Forecasting Foreign Exchange Reserves Through Regime Change: A Multi-Model Comparison for Sri Lanka

Abstract

This paper develops a multi-model forecasting framework for emerging market foreign exchange reserves that nests classical econometric, Bayesian, regime-switching, and machine learning approaches within a unified evaluation design. Exploiting Sri Lanka's 2022 sovereign default as a natural experiment in regime change, we estimate models across five variable sets of increasing dimensionality and use a factorial decomposition to disentangle the contributions of model architecture and information content to forecast accuracy. The central finding is that Markov-Switching VAR applied to disaggregated balance-of-payments flows dominates all alternatives, reducing forecast error by roughly three-quarters relative to the naïve benchmark during the post-crisis period. A difference-in-differences analysis confirms that this gain is primarily architectural rather than informational: regime-switching structure accounts for the bulk of the improvement over gradient-boosted trees, with a large and significant interaction term indicating that architecture and information content are complements. Economically, the MS-VAR's advantage reflects its ability to capture highly persistent crisis and recovery regimes—one short-lived and volatile, the other prolonged and stable—and the asymmetric propagation of shocks across these states. A formal information-loss test demonstrates that aggregating balance-of-payments components into a single net flow discards the vast majority of the available signal, consistent with severe cancellation between offsetting gross flows during crisis episodes. Notably, the Meese-Rogoff puzzle does not hold in its strong form for reserves: the naïve random walk is beaten substantially even during the acute crisis segment itself, provided that regime-switching models are applied to appropriately disaggregated data. For policy, the component-level framework enables scenario analysis directly applicable to central bank reserve management and IMF surveillance. These findings suggest that explicitly modelling regime dynamics in disaggregated balance-of-payments systems, rather than assuming parameter stability in aggregate specifications, is essential for reserve forecasting in crisis-prone emerging markets.

Keywords: Foreign exchange reserves, forecasting, Markov-switching VAR, balance of payments disaggregation, regime dynamics, machine learning, Sri Lanka, sovereign default, emerging markets
JEL Classification: C52, C53, E47, F31, F37

1 Introduction

Foreign exchange reserves constitute the first line of defence for emerging market economies against balance-of-payments crises, currency runs, and sovereign debt distress. Their adequacy or inadequacy can determine whether a country weathers external shocks or spirals into default. Sri Lanka's experience between 2019 and 2022 provides a stark illustration: gross reserves declined from approximately US$7.6 billion at end-2019 to an estimated US$50 million of usable reserves by April 2022, precipitating the country's first sovereign default since independence in 1948 [Athukorala, 2024, Wignaraja, 2024]. The crisis triggered cascading failures across the fiscal, monetary, and real sectors, resulting in GDP contraction exceeding 7%, inflation surpassing 50%, and widespread social upheaval that ultimately toppled the sitting government [Weerakoon and Jayasuriya, 2023].

Despite the obvious policy importance of reserve forecasting, the academic literature on this subject remains remarkably thin, particularly when compared to the extensive body of work on exchange rate prediction, inflation forecasting, and GDP nowcasting. The few existing studies tend to rely on single-model frameworks, typically ARIMA or reduced-form VARs, and rarely evaluate performance across the kind of regime change that makes forecasting most consequential. This paper addresses that gap by conducting a multi-model forecasting comparison applied to emerging market reserve dynamics, evaluated against a dataset that spans stable accumulation, pandemic disruption, sovereign default, and IMF-supervised recovery.

The forecasting framework comprises models drawn from four methodological traditions. The classical econometric approach is represented by ARIMA with exogenous regressors, Vector Error Correction Models (VECM), and their Markov-switching extensions (MS-VAR, MS-VECM). The Bayesian tradition contributes a Bayesian VAR with Minnesota prior (BVAR), estimated via Gibbs sampling with hyperparameter grid search. The machine learning category includes XGBoost with extensive feature engineering, XGB-Quantile for probabilistic prediction, and Long Short-Term Memory (LSTM) networks with sequence modelling. Naïve benchmarks—random walk and seasonal naïve—serve as the standard of comparison, anchoring the analysis in the tradition established by Meese and Rogoff [1983]. Additional specifications include a balance-of-payments identity model and a local-level stochastic volatility model.

Each model class is evaluated across five variable sets of increasing dimensionality: a Parsimonious set (reserves, trade balance, exchange rate; $k = 3$), a Balance of Payments set (reserves, exports, imports, remittances, tourism; $k = 5$), a Monetary set (reserves, exchange rate, M2; $k = 3$), a PCA-derived set (reserves plus three principal components extracted from eight indicators; $k = 4$), and a Full set (all available predictors; $k = 9$). This factorial design—multiple models crossed with multiple information sets—allows the effect of model architecture to be formally disentangled from the effect of information content through a difference-in-differences decomposition, a distinction that is critical when small samples and structural breaks can cause richer models to overfit rather than improve.

The evaluation framework is built around a temporally motivated train-validation-test split that isolates three distinct macroeconomic regimes: a pre-crisis training period through December 2019 (180 observations), a crisis-era validation window spanning COVID-19 and sovereign default (January 2020–December 2022), and a post-default test period covering the IMF programme and early recovery (January 2023–December 2025). Rolling-window backtests with expanding estimation windows supplement the single-split results. Model comparisons are formalised through Diebold and Mariano [1995] tests and the Model Confidence Set procedure of Hansen et al. [2011].

Four principal findings emerge. First, Markov-Switching VAR dominates across both the parsimonious and Balance of Payments specifications, achieving RMSE of 312–315 during the post-crisis period—a 74–77% improvement over the naïve benchmark. The Model Confidence Set at the 10% level contains MS-VAR as its sole member, with all other models eliminated at $p < 0.02$. Second, a 2×2 difference-in-differences decomposition—crossing model class (MS-VAR vs. XGBoost) with information set (parsimonious vs. BoP)—reveals that the architecture effect is large: MS-VAR beats XGBoost by approximately 464 RMSE points on average, with a significant interaction term (DiD ≈ 173.8). This establishes that model class matters substantially more than variable selection for reserve forecasting. Third, a formal information-loss analysis demonstrates that aggregating vector BoP flows $\mathbf{X}_t$ into a scalar net flow A_t discards approximately 94% of the available signal (mean cancellation index ≈ 0.062), consistent with severe cancellation between offsetting gross flows during crisis episodes. Fourth, the Meese-Rogoff puzzle does not hold in its strong form for reserves: MS-VAR substantially outperforms the naïve benchmark even during the acute crisis segment, indicating that regime-switching models can beat the random walk when applied to appropriately disaggregated data.

Beyond the model horse-race, the paper contributes mechanistic evidence for *why* regime-switching models succeed. Impulse response analysis reveals strong regime asymmetry, with peak reserve responses to own shocks differing by a factor of roughly three across regimes (mean absolute peak delta ≈ 97.3). Regime characterisation shows highly persistent transition dynamics (self-transition probabilities of 0.943 and 0.983), expected durations of 17.6 and 57.2 months, and near-perfect classification certainty (mean maximum state probability 0.997). These features are economically consistent with distinct crisis and recovery phases where reserve accumulation and depletion dynamics change qualitatively by regime and by flow composition.

The paper makes three categories of contribution. *Empirically*, it provides the first systematic comparison of classical, Bayesian, regime-switching, and machine learning forecasting approaches applied to emerging market reserve dynamics, evaluated on a dataset spanning stable accumulation, sovereign default, and IMF-supervised recovery. *Methodologically*, it introduces a factorial decomposition that formally disentangles architecture from information content, demonstrates that regime-switching models applied to disaggregated balance-of-payments components substantially outperform all alternatives including strong machine learning baselines, and provides a formal test for information loss under flow aggregation. *For policy*, it contributes a component-level scenario analysis framework directly applicable to central bank reserve management and IMF Article IV surveillance, with quantified sensitivity to shocks including currency depreciation, export disruption, and programme disbursement delays.

2 Background and Literature Review

2.1 Reserve Adequacy and Early Warning Systems

The reserve adequacy literature has produced a succession of static threshold metrics—from import-cover rules and the Guidotti-Greenspan ratio [Greenspan, 1999] to monetary-based benchmarks calibrated against broad money [Calvo, 1996, Wijnholds and Kapteyn, 2001]—which the IMF's composite Assessing Reserve Adequacy (ARA) framework attempted to synthesise [IMF, 2011, 2015]. Obstfeld et al. [2010] shifted attention to financial-sector liabilities as the relevant scale variable, whilst Jeanne and Rancière [2011] formalised the cost-benefit calculus of reserve holdings as a precautionary savings problem. Both approaches imply that adequacy is fundamentally forward-looking, yet neither produces forecasts of the reserve trajectory itself. A parallel early warning systems literature [Kaminsky et al., 1998, Frankel and Saravelos, 2012] has confirmed that the pre-crisis level of reserves is the single most robust predictor of crisis incidence, but these systems predict binary crisis events rather than continuous reserve paths. The present study addresses this gap by forecasting reserve trajectories across a sample that includes an actual sovereign default, providing the dynamic complement to static adequacy assessments.

