ATM CASH MANAGEMENT IN VIETNAM

with **Uncertainty-Aware Forecasting**

and **Inventory Optimization**

FLINS-ISKE 2026 · Sydney, Australia · July 2026

Huu-Thanh Phan • **Xuan-Bach Le** • **Thanh-Tho Quan**
Ho Chi Minh City University of Technology (HCMUT), VNU-HCM

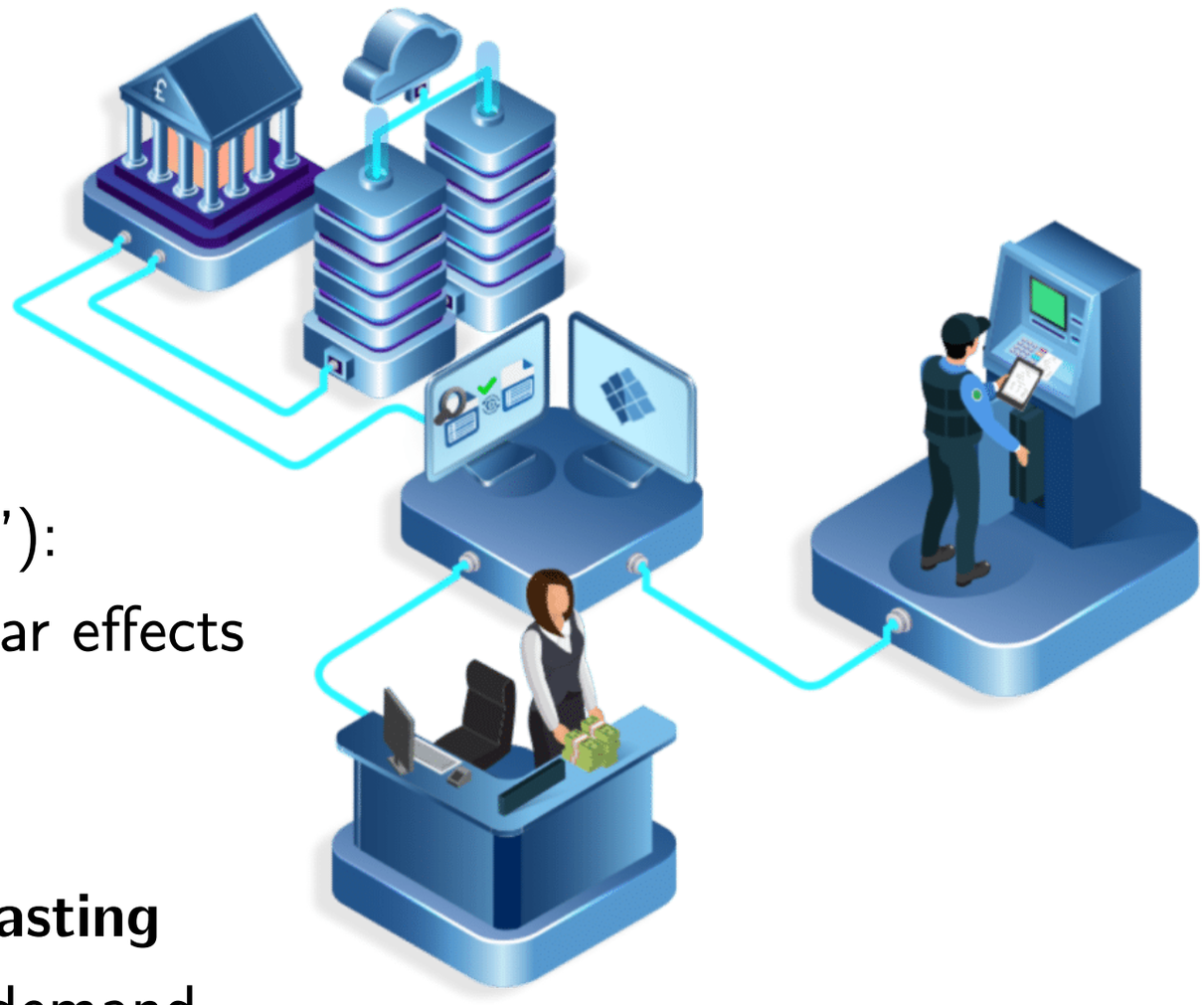
Next section

Motivation

ATMs remain critical cash-distribution infrastructure in Vietnam, ≈70% cash

Core trade-off:

