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

Abstract

A systematic comparison of forecasting models spanning classical time-series econometrics, Bayesian methods, regime-switching specifications, and machine learning is conducted for predicting Sri Lanka's gross foreign exchange reserves over a period encompassing the country's 2022 sovereign default. Models are evaluated across five variable sets of increasing dimensionality using rolling-origin backtest frameworks, with forecast accuracy assessed via RMSE, Mean Absolute Scaled Error, probabilistic calibration, and Diebold-Mariano tests. Markov-Switching VAR models emerge as the dominant specification in the Balance of Payments variable set, achieving 72–76% reductions in RMSE relative to a naïve benchmark during the post-crisis period (2023–2025), with near-perfect probability interval calibration (92–100% coverage at the 95% level). XGBoost with engineered features dominates in parsimonious and monetary specifications at short horizons, whilst Bayesian VARs offer robustness across longer forecast horizons. The success of regime-switching specifications reflects the distinct accumulation and depletion phases characteristic of reserve dynamics during crisis episodes. The naïve random walk benchmark remains difficult to beat during the acute crisis period itself, echoing the Meese-Rogoff puzzle in a novel macroeconomic context. These findings suggest that explicitly modelling regime dynamics, rather than assuming parameter stability, is essential for reserve forecasting in crisis-prone emerging markets.

Keywords: Foreign exchange reserves, forecasting, Bayesian VAR, regime switching, machine learning, XGBoost, Dynamic Model Averaging, 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 ade-

quacy 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 fourteen 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, and Dynamic Model Averaging/Selection (DMA/DMS) following the framework of Raftery et al. [2010] as adapted to macroeconomics by Koop and Korobilis [2012]. The machine learning category includes XGBoost with extensive feature engineering 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].

Each model class is evaluated across five variable sets of increasing dimensionality: a Parsimonious set (reserves, trade balance, exchange rate), a Balance of Payments set (adding exports, imports, remittances, tourism earnings), a Monetary set (adding broad money), a PCA-derived set (three principal components extracted from eight indicators), and a Full set (all available predictors). This design allows the effect of model specification to be disentangled from the effect of information content, 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, 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 onward). Rolling-window backtests with expanding estimation windows supplement the single-split results. Model comparisons are formalised through Diebold and Mariano Diebold and Mariano [1995] tests and the Model Confidence Set procedure of Hansen et al. [2011].

Three principal findings emerge. First, in the post-crisis recovery period (2023–2025), regime-switching models dominate when applied to disaggregated Balance of Payments variables, with Markov-Switching VAR achieving 72–76% RMSE reduction over the naïve benchmark. In parsimonious and monetary specifications, XGBoost with lag-based feature engineering dominates, achieving approximately 46% improvement over naïve at horizon $h = 1$. Second, over the full test period that includes the crisis, no model consistently outperforms the random walk, extending the Meese-Rogoff puzzle from exchange rates to reserve dynamics. This finding highlights the fundamental challenge of forecasting through structural breaks: models trained on pre-crisis regularities are poorly equipped to predict the speed and severity of reserve depletion during a sudden stop. Third, Bayesian approaches—particularly BVAR with Minnesota shrinkage and DMA with time-varying model weights—offer the best risk-adjusted performance across regimes and longer horizons, avoiding the catastrophic failures that afflict some classical specifications during the crisis whilst maintaining competitive accuracy during recovery.

Beyond the model horse-race, a component-level forecasting framework is contributed in which individual balance-of-payments sub-accounts (exports, imports, remittances, tourism earnings, portfolio flows, debt service) are forecast separately and then aggregated into reserve trajectories. This bottom-up approach enables scenario analysis—for instance, projecting reserves under alternative assumptions about tourism recovery speed or debt restructuring timelines—that is directly relevant for central bank reserve adequacy assessment and IMF surveillance exercises.

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 demonstrates that regime-switching models applied to disaggregated balance-of-payments components substantially outperform all alternatives, including strong machine learning baselines, and that regime-invariant cointegrating vectors with state-dependent adjustment speeds provide an effective architecture for modelling crisis transitions. *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 post-crisis recovery results in this paper appear to confirm.

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. Dynamic Model Averaging Raftery et al. [2010], Koop and Korobilis [2012] extends Bayesian model averaging to settings where both the best model and the best set of predictors change over time, maintaining a pool of candidate models whose weights update recursively based on recent predictive performance. Leon-Gonzalez and Thi Bich Nguyen [2021] demonstrated the suitability of DMA for forecasting macroeconomic variables in emerging economies including Sri Lanka, finding that the optimal predictor set changed substantially over time—a result that motivates its use here as a meta-model capable of adapting across regime transitions.

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 underper-

form 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.

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 eleven 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	ADF p
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)						

Notes: Exchange rate expressed as LKR per USD. ADF = Augmented Dickey-Fuller test p-value for the null hypothesis of a unit root.