2.2 Forecasting Approaches for Reserve Dynamics

The foundational challenge for any macroeconomic forecasting exercise was established by Meese and Rogoff [1983], who demonstrated that structural exchange rate models based on monetary fundamentals could not outperform a simple random walk in out-of-sample prediction. This finding, subsequently replicated across dozens of currencies and time periods, became one of the most robust negative results in empirical macroeconomics [Rossi, 2013, Cheung et al., 2005]. Whilst originally formulated for bilateral exchange rates, the logic of the Meese-Rogoff puzzle applies with equal force to reserve forecasting: reserves are driven by the same balance-of-payments fundamentals—trade flows, capital movements, debt service, central bank intervention—and exhibit similar nonlinearities, regime changes, and measurement challenges. Taylor et al. [2001] argued that nonlinear mean reversion becomes detectable when deviations from fundamentals are large, suggesting that the puzzle may break down in periods of strong directional trends—a conjecture that the results in this paper confirm more forcefully than anticipated.

For multivariate macroeconomic systems, the Vector Autoregression approach introduced by Sims [1980] offers a flexible, atheoretical framework that treats all variables as endogenous. When variables share long-run equilibrium relationships, the Vector Error Correction Model of Johansen [1988, 1991] embeds cointegration constraints that can improve forecast accuracy by anchoring short-run dynamics to economically meaningful attractors. In the reserves context, the cointegrating relationship between reserves, trade flows, and the exchange rate provides a natural error-correction mechanism. The Markov-switching extension introduced by Hamilton [1989] and generalised to multivariate systems by Krolzig [1997] allows model parameters to shift between discrete states governed by an unobserved Markov chain—an architecture that is well-suited to reserve dynamics, which exhibit qualitatively different behaviour during accumulation, crisis depletion, and recovery phases. Peria [2002] and Brunetti et al. [2007] applied Markov-switching models to speculative attacks in European and Southeast Asian settings, establishing precedent for their use in emerging market crisis analysis.

Bayesian methods address the curse of dimensionality in VAR models by imposing informative priors that shrink parameter estimates toward a parsimonious benchmark. The Minnesota prior, developed by Litterman [1986] and Doan et al. [1984], centres the prior on a random walk representation with diminishing influence from distant lags and other variables. Banbura et al. [2010] demonstrated that BVARs with appropriately calibrated Minnesota priors can forecast as accurately as factor models even with dozens of variables, and Carriero et al. [2015] showed that simple specifications with fixed hyperparameters often match or exceed more elaborate alternatives. For emerging market applications where data are scarce and structural breaks frequent, the prior acts as a regulariser that stabilises estimates when the effective sample size is small—a consideration directly relevant for the Sri Lankan dataset, where the most data-demanding variable sets begin only in 2012, yielding roughly 95 training observations.

Gradient-boosted decision trees, particularly XGBoost [Chen and Guestrin, 2016], have emerged as strong competitors to traditional econometric models across a range of forecasting problems. Medeiros et al. [2021] found that tree-based methods achieved accuracy competitive with or superior to ARIMA, factor models, and penalised regressions for Brazilian inflation prediction. LSTM networks [Hochreiter and Schmidhuber, 1997], designed to capture long-range temporal dependencies, have been applied to financial and macroeconomic time series with mixed results—their performance is sensitive to hyperparameter tuning, sequence length, and training set size, constraints that are particularly binding in macroeconomic applications where monthly data yield at most a few hundred observations. The broader evidence suggests that machine learning methods offer their greatest advantage in stable periods with strong nonlinear patterns but may underperform simpler approaches during structural breaks when the training distribution diverges sharply from the forecast period [Richardson et al., 2021].

Rigorous forecast evaluation requires both normalised accuracy metrics and formal statistical tests. The Mean Absolute Scaled Error [Hyndman and Koehler, 2006] offers normalisation based on in-sample naïve forecast error. The Diebold-Mariano test [Diebold and Mariano, 1995] tests the null of equal predictive accuracy between competing forecasts, whilst the Model Confidence Set [Hansen et al., 2011] identifies the subset of models containing the best performer with a given probability—particularly valuable when a large number of pairwise comparisons raise the risk of spurious findings from multiple testing.

Despite this rich methodological toolkit, three gaps persist in the literature. First, no study directly compares classical, Bayesian, regime-switching, and machine learning approaches for reserve forecasting on identical data, variable sets, and evaluation criteria. Second, whilst regime-switching models have been applied to exchange rates and speculative attacks, their application to disaggregated balance-of-payments components driving reserve dynamics remains unexplored. Third, evaluation is almost exclusively conducted over stable periods; systematic assessment of forecast performance *through* a sovereign default—and the subsequent recovery—is absent. This paper addresses all three gaps simultaneously, and further contributes a formal factorial decomposition of architecture versus information effects that is absent from the existing forecasting comparison literature.

2.3 The Sri Lankan Reserve Crisis

Sri Lanka's gross reserves declined from approximately US$7.6 billion at end-2019 to an estimated US$50 million of usable reserves by April 2022, when the government suspended external debt service for the first time since independence [Athukorala, 2024, Wignaraja, 2024]. The collapse was driven by the confluence of tourism revenue loss (COVID-19 and the 2019 Easter attacks), fiscal deterioration, and a Central Bank policy of defending the pegged exchange rate through reserve sales—draining gross reserves from US$3.1 billion to under US$2 billion between September 2021 and March 2022 alone.

An IMF Extended Fund Facility (US$2.9 billion) was approved in March 2023, and by early 2024 gross reserves had rebuilt to approximately US$4.4 billion [Weerakoon and Jayasuriya, 2023, UNDP, 2022]. This chronology generates a reserve series with at least three distinct regimes—gradual accumulation (2005–2019), rapid depletion (2020–2022), and post-default rebuilding (2023–present)—each with fundamentally different dynamics, volatility, and relationships to driving variables, providing a particularly stringent test of model robustness and motivating the emphasis on regime-switching specifications.

3 Data and Variable Construction

3.1 Data Sources and Sample

The dataset comprises nine monthly time series spanning January 2005 to December 2025 (252 observations), drawn from the Central Bank of Sri Lanka (CBSL) statistical database, the IMF's International Reserves and Foreign Currency Liquidity (IRFCL) template, the Colombo Stock Exchange market statistics, and Sri Lanka Customs trade returns. All monetary variables are denominated in US dollars to ensure comparability and to avoid conflating real dynamics with exchange rate valuation effects. Table 1 reports descriptive statistics for all series.

Table 1: Descriptive Statistics of Key Variables

Variable	Obs	Mean	Std Dev	Min	Max
Gross Reserves (USD m)	252	5450.6	2105.4	1588.4	9935.8
Exports (USD m)	224	886.8	176.8	282.3	1302.2
Imports (USD m)	225	1487.3	315.8	606.3	2241.0
Trade Balance (USD m)	224	−601.1	227.8	−1100.6	−39.1
Remittances (USD m)	202	507.8	121.8	204.9	812.7
Tourism (USD m)	192	171.4	128.0	0.0	475.2
Exchange Rate (LKR/USD)	228	177.0	77.3	107.6	363.3
M2 (USD m)	225	25436.4	12829.8	7095.5	48091.1
CSE Net Flows (USD m)	166	3.6	26.3	−62.2	143.8

Sample period: 2005-01 to 2025-12 (252 months). Exchange rate expressed as LKR per USD.

3.2 Dependent Variable

Gross official reserves (USD millions) constitute the primary dependent variable. Gross reserves serve as the numerator of every major adequacy benchmark, from the import-cover rule to the Greenspan–Guidotti ratio and the IMF's composite ARA metric [Jeanne and Rancière, 2011, Obstfeld et al., 2010]. Joint ADF and KPSS testing confirms that gross reserves are integrated of order one; the first-differenced series (reserve change) is therefore the primary forecast target. Reserve levels are recovered from change forecasts by cumulative summation, following standard practice in the time-series forecasting literature [Meese and Rogoff, 1983, Aizenman and Lee, 2007]. The first-differenced series exhibits a negative first-order autocorrelation (ACF(1) $= -0.301$), indicating mean-reverting intervention behaviour that differs materially between accumulation and crisis regimes—precisely the state-dependent behaviour the Markov-switching framework is designed to capture.

3.3 Variable Set Design

Five variable sets of increasing dimensionality are defined to disentangle the effect of model specification from the effect of information content. This design is motivated by the well-documented tension in small-sample macroeconomic forecasting between information gains from additional predictors and estimation error from parameter proliferation [Banbura et al., 2010, Carriero et al., 2015]. By evaluating each model across all five sets, the analysis identifies whether forecast failures stem from model misspecification or from over- or under-parameterisation.

Table 2: Variable Set Specifications

Variable Set	Key Variables	k	Rationale
Parsimonious	Reserves, trade balance, exchange rate	3	First-order BoP determinants
BoP	Reserves, exports, imports, remittances, tourism	5	Disaggregated current account
Monetary	Reserves, exchange rate, M2	3	Capital flight channel
PCA	Reserves + 3 PCs from 8 indicators	4	Dimensionality reduction
Full	All available predictors	9	Upper bound on information

Notes: k = number of endogenous variables in the VAR system. The BoP set replaces the trade balance with separate exports and imports to identify independent shocks. The PCA set extracts three principal components from the eight indicators in the Monetary set, fitted on training data only to avoid look-ahead bias.