- Understocking
  → Customer dissatisfaction
- Overstocking
  → High holding cost



Traditional heuristics
("refill every Monday", "top up at 20%"):

- Ignore stochastic demand and calendar effects
- Do not leverage historical data

What's needed: Decision-centric forecasting

- Lead time → must plan with future demand
- Asymmetric costs → tail risk → uncertainty matters

Research Gaps

1. **No systematic benchmark of global/deep models for ATM demand**
   Most studies: ARIMA, exponential smoothing, basic ML
2. **Uncertainty rarely linked to business outcomes**
   Prediction intervals reported but not tied to cost/service
3. **Forecasting and replenishment studied separately**
   Few compare models under a common cost framework
4. **Vietnamese context underrepresented**

Our Contribution

A **forecast-then-optimize** framework benchmarking **30 configurations** across:

Architectures
Statistical
Tree-based
Neural (N-BEATS, TFT)

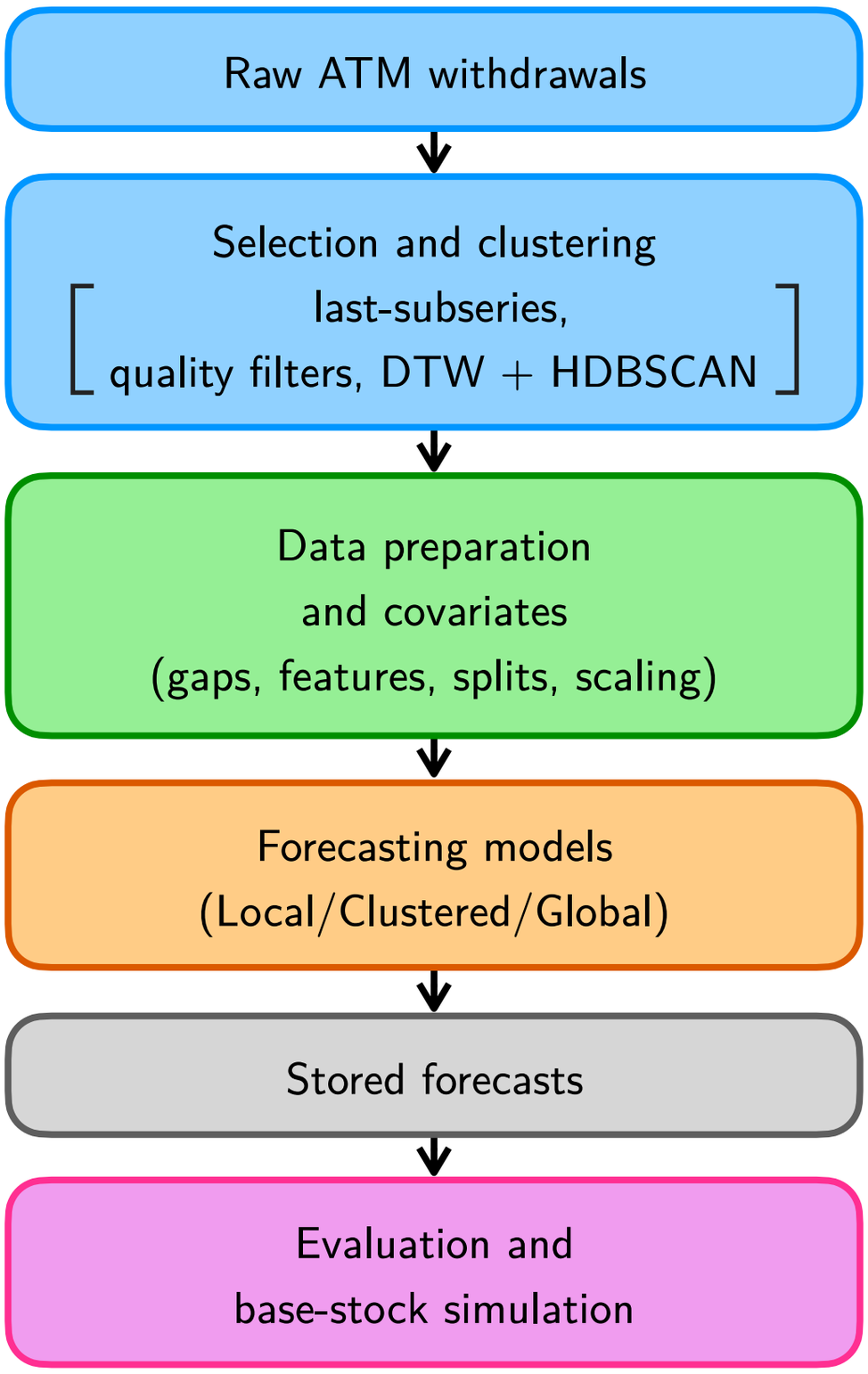
Uncertainty
Point forecasts
Quantile regression
MC Dropout

