

Leakage 28 29 41 67 74 75 111 112 119 143 145 152 158 161 181 183 189 190 202 214 217

Local model 135 140

MASE 28 29 44 45 76 134 136 164 175 192 193 196 209 221

Override 29 165 195 200 202 205 206 222

Partial pooling 28 53 57 72 105 149 150 216 218

Prediction interval 92 147 151 180 202

Prior 16 17 18 19 21 23 24 25 27 28 32 33 34 36 38 39 45 48 49 54 55 56 57 58 59 60 63 64 65 78 81 82 83 85 86 87 88 89 91 92 93 94 95 98 101 106 107 109 119 122 123 125 131 133 135 136 142 145 146 148 150 153 155 159 164 165 170 175 180 181 182 183 187 188 189 190 196 197 208 214 215 216 217 218 219 221 222 224 225 226

Probabilistic forecast 173

Product attributes 19 71 75 134 158 159 186 202 207 222

Quantile 23 24 29 41 45 59 76 89 91 92 145 157 169 170 171 172 174 177 178 179 180 181 183 193 194 204 205 206 221

Reference class 48 50 51 52 55 58 90 117 132 136 180 215

Rolling-origin evaluation 253

Scenario 16 29 52 55 56 58 59 65 83 89 90 115 130 147 151 180 181 183 201 202 204 205 208

Seasonal-naive forecast 125

Sell-through 21 22 28 32 33 34 35 42 45 50 54 63 64 65 88 95 98 105 128 159 169 195 220 226

SKU 18 19 20 21 22 35 38 39 40 63 69 71 73 76 78 95 116 117 176

Stockout 17 18 20 21 22 23 24 25 28 33 34 36 37 39 44 45 49 53 60 89 90 104 114 115 125 128 145 151 172 174 178 181 183 194 195 205 214 217 221 225

Transfer learning 253

Zero-shot forecast 253

A reproducible edition keeps its limits visible.

Production is the warp. Measurement is the weft.

Green Future Tech prepared this white paper from published sources opened during research, each recorded with its URL, publisher, date and a quoted sentence in the evidence register, and from GFT's Cameroon textile programme evidence register (R entries, transcribed from the programme's data annex and flagged where refreshed in 2026). Derived figures show their inputs; illustrative models are labelled with their basis. No figure is presented as audited, and nothing here is legal, tax or investment advice.

Source treatment preserves uncertainty.

Inherited records reproduce the supplied finding and attribution only where cited. Their source-text status is preserved. Original publication dates and exact source URLs are incomplete in that register. These records require independent validation before external reliance. New chapter-local sources carry a publication date, complete URL, publisher, exact supporting quotation and access date.

The exports share a content model.

The PDF and web edition use the same parsed chapter blocks and SVG exhibits. The build records input hashes, artifact hashes, layout measurements and elapsed time. Cross-references, source notes and the term index are generated from identifiers rather than manually typed page numbers.

IMPRINT

Issued by Green Future Tech Limited, Hong Kong SAR.
Engineering the sustainable production of tomorrow.

GFT Research. Edition 2026-10-30. Document wp04. Locally generated proof. No named author, office, certification or client endorsement is implied.

GREEN FUTURE TECH

LIMITED / HONG KONG SAR

The factory that makes the garment should make the evidence.

Engineering the sustainable production of tomorrow.