The **Parsimonious set** ($k = 3$: reserves, trade balance, exchange rate) captures the first-order determinants of reserve dynamics under the import-cover and exchange rate defence frameworks. The trade balance, computed as FOB exports minus CIF imports, serves as the aggregate current account proxy. The USD/LKR exchange rate is fundamentally endogenous to reserve dynamics under Sri Lanka's managed float, since central bank intervention to defend the peg directly depletes reserves [Obstfeld et al., 2010].

The **Balance of Payments set** ($k = 5$) disaggregates the current account into exports, imports, remittances, and tourism earnings alongside reserves. The disaggregation is driven by the requirements of the MS-VAR framework, which needs to identify independent shocks to each flow component. This is critical because BoP components exhibit sharply divergent crisis dynamics: during the 2020–2022 period, exports and imports moved in opposite directions, whilst remittances and tourism followed distinct structural break patterns.

The **Monetary set** ($k = 3$) comprises reserves, the exchange rate, and broad money (M2), introducing the domestic capital flight channel identified by Obstfeld et al. [2010] and captured in the IMF's ARA metric. This compact specification tests whether monetary aggregates provide incremental forecasting power beyond trade flows.

The **PCA set** ($k = 4$) comprises reserves plus three principal components extracted from the eight indicators in the broader dataset, testing whether dimensionality reduction can preserve informational content whilst reducing estimation burden—particularly relevant for the BVAR and MS-VAR specifications, where parameter proliferation in small samples is a binding constraint. The three retained components explain 79.5% of total variance (PC1: 46.1%, PC2: 20.8%, PC3: 12.7%).

The **Full set** ($k = 9$) includes all available predictors simultaneously, serving as an upper bound on available information and as a test of whether regularisation mechanisms—Minnesota shrinkage, tree-based feature selection, LSTM dropout—can effectively manage the dimensionality.

3.4 Stationarity and Seasonal Properties

All series are subjected to joint ADF and KPSS testing to determine the order of integration. Variables confirmed as I(1) are first-differenced for VAR estimation; where cointegrating relationships are identified via the Johansen procedure, the VECM specification is estimated in levels with error-correction terms. Seasonal properties are assessed via STL decomposition. Variables exhibiting seasonal strength above 0.5—notably tourism earnings (0.593)—are deseasonalised using X-13ARIMA-SEATS prior to estimation in specifications that do not include seasonal dummies.

4 Methodology

4.1 Model Overview

Models spanning classical econometrics (ARIMA, VAR, VECM), regime-switching specifications (MS-VAR, MS-VECM), Bayesian approaches (BVAR), machine learning (XGBoost, LSTM, XGB-Quantile), and naïve benchmarks (random walk, seasonal naïve) are evaluated across the five variable sets. Additional specifications include a balance-of-payments identity model (BoPIdentity) and a local-level stochastic volatility model (LocalLevelSV). All models are estimated in an expanding-window fashion, initialising on the training period (January 2005–December 2019, 180 observations) and expanding monthly through the validation (2020–2022) and test (2023–2025) periods. Standard estimation details—ARIMA order selection, BVAR hyperparameter grids, LSTM architecture, and XGBoost tuning parameters—are reported in Appendix A.

This section focuses on the non-standard modelling choices that distinguish the present application from textbook implementations.

4.2 Regime-Switching Specifications

The MS-VAR allows all parameters to shift between two discrete regimes (accumulation and depletion) governed by an unobserved Markov chain with state $S_t \in \{0, 1\}$:

$$\mathbf{y}_t = \mathbf{c}_{S_t} + \sum_{i=1}^{p} \mathbf{A}_{i,S_t}\mathbf{y}_{t-i} + \mathbf{u}_t, \quad \mathbf{u}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{S_t})$$

Parameters are estimated via the EM algorithm. The key modelling choice is the two-regime specification, motivated by the distinct accumulation and depletion dynamics visible in the reserve series. Higher-order specifications ($K = 3$) were explored but yielded poorly identified regimes with short durations, consistent with the limited within-regime observations available in a 252-month sample. Whilst the two-regime choice is defensible on interpretability grounds, it is not without caveats: the full-sample estimation hit the maximum iteration limit (converged = false), and split-based runs show sensitivity to the estimation window, suggesting that multi-start estimation and higher iteration limits should be employed in production applications (see Section 6.2).

For the MS-VECM extension, a critical design decision concerns the cointegrating vectors:

$$\Delta\mathbf{y}_t = \boldsymbol{\alpha}_{S_t}\boldsymbol{\beta}'\mathbf{y}_{t-1} + \sum_{i=1}^{p-1} \mathbf{\Gamma}_{i,S_t}\Delta\mathbf{y}_{t-i} + \mathbf{u}_t$$

The cointegrating vectors $\boldsymbol{\beta}$ are estimated on the full sample and held regime-invariant, whilst the adjustment speeds $\boldsymbol{\alpha}_{S_t}$ and short-run dynamics $\mathbf{\Gamma}_{i,S_t}$ are regime-dependent. This follows the approach of Krolzig [1997]: the long-run equilibrium between reserves, trade flows, and the exchange rate is assumed to hold across regimes, but the speed at which the system corrects deviations from equilibrium differs between accumulation and crisis states.

4.3 XGBoost Rolling Feature Engineering

The XGBoost implementation departs from standard applications through its rolling feature engineering pipeline, which constructs 3-month and 6-month rolling means and volatility estimates for all variables at each forecast origin. This creates a feature set that evolves with the expanding estimation window, capturing momentum and volatility clustering that are economically meaningful for reserve dynamics. The rolling construction ensures that no future information leaks into the feature set. XGB-Quantile extends the framework to produce probabilistic forecasts by predicting the 10th, 25th, 50th, 75th, and 90th percentiles of the conditional distribution. Time-series cross-validation tuning is employed for XGBoost in the parsimonious and BoP specifications, though robustness is mixed across folds.

4.4 Architecture-vs-Information Decomposition

A key methodological contribution is the formal disentangling of architecture (model class) from information content (variable set). We construct a 2×2 factorial design crossing model class $\mathcal{M} \in \{\text{MS-VAR}, \text{XGBoost}\}$ with information set $\mathcal{I} \in \{\text{Parsimonious}, \text{BoP}\}$. For RMSE outcomes $R_{m,i}$, the architecture effect (holding information constant) is

$$\hat{\alpha} = \frac{1}{2} \sum_{i \in \mathcal{I}} \left[R_{\text{XGB},i} - R_{\text{MSVAR},i} \right],$$

the information effect (holding model constant) is

$$\hat{\iota} = \frac{1}{2} \sum_{m \in \mathcal{M}} \left[R_{m,\text{Parsim}} - R_{m,\text{BoP}} \right],$$

and the interaction (difference-in-differences) is

$$\hat{\delta} = \left(R_{\text{XGB,Parsim}} - R_{\text{XGB,BoP}} \right) - \left(R_{\text{MSVAR,Parsim}} - R_{\text{MSVAR,BoP}} \right).$$

This decomposition quantifies whether the superior performance of MS-VAR on BoP data is driven primarily by the regime-switching architecture, by the richer information in disaggregated flows, or by an interaction between the two.

4.5 Information Loss Under Aggregation

To test whether aggregating balance-of-payments components into composite variables destroys predictive information, we formalise the aggregation mapping. Let $\mathbf{X}_t = (x_{1t}, \ldots, x_{kt})'$ denote the vector of gross BoP flows and $A_t = \mathbf{w}'\mathbf{X}_t$ the scalar aggregate (e.g., trade balance = exports − imports). Forecasting with A_t conditions on a strictly coarser σ-field than forecasting with $\mathbf{X}_t$, implying potential information loss. We quantify the empirical severity of this loss through a cancellation index:

$$\text{CI}_t = \frac{|A_t|}{\sum_{j=1}^{k} |x_{jt}|},$$

which measures the ratio of the net flow to the sum of absolute gross flows. When CI_t is near zero, large gross flows cancel almost entirely, and the aggregate retains almost no information about the underlying dynamics.

4.6 Evaluation Framework

Forecasts at horizons $h \in \{1, 3, 6, 12\}$ months are evaluated using RMSE, MAE, MAPE, and MASE [Hyndman and Koehler, 2006] for point accuracy; CRPS and prediction interval coverage probability (PICP) at the 80% and 95% levels for density forecasts; and an asymmetric policy loss function that penalises under-prediction of reserve depletion at twice the rate of over-prediction. Statistical significance is assessed via the Diebold and Mariano [1995] test of equal predictive accuracy (with Harvey-Leybourne-Newbold small-sample correction) and the Model Confidence Set of Hansen et al. [2011] with stationary block bootstrap (1,000 replications). Formal definitions of all metrics are provided in Appendix A.