Scope
Local (per-ATM)
Clustered
Global (pooled)

Evaluated on **84 ATMs** from a Vietnamese bank over **3.7 years** using both **accuracy metrics** and **inventory cost simulation**

Next section

Framework Overview



Forecast-then-optimize: Isolates the effect of forecast quality on inventory cost

- All 30 model configurations share the same data splits, covariates, and policy
- Evaluation decoupled from training (stored forecasts)
- Simulation applies identical base-stock policy across models

Inventory Model

Minimize expected total cost across all ATMs over the evaluation horizon

$$\min \mathbb{E}\left[\sum_{a\in\mathcal{A}}\left(\underbrace{\sum_{t=1}^{H} c_h I_{a,t}}_{\text{Holding cost}} + \underbrace{\sum_{t=1}^{H} c_s U_{a,t}}_{\text{Shortage cost}} + \underbrace{c_r N_{a,\text{orders}}}_{\text{Replenishment cost}}\right)\right]$$

Order-up-to level at each review epoch:

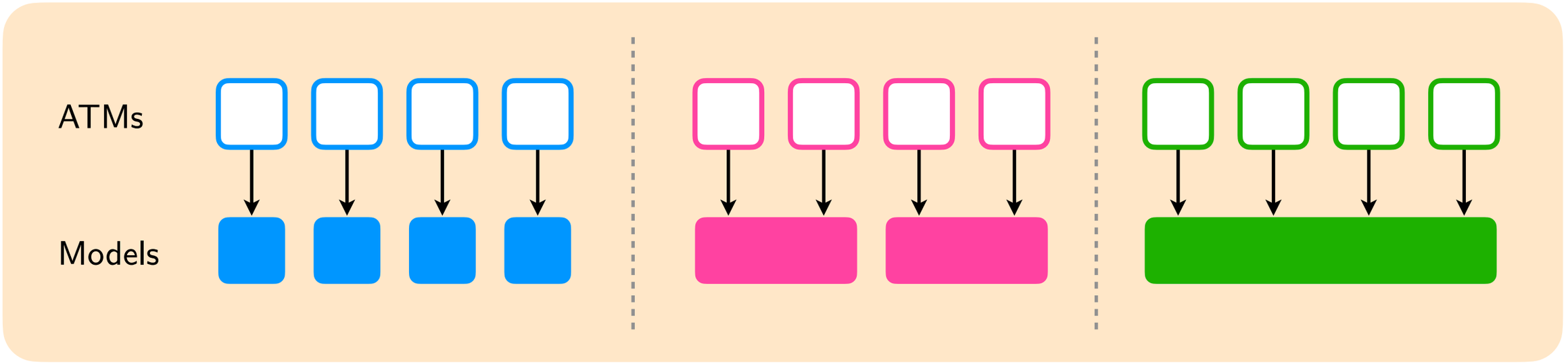
$$B_{a,t} = Q_{\rho^*}\left(\sum_{i=1}^{R+L} Y_{a,t+i}\right)$$

$c_h = 0.00026$ VND/VND-day (≈9.5% annual)

$c_s = 0.02574$ VND/VND (reputational penalty)

$c_r = 2000000$ VND/order (dispatch cost)

30 Model Configurations



Point = single trajectory (median); safety stock from scaled residuals

Quantile = direct ρ^*-quantile for base-stock decisions

Critical ratio $\rho^* = \dfrac{c_s}{c_s + c_h} \approx 0.99 \rightarrow$ cost-justified ≈99% service target

Next section

Top Models by Cost

Top tier (by total cost):

1. TFT Q G
2. N-BEATS Q G
3. N-BEATS Q C
4. TFT Q C
5. RF P L

Best: TFT Q G

- Cost: ≈16.0M VND/ATM/30d
- Fill rate: ≈99.1%, MASE: ≈0.51

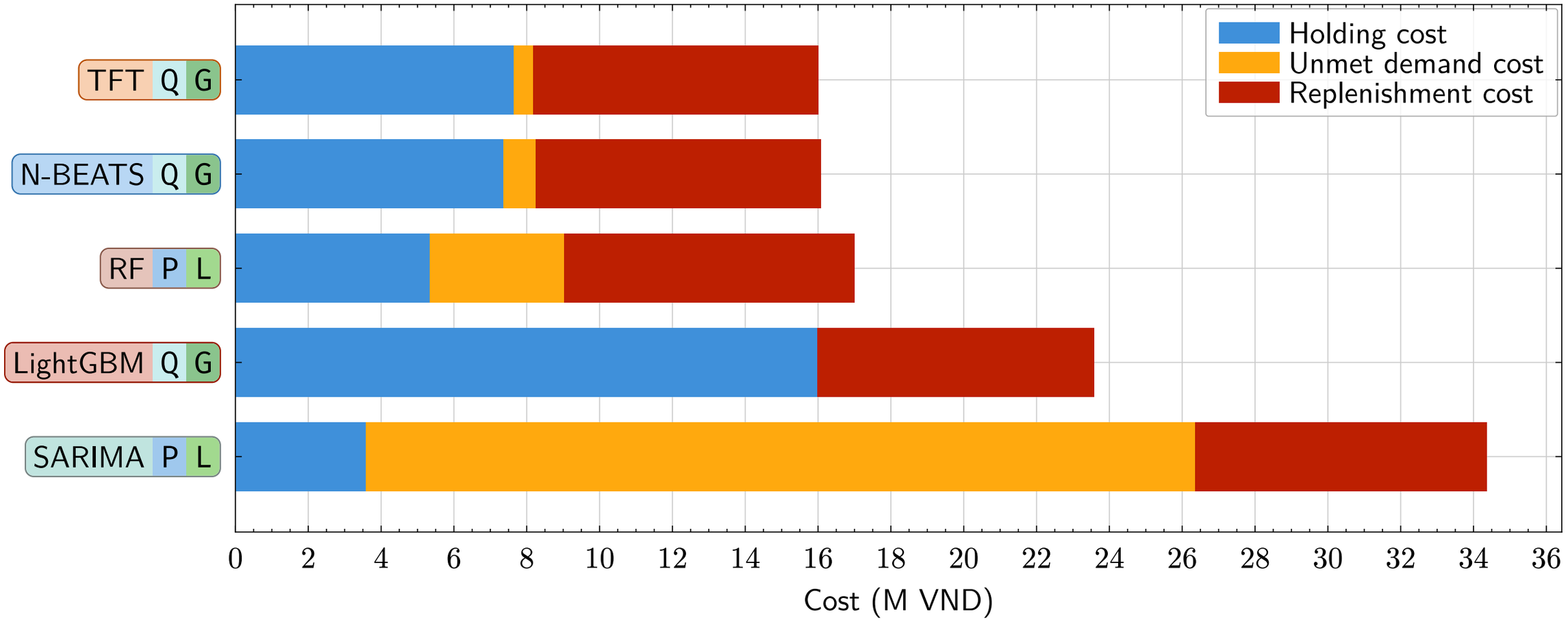
SARIMA P L **baseline:**

- Total cost:
  34.4M → 16.0M VND
  (roughly **halved**)
- Fill rate:
  86% → 99%
  (from frequent stockouts to rare)