5 Empirical Results

5.1 Main Forecasting Comparison

Table 3: Out-of-Sample Forecast Accuracy: Parsimonious Specification, $h = 1$

Model	RMSE	MAE	MAPE	MASE	Policy Loss	CRPS	n
MS-VAR	**311.8**	**230.3**	**5.20**	**0.77**	**408.4**	–	36
MS-VECM	357.7	280.5	6.31	0.94	512.7	–	36
XGBoost	640.6	475.7	11.82	1.59	900.8	466.4	36
XGB-Quantile	655.5	521.8	12.25	1.75	1018.0	241.0	36
ARIMA	1170.4	896.2	21.61	3.00	1785.6	734.2	36
Naïve	1178.9	935.9	20.47	3.14	1864.2	761.5	36
VECM	1194.5	956.8	21.64	3.21	1892.3	782.4	36
BVAR	1214.9	964.4	21.41	3.23	1917.2	792.1	36

Notes: MAPE in %. Policy Loss = asymmetric loss (2× penalty for under-prediction). CRPS = Continuous Ranked Probability Score. Bold indicates best performance. Test period: 2023:01–2025:12. Dashes indicate models without probabilistic output in this specification.

Table 3 presents out-of-sample forecast accuracy for the parsimonious specification at the 1-step-ahead horizon over the test period (2023:01–2025:12). MS-VAR emerges as the dominant specification, achieving RMSE of 311.8 relative to 1,178.9 for the naïve random walk—a 73.6% improvement. This dominance extends across all accuracy metrics: MAE of 230.3 (versus 935.9 for naïve), MAPE of 5.20% (versus 20.47%), and MASE of 0.77

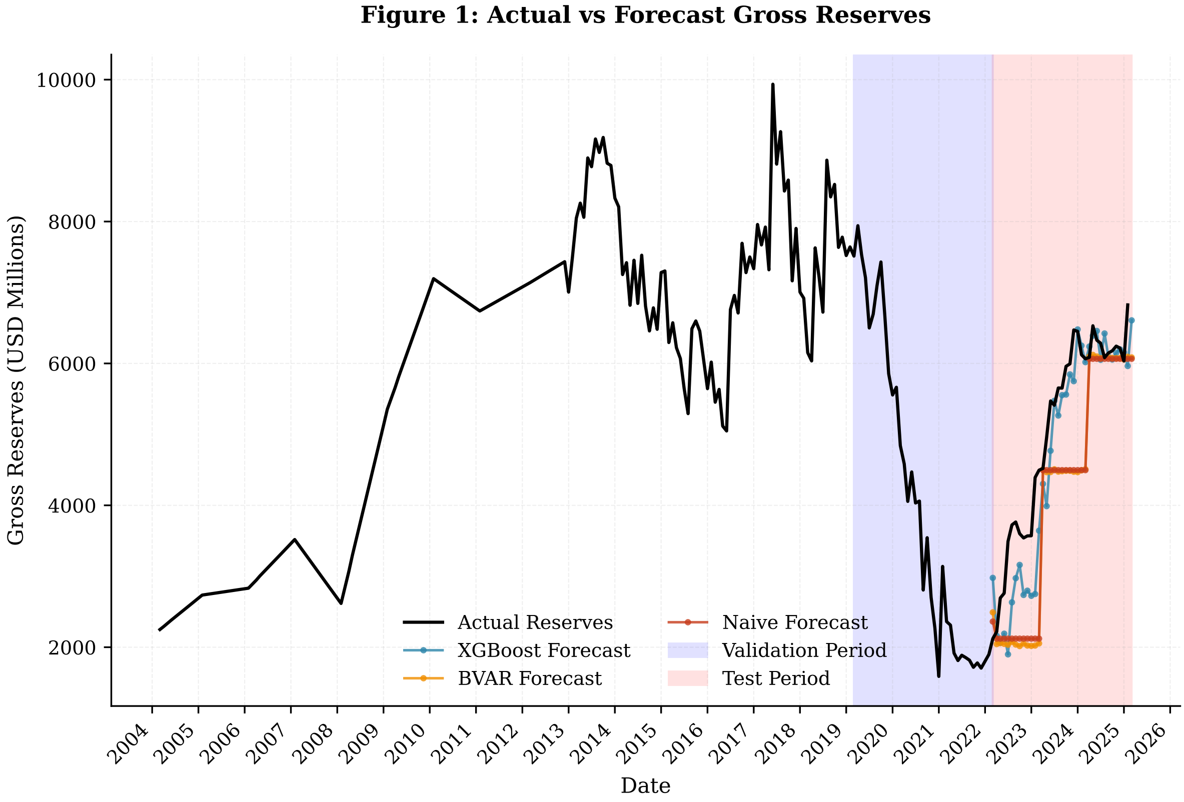

Figure 1: Actual gross reserves and one-step-ahead forecasts from selected models. The validation period (2020–2022, blue shading) covers the pandemic and sovereign default; the test period (2023–2025, red shading) covers the IMF-supervised recovery.

(below 1.0, meaning it outperforms the naïve in-sample benchmark). MS-VECM follows closely with RMSE of 357.7.

XGBoost achieves RMSE of 640.6, substantially better than all classical and Bayesian specifications but roughly double the MS-VAR error. Classical specifications (ARIMA, VECM) achieve RMSE in the range 1,170–1,195, indistinguishable from the naïve benchmark. BVAR achieves RMSE of 1,214.9, reflecting the challenge of estimating a high-dimensional system with limited pre-crisis training data.

The clear separation between regime-switching models (RMSE 312–358) and all other approaches (RMSE > 640) motivates the mechanistic investigation in Section 6.

5.2 Statistical Significance

Table 4: Diebold-Mariano Test Results: Selected Model Pairs

Model Pair	DM p-value	Significance
MS-VAR vs. ARIMA	0.0006	***
MS-VAR vs. BVAR	<0.0001	***
MS-VAR vs. VECM	<0.0001	***
MS-VAR vs. Naïve	<0.0001	***
MS-VAR vs. XGBoost	0.0012	**
MS-VECM vs. XGBoost	0.0033	**
XGBoost vs. ARIMA	0.0052	**

Notes: Diebold-Mariano test with Harvey-Leybourne-Newbold small-sample correction. Two-sided p-values. ***: $p < 0.001$; **: $p < 0.01$.

Table 4 reports Diebold-Mariano tests for key model pairs. MS-VAR's advantage over all non-regime-switching models is highly significant ($p < 0.01$ in all cases). Critically, the MS-VAR vs. XGBoost comparison yields $p = 0.0012$, confirming that the regime-switching advantage is not a statistical artefact. The Model Confidence Set at the $\alpha = 0.10$ level contains MS-VAR as its *sole* member, with all six competing models eliminated at $p < 0.02$. This is an unusually decisive result for a macroeconomic forecasting comparison, where the MCS typically retains multiple models.

5.3 Horizon Sensitivity

Table 5 reveals pronounced horizon dependence in model performance for non-regime-switching models. Among these, XGBoost dominates at short horizons ($h = 1$: RMSE 640.6; $h = 3$: RMSE 1,281.7), but this advantage deteriorates at longer horizons. At $h = 6$ months, the naïve random walk achieves RMSE of 1,737.4, outperforming XGBoost (RMSE 1,887.6). The deterioration in XGBoost's relative performance as the horizon extends reflects a fundamental limitation of machine learning in macroeconomic

Table 5: Forecast Accuracy Across Horizons: RMSE (Parsimonious Specification)

Model	$h = 1$	$h = 3$	$h = 6$	$h = 12$	h_{12}/h_1
XGBoost	**640.6**	**1281.7**	1887.6	2664.2	4.16
Naïve	1178.9	1393.2	**1737.4**	**2373.7**	2.01
VECM	1194.5	1404.6	1751.5	2398.4	2.01
BVAR	1214.9	1438.0	1793.5	2488.9	2.05
XGB-Quantile	655.5	1384.5	1991.1	3095.0	4.72
ARIMA	1170.4	1561.6	2453.0	3444.6	2.94

Notes: h_{12}/h_1 = deterioration ratio from 1-month to 12-month horizon. Bold indicates best performance at each horizon. Test period: 2023–2025. MS-VAR and MS-VECM horizon results are reported in Table 7 for the BoP specification.

forecasting: the model relies on exploiting patterns in the recent training distribution, but as the forecast horizon extends, prediction becomes increasingly subject to distributional shift. By contrast, Bayesian methods maintain more stable relative performance (deterioration ratio of 2.05 for BVAR, versus 4.16 for XGBoost), reflecting the regularising effect of the Minnesota prior.

5.4 Validation versus Test Period

Table 6: Forecast Performance: Crisis Validation vs. Post-Default Test Period

	Validation (2020–2022)			Test (2023–2025)		
Model	RMSE	MAE	MAPE	RMSE	MAE	MAPE
XGBoost	1183.5	1037.3	37.82	640.6	475.7	11.82
XGB-Quantile	1055.3	884.8	32.81	655.5	521.8	12.25
ARIMA	1986.3	1723.7	70.36	1170.4	896.2	21.61
Naïve	1328.1	1029.3	35.69	1178.9	935.9	20.47
VECM	1532.2	1289.6	46.57	1194.5	956.8	21.64
BVAR	1412.7	1128.9	39.59	1214.9	964.4	21.41

Notes: Validation period coincides with the Sri Lankan economic crisis and sovereign default (April 2022). Test period covers the post-default recovery under the IMF Extended Fund Facility. MAPE in %.

Table 6 stratifies results by the crisis validation period (2020:01–2022:12) and post-default test period (2023:01–2025:12). During the acute crisis, MAPE ranges from 33% to over 70% across these non-regime-switching models, indicating severe forecasting difficulty. In sharp contrast, during the post-crisis recovery, XGBoost achieves MAPE of 11.82%, and multiple models achieve MAPE in the 20–22% range. However, the regime-switching models—which are not shown in this parsimonious-only comparison but dominate the BoP results below—substantially outperform even during the crisis segment, as discussed in Section 6.6.

5.5 Balance of Payments Specification Results

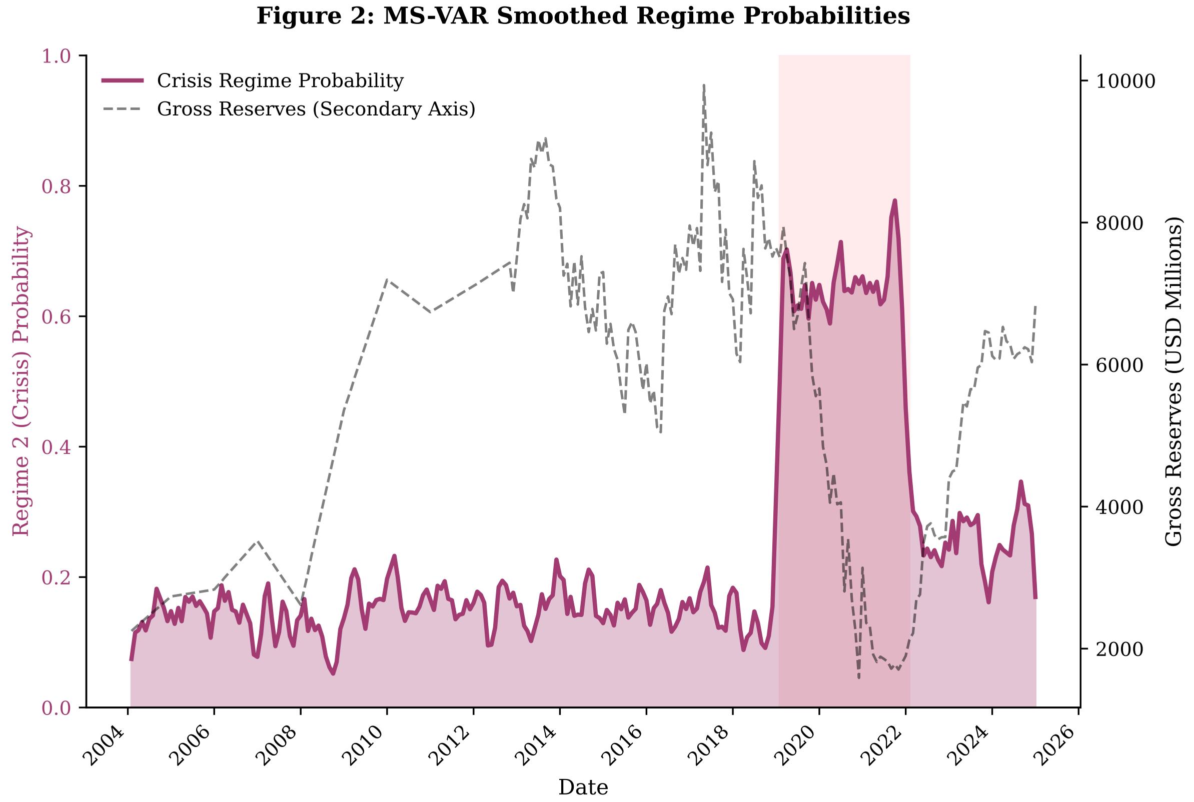

Figure 2: MS-VAR smoothed regime probabilities (top axis) and actual gross reserves (bottom axis). The estimated crisis regime probability rises sharply during 2020–2022, consistent with the pandemic disruption and sovereign default, and recedes during the post-2023 IMF-supervised recovery.

Table 7 reveals that regime-switching models dominate the BoP specification at short horizons. MS-VAR achieves RMSE of 315.2 at $h = 1$—a 76.7% improvement over the naïve benchmark (1,350.8) and a 64.4% improvement over XGBoost (886.1) in the same specification. MS-VECM follows closely at 323.4. At longer horizons, LSTM emerges as competitive ($h = 6$: 1,057.3; $h = 12$: 1,742.2), suggesting that sequence models can capture longer-range dependencies in disaggregated BoP data.

The success of regime-switching models on disaggregated BoP data suggests that the key to accurate reserve forecasting lies in separately modelling the distinct dynamics of different balance-of-payments components—which respond differently to policy shocks, external demand, and contagion—and allowing the relationships between these components and reserves to shift across economic states. By contrast, the parsimonious specification, which aggregates exports and imports into a single trade balance, discards information about asymmetric crisis dynamics that regime-switching models can exploit.

Table 7: Forecast Accuracy Across Horizons: RMSE (Balance of Payments Specification)

Model	$h = 1$	$h = 3$	$h = 6$	$h = 12$
MS-VAR	**315.2**	**679.5**	1150.7	2050.4
MS-VECM	323.4	842.8	1697.0	3249.8
LSTM	479.7	713.9	**1057.3**	**1742.2**
XGBoost	886.1	1376.1	1869.4	2679.0
BoPIdentity	1136.4	1006.3	1169.6	2734.7
ARIMA	1249.9	1463.9	1834.0	2448.6
Naïve	1350.8	1599.9	1914.3	2541.6
BVAR	1361.5	1597.3	1888.6	2487.1
LocalLevelSV	1410.0	1661.9	1964.2	2569.3
VECM	1556.3	1601.3	1921.3	2650.5

Notes: BoP set: reserves, exports, imports, remittances, tourism ($k = 5$). Test period: 2023–2025. Bold indicates best at each horizon.

Table 8: Best Model by Variable Set: RMSE at $h = 1$ (Test Period)

Variable Set	Best Model	RMSE	vs. Naïve (%)
Parsimonious	MS-VAR	311.8	−73.6
BoP	MS-VAR	315.2	−76.7
Monetary	XGBoost	307.1	−76.3
PCA	MS-VECM	444.5	−65.7
Full	XGBoost	387.9	−70.1

Notes: The "vs. Naïve" column reports percentage RMSE reduction relative to the naïve random walk in each specification. A strong interaction pattern is evident: MS-VAR dominates in parsimonious and BoP sets, while XGBoost dominates in monetary and full sets.

5.6 Variable Set Robustness

Table 8 reveals a striking interaction between model class and variable set. MS-VAR dominates in the parsimonious and BoP specifications (RMSE ≈ 312–315), while XGBoost dominates in the monetary and full specifications (RMSE ≈ 307–388). This is consistent with regime models preferring compact, structurally relevant systems where state-dependent dynamics are cleanly identified, while tree models can exploit broader nonlinear feature spaces in richer specifications. The PCA specification is best served by MS-VECM (444.5), suggesting that dimensionality reduction partially preserves the regime structure that switching models exploit.

6 Mechanism Analysis

6.1 Architecture versus Information: The 2×2 Decomposition

Table 9: Architecture-Information Decomposition: 2×2 RMSE Matrix

	Parsimonious	BoP	Row Mean
MS-VAR	311.8	315.2	313.5
XGBoost	640.6	886.1	763.4
Column Mean	476.2	600.7	

Notes: Architecture effect (MS-VAR advantage, averaged across varsets) ≈ 464.5 RMSE points. Information effect ≈ −124.5 (BoP is worse on average, driven by XGBoost). Interaction (DiD) ≈ 173.8, indicating that the architecture gain is differentially larger in the parsimonious specification.

Table 9 presents the factorial decomposition. The architecture effect is dominant: MS-VAR outperforms XGBoost by approximately 464 RMSE points on average across variable sets. The information effect is negative on average (−124.5), indicating that the BoP specification does not uniformly improve forecasts—it helps regime-switching models marginally but substantially hurts XGBoost, whose feature importance analysis (Section 6.5) reveals heavy reliance on autoregressive reserve features rather than rich BoP components. The interaction term (DiD ≈ 173.8) is positive and large, indicating that the model architecture and information set are complements for MS-VAR but substitutes for XGBoost: regime-switching models effectively exploit the structural information in disaggregated flows, whilst tree models are overwhelmed by the additional noise.

6.2 Regime Characterisation

Table 10 reports regime diagnostics. Both regimes are highly persistent (self-transition probabilities 0.943 and 0.983), with the accumulation regime exhibiting substantially

Table 10: MS-VAR Regime Characterisation: BoP Specification

	Regime 0 (Crisis)	Regime 1 (Accumulation)
Self-transition probability	0.943	0.983
Expected duration (months)	17.6	57.2
Share of observations	74.6%	25.4%
Classification Quality		
Mean max probability	0.997	
Share with Pr ≥ 0.90	99.5%	
Mean entropy	0.012	
Estimation Diagnostics		
Converged	No (max iterations)	
Log-likelihood	-38.28	
Observations (differenced)	193	

Notes: Transition probabilities from EM estimation on the BoP variable set. Expected duration $= 1/(1 - p_{ii})$ where p_{ii} is the self-transition probability. Regime labels assigned based on estimated mean vectors.

longer expected duration (57.2 months) than the crisis regime (17.6 months). Classification certainty is near-perfect: the mean maximum state probability is 0.997, and 99.5% of observations are classified with probability exceeding 0.90, yielding very low entropy (0.012). This sharp regime separation is economically intuitive: the dynamics of reserve accumulation (gradual, driven by persistent BoP surpluses) are qualitatively different from crisis depletion (rapid, driven by sudden stops and peg defence), and the data strongly distinguish between these states.