Global quantile neural

→ substantial gains

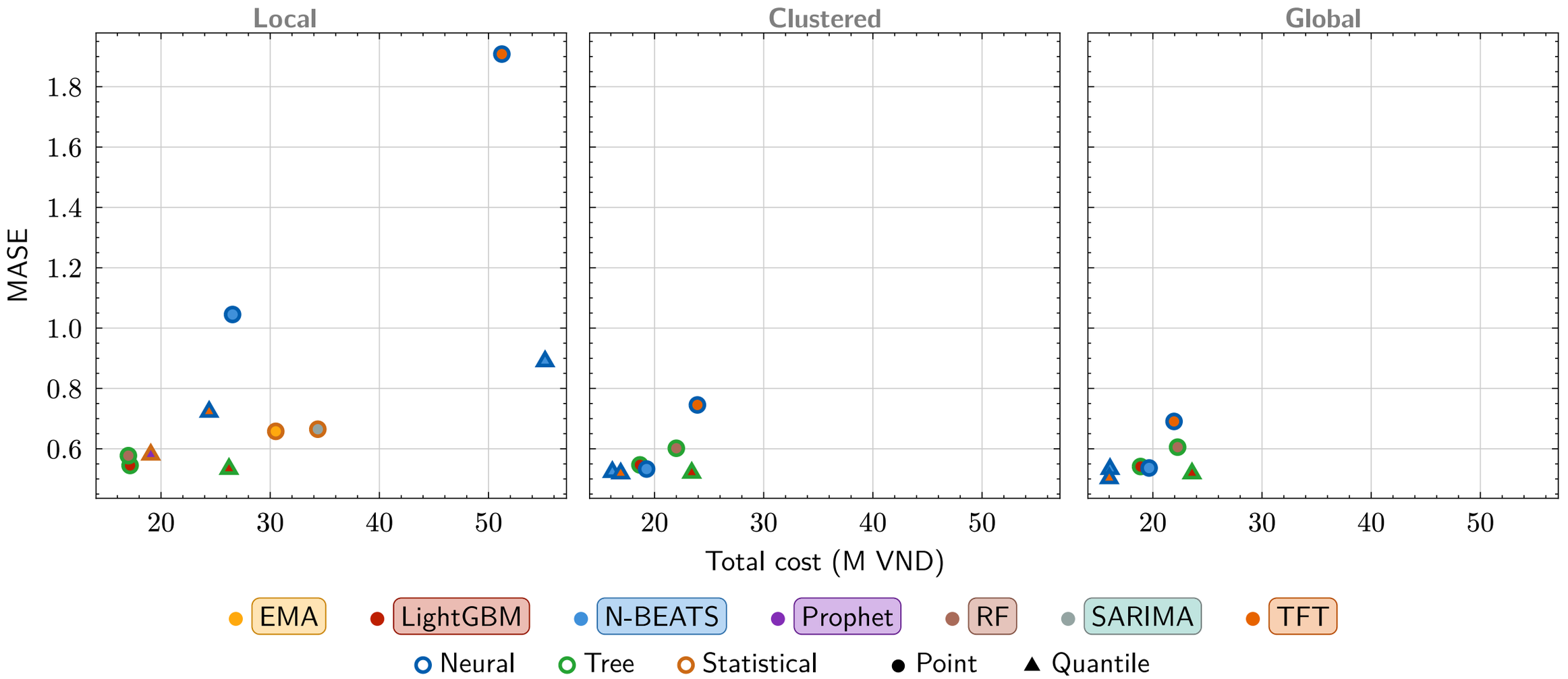
Top Models by Cost



Representative models only → one visual per family

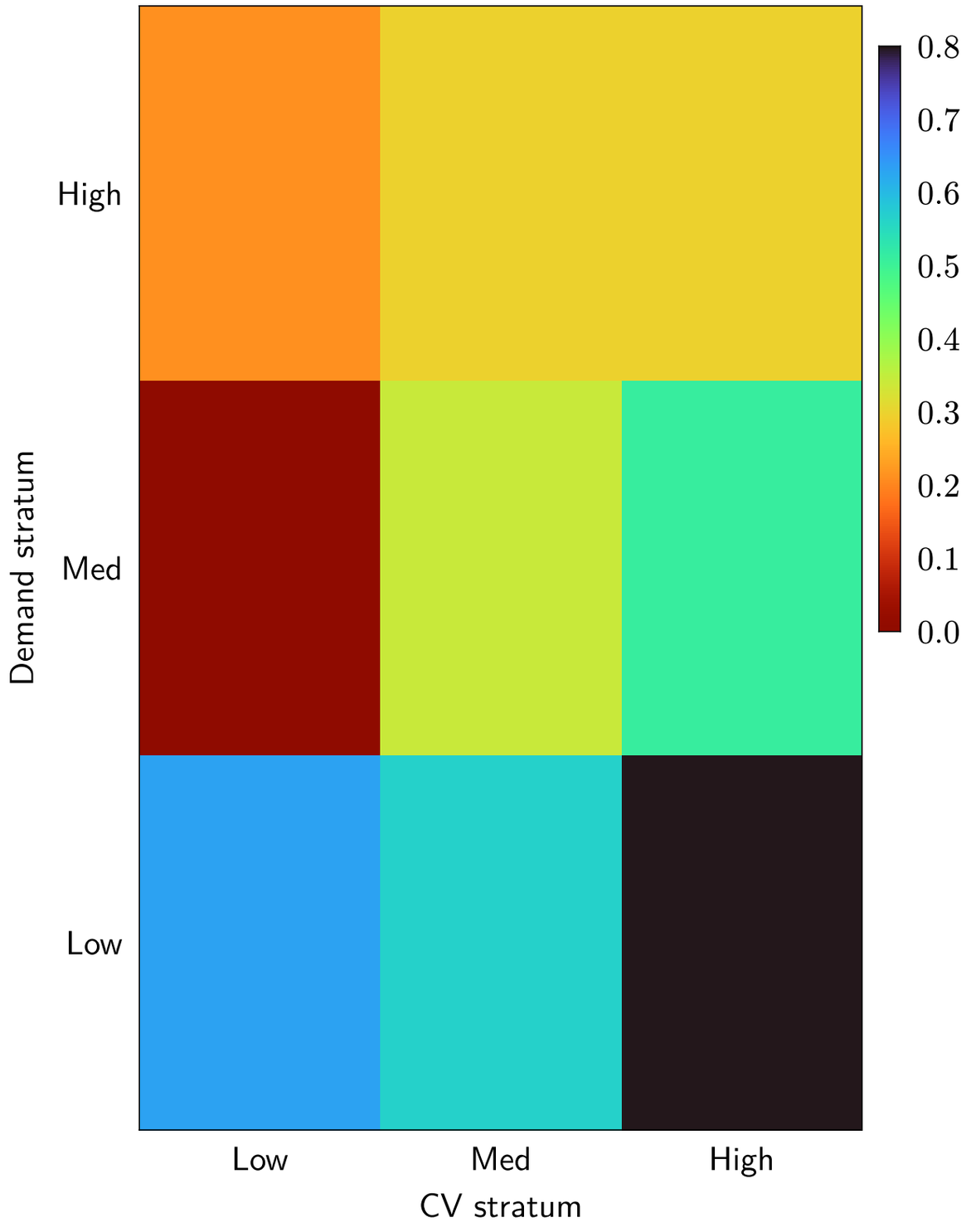
- Neural quantile models balance **holding** and **unmet demand** cost
- LightGBM Q G buys service with a **holding-cost premium**
- RF P L stays competitive, but with more stockout risk
- SARIMA P L under-orders → shortage cost dominates

Accuracy ≠ Cost



Why? Calibration measures coverage; cost depends on sharpness — the width of prediction intervals. LightGBM Q G: best CRPS, best calibration, but 47% higher cost than TFT Q G.

Heterogeneity Across ATMs



Best on average ≠ best everywhere

- TFT Q G: lowest mean total cost across fleet
- N-BEATS Q G: cheaper on ≈60% of ATMs
- Cell color = share of ATMs where TFT is cheaper than N-BEATS

Where N-BEATS tends to win

- Low-demand and high-CV ATM groups lean toward N-BEATS Q G
- TFT Q G wins only 4/28 low-demand ATMs and 8/28 high-CV ATMs
- No clean fixed rule → monitor and reassign per ATM

Next section

Key Findings

1. Global neural quantile models dominate

TFT Q G and N-BEATS Q G reduce cost by >50% vs classical baselines, achieving ≥ 99% fill rates

2. Accuracy and calibration ≠ business performance

Models with the best CRPS/calibration can still incur high cost — **sharpness** of prediction intervals matters, not just coverage

3. No universal winner across ATMs

Pronounced heterogeneity argues for adaptive, per-ATM model selection rather than one-size-fits-all deployment

Practical Implications

For ATM operators and banks:

- Replace heuristic replenishment with **data-driven**, **uncertainty-aware** policies
- Choose models using **business cost**, not forecast accuracy alone
- Use **monitoring-based reassignment** across the ATM fleet

Main limits and next steps:

- Results come from one bank and one fixed cost setting
- Next steps: sensitivity analysis, routing constraints, and decision-focused learning

Bottom line: Well-calibrated probabilistic deep models can **materially improve** ATM cash management, but model choice only makes sense when the evaluation matches the business goal

Thank you.