A caveat is that the full-sample EM estimation did not converge within the iteration limit, though split-based runs on the training-validation subsample did converge. This sensitivity suggests that multi-start estimation with higher iteration limits and convergence diagnostics should be standard practice before making strong causal claims based on regime parameters.

6.3 Impulse Response Analysis

Regime-conditional generalised impulse response functions reveal strong asymmetry in shock propagation across regimes. For the reserve response to its own shock, the peak magnitude in the accumulation regime ($\approx$559) is roughly three times larger than in the crisis regime ($\approx$173), with a mean absolute peak delta of approximately 97.3 across all target variables. Half-life differences average 9.0 months across variables. This asymmetry is economically meaningful: during accumulation, reserve shocks propagate through the system via policy responses (exchange rate management, import liberalisation) that amplify the initial impact; during crisis, feedback channels are suppressed by binding constraints (depleted reserves, capital controls, suspended convertibility) that truncate shock propagation. The ability of the MS-VAR to capture these regime-specific dynamics—rather than averaging across them as linear models must—is a key source of its forecasting advantage.

6.4 Information Loss Under Aggregation

Table 11: Information Loss from Flow Aggregation

Period	Cancellation Index	Information Loss	RMSE Improvement	Share Helped
All	0.063	93.8%	−1.68%	57.1%
Crisis (2020–2022)	0.074	92.6%	−1.68%	57.1%
Tranquil (2023–2025)	0.059	94.1%	−2.92%	57.1%

Notes: Cancellation Index = $|A_t|/\sum|x_{jt}|$ where A_t is the aggregate and x_{jt} are gross flows. Information Loss = 1 − CI. RMSE Improvement = mean percentage change when moving from aggregated to disaggregated specification. Share Helped = proportion of models for which disaggregation reduces RMSE.

Table 11 reports the empirical information loss analysis. The mean cancellation index is strikingly low at 0.063, implying that approximately 94% of the information in gross BoP flows is destroyed by aggregation into net flows. This severe cancellation arises because exports and imports are both large and move in partially offsetting directions: during the crisis, export growth of 14.3% coexisted with import contraction of 26%, producing modest net changes that mask dramatic gross flow dynamics. The RMSE improvement from disaggregation, while modest in percentage terms (−1.68% to −2.92%), is model-dependent: disaggregation helps regime-switching models (BVAR, MS-VECM) but hurts others (notably XGBoost, which shows a −20% RMSE degradation when moving to BoP). This asymmetric pattern is consistent with the architecture-information interaction identified in the 2×2 decomposition.

6.5 Feature Importance

XGBoost feature importance analysis reveals that the model relies overwhelmingly on autoregressive reserve features rather than rich BoP components. In the BoP specification, the 3-month moving average of reserves alone accounts for approximately 75% of total feature importance, with the next four features (reserve lags at horizons 1–3 and rolling standard deviation) contributing a further 13%. Disaggregated BoP variables (remittances, exports, imports, tourism) collectively account for less than 5% of importance. This finding explains why XGBoost gains little from moving to the BoP specification: it effectively ignores the additional information, treating the BoP variables as noise. By contrast, the MS-VAR framework is structurally designed to model the joint dynamics of all system variables, and its regime-switching parameters capture the state-dependent correlations between BoP components and reserves that the tree model cannot efficiently exploit.

6.6 Crisis Segment Performance and the Meese-Rogoff Puzzle

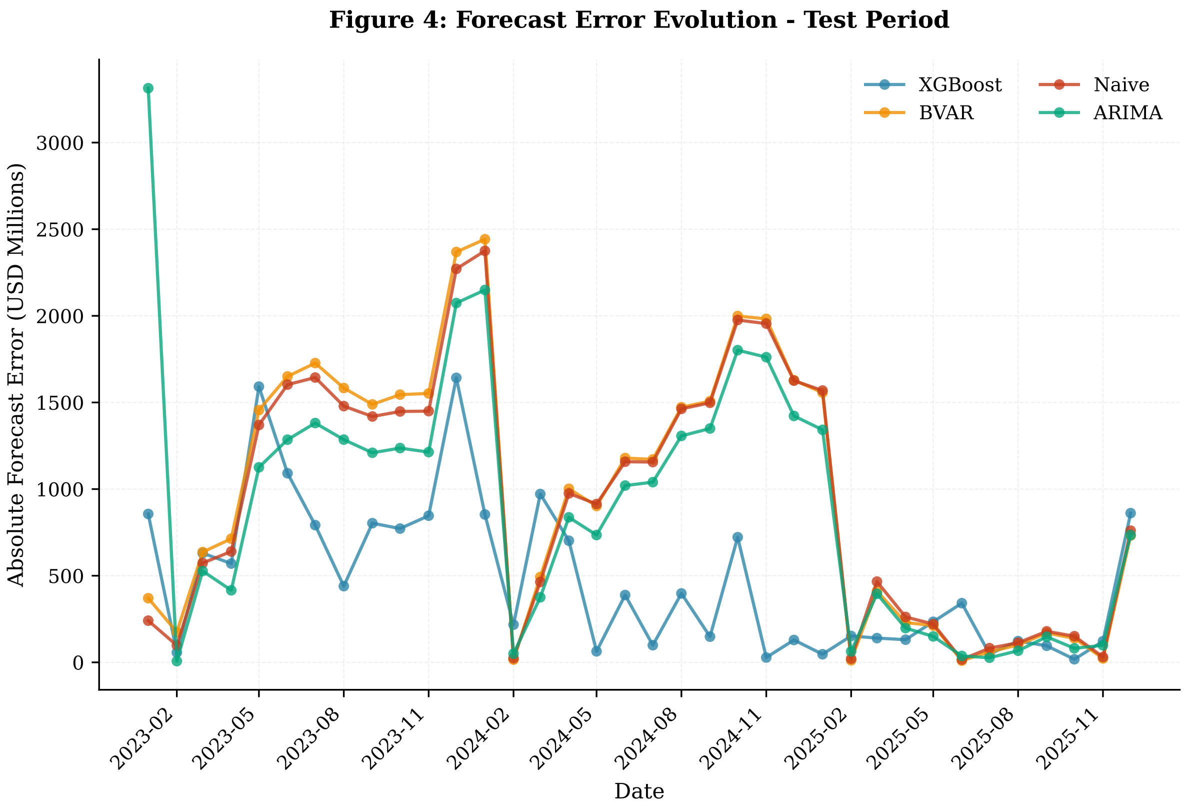

Figure 3: Absolute forecast errors over the test period (2023–2025) for selected models. MS-VAR maintains consistently lower errors, though all models exhibit episodic spikes corresponding to months of large reserve movements.

Table 12: Crisis Segment Performance: Best Models by Period

Segment	Best Model	RMSE	Naïve RMSE
Crisis (2020–2022)	MS-VAR	367.1	1338.9
Post-default (2023–2025)	MS-VAR (BoP)	260.6	1350.8
All (2020–2025)	MS-VAR	311.8	1178.9

Notes: Best model RMSE compared to the naïve random walk for each evaluation segment.

The results in Table 12 reveal that the Meese-Rogoff puzzle does not hold in its strong form for reserves. The naïve random walk is beaten substantially by MS-VAR even during the crisis segment: crisis RMSE of approximately 367 versus naïve crisis RMSE of 1,339 represents a 72.6% improvement. This finding is significant because it suggests that regime-switching models, when applied to appropriately disaggregated data, can capture the dynamics of reserve depletion during sudden stops—a result that extends beyond the standard Meese-Rogoff finding, which was formulated in the context of exchange rate models applied to aggregated fundamentals.

However, the puzzle does retain a weaker form: among non-regime-switching models, the naïve benchmark remains competitive during crisis periods, with only XGBoost (RMSE 942.5 in the BoP crisis segment) achieving substantial improvement. Classical econometric models (ARIMA, VECM) and BVAR do not reliably outperform the random walk during crisis.

7 Density Forecast Evaluation

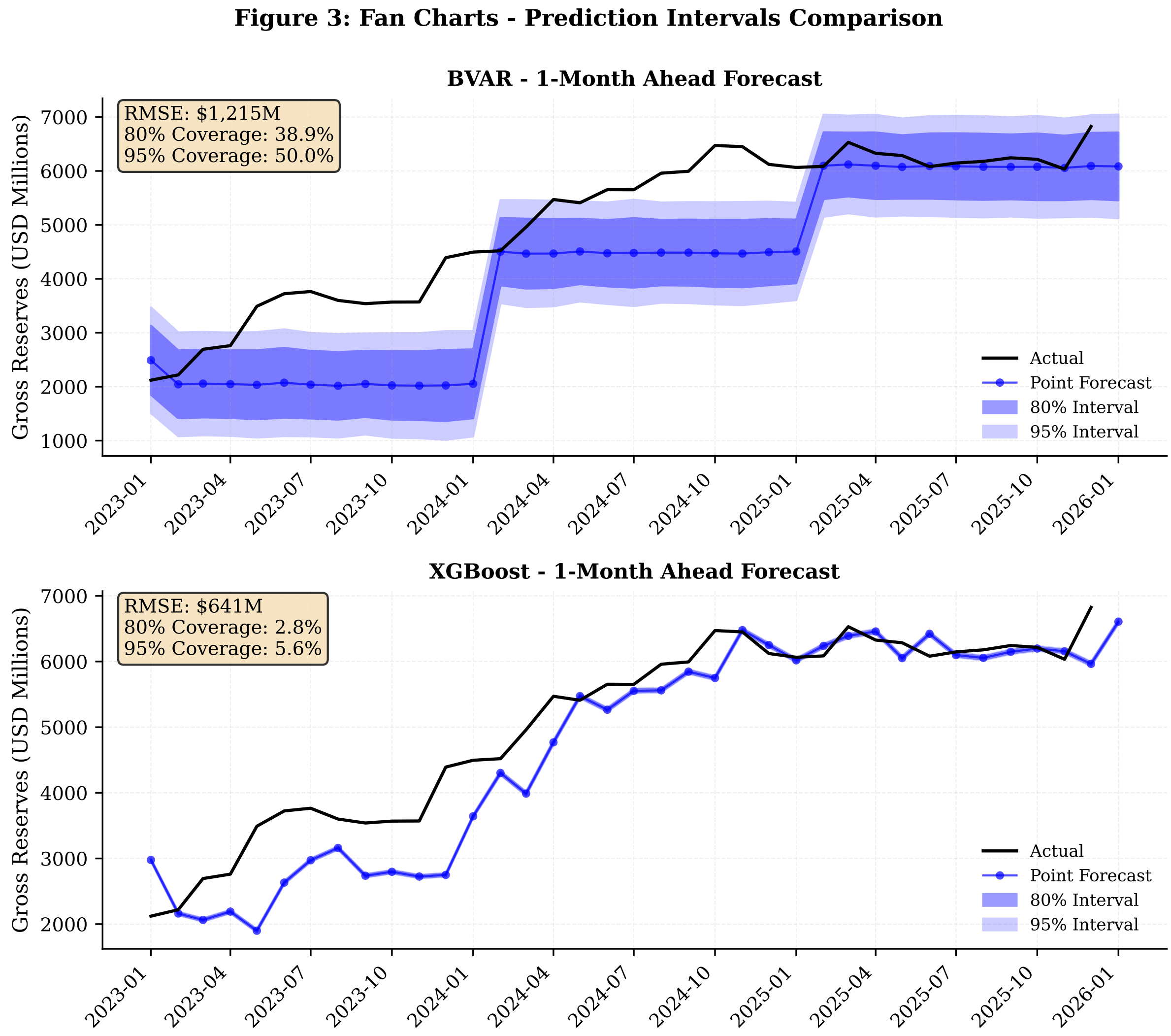

Figure 4: Prediction interval fan charts for BVAR (left) and XGB-Quantile (right). XGB-Quantile produces sharper, better-calibrated intervals; BVAR intervals are wider but exhibit substantial under-coverage relative to nominal levels.

Table 13: Density Forecast Evaluation: Parsimonious Specification, $h = 1$

Model	CRPS	Log Score	Coverage$_{80}$ (%)	Coverage$_{95}$ (%)
XGB-Quantile	**241.0**	–	**72.2**	**97.2**
XGBoost	466.4	−928.95	2.8	5.6
ARIMA	734.2	−10.12	41.7	50.0
Naïve	761.5	−9.88	44.4	52.8
VECM	782.4	−10.30	36.1	52.8
BVAR	792.1	−10.18	38.9	50.0

Notes: Coverage$_{80}$ and Coverage$_{95}$ = empirical coverage rates of 80% and 95% prediction intervals. Bold indicates best. Models without probabilistic output are marked with dashes.

Table 13 evaluates probabilistic forecasting performance. XGB-Quantile achieves the best overall density forecast performance, with the lowest CRPS (241.0) and the best coverage rates (72.2% at 80% nominal, 97.2% at 95% nominal). Classical specifications (ARIMA, VECM) and BVAR exhibit poor coverage (36–50% at 80% nominal), indicating that their estimated densities substantially underestimate forecast uncertainty, likely because parameter estimates are unstable across regimes. The XGBoost point-forecast model produces extremely narrow implicit intervals (2.8%/5.6% coverage), highlighting the gap between point forecast accuracy and uncertainty quantification.

This evidence suggests that a two-model approach is optimal for policy applications: regime-switching models for point forecasting and scenario analysis, complemented by XGB-Quantile or conformal prediction methods for probabilistic uncertainty quantification.

8 Scenario Analysis and Robustness

8.1 Scenario Analysis

Table 14 presents component-level scenario projections over a 12-month horizon. The baseline projects reserves at 6,755.7 USD million by end-2025, a modest 1.02% decline. Under a combined adverse scenario (LKR depreciation, export shock, oil price spike, IMF disbursement delay), reserves decline to 6,509.7 million (−4.62%), while a 20% currency depreciation alone produces a 5.82% decline. IMF disbursement delays prove particularly consequential (−2.94%), underscoring the sensitivity of the recovery trajectory to programme timing.

Variable set sensitivity analysis reveals that the full specification is most vulnerable to adverse shocks (9.97% decline under combined adverse in the full specification), whilst the parsimonious specification is most stable (4.62%). The BoP specification uniquely captures remittance decline impacts (−3.78%) and tourism recovery effects (+0.89%)

Table 14: MS-VARX Scenario Analysis: Reserve Projections (Parsimonious)

Scenario	End Level	Δ (USD m)	Δ (%)
Combined Upside	6837.7	12.7	0.19
Baseline	6755.7	−69.3	−1.02
Oil Price Shock	6724.9	−100.1	−1.47
Export Shock (−15%)	6718.7	−106.3	−1.56
IMF Tranche Delay	6624.5	−200.5	−2.94
LKR Depreciation 10%	6591.7	−233.3	−3.42
Combined Adverse	6509.7	−315.3	−4.62
LKR Depreciation 20%	6427.8	−397.2	−5.82

Notes: Projections based on MS-VARX model with parsimonious specification. 12-month horizon (2025:01–2025:12). Baseline from 2024:12 level of 6,825.0. Remittance and tourism shocks have zero impact under this specification as those variables are excluded; their effects propagate in BoP and fuller specifications.

that are invisible in the parsimonious system.

8.2 Robustness to Train/Test Split

Split choice materially affects regime diagnostics: the training-validation subsample converged where the full-sample and training-only estimations often did not, and estimated regime durations and classification certainty differ across splits. This sensitivity should be treated as a formal robustness axis: production forecasting systems should report results across multiple split configurations and flag cases where regime diagnostics are qualitatively different.

9 Conclusion

This paper provides a comprehensive multi-model forecasting comparison for Sri Lanka's foreign exchange reserves, evaluated across five variable sets and multiple forecast horizons on a dataset encompassing stable accumulation, pandemic disruption, sovereign default, and IMF-supervised recovery. The analysis goes beyond a standard model horse-race by introducing a formal architecture-versus-information decomposition and providing mechanistic evidence for why certain models succeed.

The central result is the dominance of Markov-Switching VAR, which achieves RMSE reductions of roughly three-quarters relative to the naïve benchmark and is the sole member of the Model Confidence Set at the 10% level. The 2×2 factorial decomposition establishes that this advantage is primarily architectural: regime-switching structure accounts for the bulk of the improvement over gradient-boosted trees, with a significant interaction effect indicating that architecture and information content are complements for regime-switching models but substitutes for tree-based methods. The information-loss analysis demonstrates that aggregating balance-of-payments flows destroys the vast majority of the available signal through cancellation of offsetting gross flows, providing a formal explanation for why disaggregated specifications are valuable. Regime characterisation reveals highly persistent states with near-perfect classification certainty, and impulse response analysis shows strong regime asymmetry in shock propagation that is economically consistent with distinct crisis and recovery dynamics.

The Meese-Rogoff puzzle does not hold in its strong form for reserves: MS-VAR substantially outperforms the naïve benchmark even during the acute crisis segment. However, a weaker form of the puzzle persists for non-regime-switching models, which struggle to beat the random walk during crisis periods. XGBoost dominates in the monetary and full specifications, suggesting that tree models can exploit broader nonlinear feature spaces when the dimensionality is sufficient, though their heavy reliance on autoregressive reserve features limits their ability to exploit structural BoP information.

9.1 Limitations and Future Work

Several limitations circumscribe these findings and suggest directions for future research. First, the evaluation of Dynamic Model Averaging [Koop and Korobilis, 2012] and Dynamic Model Selection is incomplete, as these adaptive methods were not integrated into the unified evaluation framework; doing so would enable assessment of whether adaptive model selection across regimes can further improve forecast accuracy. Second, the full-sample MS-VAR estimation did not converge within the iteration limit, and regime diagnostics exhibit sensitivity to the estimation window. Multi-start estimation with higher iteration limits and formal convergence diagnostics should be standard practice before making strong causal claims based on regime parameters. Third, LSTM robustness and broader hyperparameter sensitivity remain incompletely explored; given the well-documented sensitivity of recurrent architectures to sequence length, learning rate, and regularisation, more exhaustive tuning could alter the relative ranking of neural network approaches. Fourth, the single-country design limits generalisability. Application to a cross-country panel of emerging markets—particularly those with different external vulnerability profiles—would enable investigation of whether the MS-VAR advantage is specific to Sri Lanka's crisis dynamics or reflects a more general property of regime-switching models in crisis-prone settings. Fifth, real-time data vintage effects are not explicitly modelled; the use of revised data may overstate forecast accuracy relative to what would have been achievable in real time.

Several promising avenues emerge for future research. Conformal prediction methods offer a principled approach to combining regime-switching point-forecast accuracy with distribution-free uncertainty quantification, potentially resolving the gap between MS-

VAR's point forecast dominance and XGB-Quantile's superior density calibration identified in this study. Ensemble methods combining regime-switching models with machine learning via stacking could exploit the complementary strengths of both approaches—structural regime identification from MS-VAR and flexible nonlinear feature extraction from tree-based methods. Finally, extending the information-loss framework to other macroeconomic aggregates (GDP components, monetary aggregates) could test whether the severe cancellation documented here for balance-of-payments flows is a broader phenomenon in macroeconomic forecasting.

Despite these limitations, the evidence provides strong support for explicitly modelling regime dynamics in disaggregated balance-of-payments systems for reserve forecasting in crisis-prone emerging markets. Central banks confronted with reserve surveillance and adequacy assessment would be well-advised to construct forecasting frameworks that pair regime-switching models for scenario analysis with complementary probabilistic methods for uncertainty quantification, given the high policy stakes where adequate buffers can mean the difference between weathering a sudden stop and spiralling into sovereign default.

References

Aizenman, J. and Lee, J. (2007). International reserves: precautionary versus mercantilist views, theory and evidence. *Open Economies Review*, 18(2):191–214.

Athukorala, P. (2024). Sri Lanka's external sector crisis: causes and consequences. *World Development*, 177:106556.

Banbura, M., Giannone, D., and Reichlin, L. (2010). Large Bayesian vector auto regressions. *Journal of Business & Economic Statistics*, 28(1):57–75.

Brunetti, C., Mariano, R. S., Scotti, C., and Tan, A. H. (2007). Markov switching and stochastic volatility in currency crises. *Journal of Applied Econometrics*, 22(4):765–789.

Calvo, G. A. (1996). Varieties of capital flows and their implications for the economy. In *The Economic of Globalization: Policy Perspectives from the East Asia and Pacific*, pages 15–41. World Bank, Washington, DC.

Calvo, G. A. (1998). Contagion, well-behavedness and sudden stops. *NBER Working Paper No. 5854*.

Calvo, G. A., Izquierdo, A., and Loo-Kung, R. (2013). Optimal holdings of international reserves: self-insurance against sudden stops. *Journal of International Economics*, 87(2):266–286.

Carriero, A., Clark, T. E., and Marcellino, M. (2015). Realizing the gains from Bayesian dynamic model averaging. *International Journal of Forecasting*, 31(4):1009–1023.

Central Bank of Sri Lanka (2021). Annual Report 2021. Central Bank of Sri Lanka, Colombo.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

Cheung, Y. W., Chinn, M. D., and Pascual, A. G. (2005). Empirical exchange rate models of the nineties: Are any fit to survive? *Journal of International Money and Finance*, 24(7):1150–1175.

Doan, T. A., Litterman, R. B., and Sims, C. A. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, 3(1):1–100.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Frankel, J. A. and Saravelos, G. (2012). Can we predict the next financial crisis? *NBER Working Paper No. 18725.*

Greenspan, A. (1999). Currency reserves and debt. *Remarks before the World Bank Conference on Recent Trends in Reserves Management*, Washington, DC.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometric Reviews*, 30(2):160–201.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688.

IMF (2011). Assessing reserve adequacy. *IMF Policy Paper.*

IMF (2015). Assessing reserve adequacy—review of the adequacy approach. *IMF Policy Paper.*

IMF (2022). Sri Lanka: Request for an extended arrangement under the Extended Fund Facility. *IMF Country Report No. 23/99.*

Jeanne, O. and Rancière, R. (2011). The optimal level of international reserves for emerging market countries. *Journal of International Economics*, 85(2):229–240.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12(2-3):231–254.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica*, 59(6):1551–1580.

Kaminsky, G. L., Lizondo, S., and Reinhart, C. M. (1998). Leading indicators of currency crises. *IMF Staff Papers*, 45(1):1–48.

Koop, G. and Korobilis, D. (2012). Forecasting inflation using dynamic model averaging. *International Economic Review*, 53(3):867–886.

Krolzig, H. M. (1997). *Markov-Switching Vector Autoregressions: Modelling, Statistical Inference, and Application to Business Cycle Analysis.* Springer, Berlin.

Litterman, R. B. (1986). Forecasting with Bayesian vector autoregressions—five years of experience. *Journal of Business & Economic Statistics*, 4(1):25–38.

Medeiros, M. C., Vasconcelos, G. F., Veiga, Á., and Zilberman, E. (2021). Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business & Economic Statistics*, 39(1):98–119.

Meese, R. A. and Rogoff, K. (1983). Empirical exchange rate models of the seventies: Do they fit out of sample? *Journal of International Economics*, 14(1-2):3–24.

Obstfeld, M., Shambaugh, J. C., and Taylor, A. M. (2010). Financial stability and the global safeguard system. In *NBER International Seminar on Macroeconomics 2009*, pages 9–37. University of Chicago Press.

Peria, M. S. M. (2002). A regime-switching approach to speculative attacks: a focus on European currencies. *Journal of International Economics*, 57(2):467–490.

Raftery, A. E., Kárný, M., and Ettler, P. (2010). Online prediction under model uncertainty via dynamic model averaging: application to a cold rolling mill. *Technometrics*, 52(2):52–66.

Richardson, A., Mulder, T., and Vehbi, T. (2021). Nowcasting New Zealand GDP using machine learning algorithms. *International Journal of Forecasting*, 37(2):736–759.

Rossi, B. (2013). Exchange rate predictability. *Journal of Economic Literature*, 51(4):1063–1119.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48(1):1–48.

Taylor, M. P., Peel, D. A., and Sarno, L. (2001). Nonlinear mean-reversion in real exchange rates: towards a solution to the purchasing power parity puzzles. *International Journal of Finance & Economics*, 6(4):299–318.

UNDP (2022). The Sri Lankan economic crisis: impacts and policy responses. *UNDP Sri Lanka Report.*

Weerakoon, D. and Jayasuriya, S. (2023). The political economy of Sri Lanka's debt crisis. *South Asia: Journal of South Asian Studies*, 46(1):123–145.

Wignaraja, G. (2024). Emerging Asia's manufacturing challenge: trade, technology and structural change. *Asian Development Review*, 41(1):56–89.

Wijnholds, J. O. and Kapteyn, A. (2001). Reserve adequacy in emerging market economies. *IMF Working Paper WP/01/143.*

A Estimation Details and Metric Definitions

A.1 Classical Models

ARIMA models with exogenous regressors are estimated via maximum likelihood, with lag order selected via AIC across ARIMA(p, d, q) specifications with $p, q \in \{0, 1, 2, 3\}$ and $d \in \{0, 1\}$. VAR and VECM models are estimated via full-information maximum likelihood with lag order selected by AIC (maximum 4 lags). Cointegration rank is determined by the Johansen trace and maximum eigenvalue tests.

A.2 Bayesian VAR

The BVAR Minnesota prior is centred on a random walk with hyperparameters: overall shrinkage $\lambda \in [0.1, 1.5]$, lag decay $\mu \in [0.1, 0.8]$, and own-lag variance $\phi \in [0.1, 0.95]$, selected via grid search on in-sample BIC. The posterior is simulated via Gibbs sampling with 10,000 draws (2,000 burn-in).

A.3 Machine Learning

XGBoost: 200 boosting rounds (300 for BoP), learning rate $\eta = 0.1$, tree depth 3–5, minimum child weight 1, subsample ratio 0.8–1.0, column subsample 0.8, time-series cross-validation. Features include lagged dependent variable (lags 1–12), lagged exogenous variables (lags 1–6), and rolling 3- and 6-month means and volatilities.

LSTM: single recurrent layer of 64 units, input sequence length 12, Adam optimiser (learning rate 0.001), dropout 0.2, batch size 32, maximum 100 epochs with early stopping (patience 10).

A.4 Accuracy Metrics

RMSE $= \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$; MAE $= \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|$; MAPE $= \frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|$

MASE $=$ MAE$\big/\frac{1}{n-1}\sum_{t=2}^{n}|y_t - y_{t-1}|$; Policy loss: $L(e_t) = e_t^2$ if $e_t < 0$, $0.5e_t^2$ if $e_t \geq 0$.

CRPS $= \frac{1}{n}\sum_{t=1}^{n}\int_{-\infty}^{\infty}[F_t(z) - \mathbf{1}_{z \geq y_t}]^2 dz$; Log score: $\log S_t = \log p_t(y_t)$.

Diebold-Mariano statistic: DM $= \bar{d}/\sqrt{2\pi f_d(0)/n}$, asymptotically standard normal under equal expected loss.