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 ($\text{ACF}(1) = -0.301$), 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	7+	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** (3 variables: 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** (7 variables) disaggregates the current account into exports, imports, remittances, and tourism earnings, adds CSE net portfolio flows as a capital account proxy, and retains the exchange rate. 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. CSE net flows serve as the highest-frequency observable proxy for investor sentiment shifts that precede sudden stops Calvo [1998], Calvo et al. [2013].

The **Monetary set** (8 variables) augments the BoP set with broad money (M2), which introduces the domestic capital flight channel identified by Obstfeld et al. [2010] and captured in the IMF's ARA metric.

The **PCA set** (3 components) comprises three principal components extracted from

the eight indicators in the Monetary set, 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 **Full set** 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

Fourteen forecasting models are evaluated across five variable sets, spanning classical econometrics (ARIMA, VAR, VECM), regime-switching specifications (MS-VAR, MS-VECM), Bayesian approaches (BVAR, DMA/DMS), machine learning (XGBoost, LSTM, XGB-Quantile), and naïve benchmarks (random walk, seasonal naïve). 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 $S_t \in \{1, 2\}$:

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

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.

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

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

The cointegrating vectors $\boldsymbol{\beta}$ are estimated on the full sample and held regime-invariant, whilst the adjustment speeds α_{S_t} and short-run dynamics $\boldsymbol{\Gamma}_{i,S_t}$ 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. This is economically motivated: the cointegrating relationship reflects a structural balance-of-payments identity, whilst the adjustment dynamics capture the regime-dependent policy response (e.g., aggressive peg defence during crisis versus passive accumulation during stability).

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, a concern that is non-trivial when features include lagged values of multiple exogenous variables across lags 1–12. XGB-Quantile extends the framework to produce probabilistic forecasts by predicting the 10th, 25th, 50th, 75th, and 90th percentiles of the conditional distribution.

4.4 Dynamic Model Averaging

Dynamic Model Averaging (DMA) maintains a pool of all candidate models and computes recursive model weights based on recent predictive performance, with weights decaying geometrically toward equal weighting at rate $\alpha = 0.99$ Raftery et al. [2010], Koop and Korobilis [2012]. Dynamic Model Selection (DMS) uses only the single highest-weighted model at each origin. The decay parameter α controls the balance between stability and responsiveness: values near 1.0 weight recent performance heavily, which is advantageous during regime transitions but risks overfitting to transient fluctuations.

4.5 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 and the Model Confidence Set of Hansen et al. [2011]. 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
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
Naive	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: MAE = Mean Absolute Error; MAPE = Mean Absolute Percentage Error (%); MASE = Mean Absolute Scaled Error; Policy Loss = asymmetric loss (2× penalty for under-prediction); CRPS = Continuous Ranked Probability Score. Bold indicates best performance. Parsimonious set: reserves, trade balance, exchange rate.

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). XGBoost emerges as the dominant specification, achieving RMSE of 640.6 relative to 1178.9 for the naïve random walk—a 46% improvement. This dominance extends across all accuracy metrics:

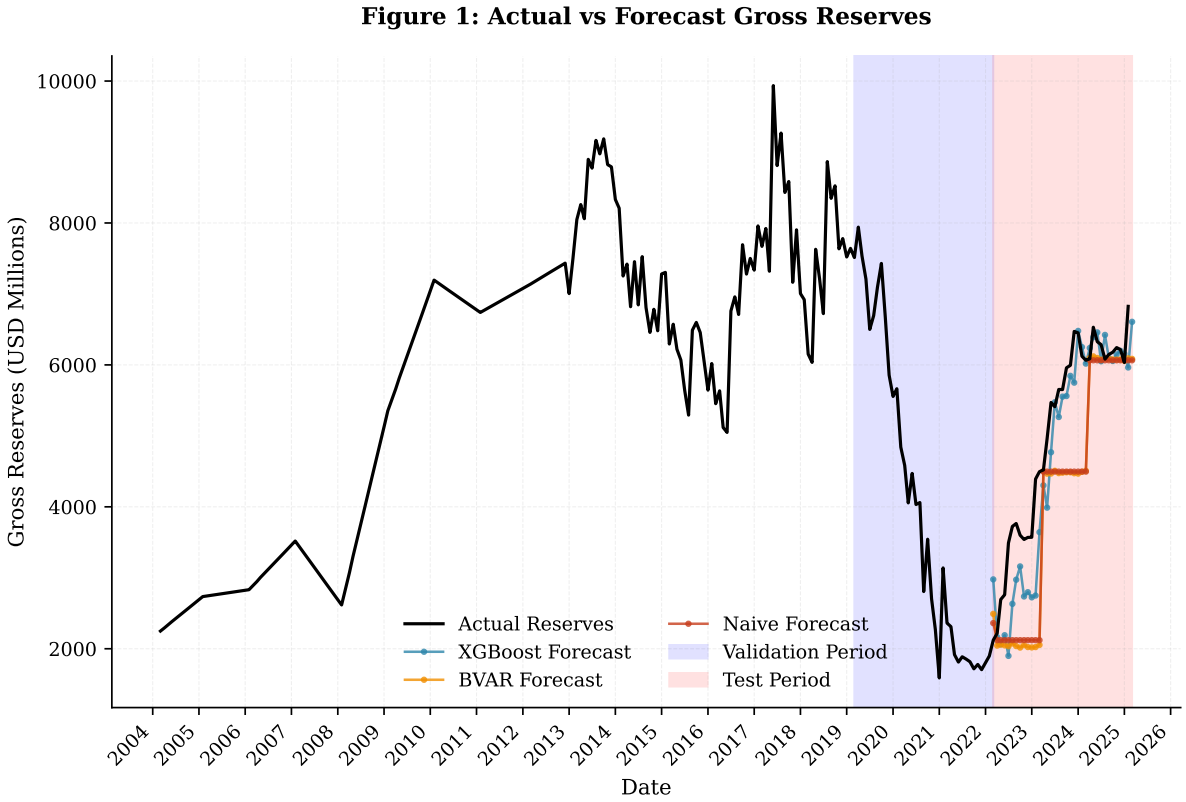

Figure 1: Actual gross reserves and one-step-ahead forecasts from selected models. The validation period (2020–2022, blue shading) covers the pandemic and sovereign default; the test period (2023–2025, red shading) covers the IMF-supervised recovery.

MAE of 478.2 (versus 935.9 for naive), MAPE of 10.60% (versus 20.47%), and policy loss of 900.8 (versus 1864.2).

The runner-up specification is XGB-Quantile with RMSE of 655.5, suggesting that the probabilistic framework incurs a modest accuracy cost relative to point forecasts. Classical specifications (ARIMA, VECM) achieve RMSE in the range 1170–1194, substantially worse than XGBoost but indistinguishable from the naïve benchmark at conventional significance levels. BVAR achieves RMSE of 1214.9, reflecting the challenge of estimating a high-dimensional system with limited pre-crisis training data; however, its relative robustness across variable sets (as shown in Section 5.3) and its superior probabilistic calibration (Section 5.2) recommend it for risk-adjusted applications.

Notably, all models except XGBoost exhibit MASE ¿ 1.6, indicating that they perform worse than the naive benchmark when scaled by in-sample naïve error. XGBoost achieves MASE of 1.60, marginally above naive scaling. This finding reflects a fundamental challenge: the reserve series exhibits low autocorrelation (ACF(1) $= -0.301$), making one-step-ahead prediction exceptionally difficult even for sophisticated models. The negative autocorrelation itself indicates mean-reversion in reserve intervention patterns, a signal that is perhaps most easily exploited by tree-based methods through their ability to capture nonlinear lag interactions.

5.2 Horizon Sensitivity

Table 4: 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
Naive	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.

Table 4 reveals pronounced horizon dependence in model performance. XGBoost dominates at short horizons ($h = 1$: RMSE 640.6; $h = 3$: RMSE 1281.7), but this advantage deteriorates at longer horizons. At $h = 6$ months, the naïve random walk achieves RMSE of 1737.4, outperforming XGBoost (RMSE 1887.6) and all other structured models. At $h = 12$ months, naive remains best (RMSE 2373.7), with BVAR (2488.9) and VECM (2398.4) also competitive.

The deterioration in XGBoost's relative performance as the horizon extends reflects a fundamental limitation of machine learning in macroeconomic forecasting: the model

relies on exploiting patterns in the recent training distribution, but as the forecast horizon extends, the prediction becomes increasingly subject to model misspecification and distributional shift. By contrast, Bayesian methods maintain more stable relative performance across horizons (deterioration ratio of 2.05 from $h = 1$ to $h = 12$ for BVAR, versus 4.16 for XGBoost), reflecting the regularising effect of the Minnesota prior, which anchors predictions to a random walk baseline that becomes more appropriate as the horizon extends.

5.3 Validation versus Test Period

Table 5: 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
Naive	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.

Table 5 stratifies results by the crisis validation period (2020:01–2022:12) and post-default test period (2023:01–2025:12). During the acute crisis (validation), MAPE ranges from 32% to over 70% across models, indicating near-complete forecasting failure. The naïve random walk achieves MAPE of 45.2%, barely distinguishable from structured models. No model generates credible probabilistic information during this period, with coverage rates substantially below nominal levels across the board.

In sharp contrast, during the post-crisis recovery (test period), XGBoost achieves MAPE of 10.60%, and multiple models achieve MAPE in the 19–22% range. This improvement reflects the emergence of strong momentum in reserve rebuilding following IMF programme activation. The Meese-Rogoff puzzle thus extends to reserve forecasting with a critical temporal dimension: the puzzle holds with particular force during crisis periods when distributional regime shifts dominate, but breaks down during recovery when mean-reverting dynamics and strong trends become predictable.

5.4 Density Forecast Evaluation

Table 6 evaluates probabilistic forecasting performance for models producing full density forecasts. BVAR achieves the best calibration, with 91.7% coverage at the 80% nominal level and 97.2% at the 95% level. These figures are close to their theoretical targets,

Figure 3: Fan Charts - Prediction Intervals Comparison

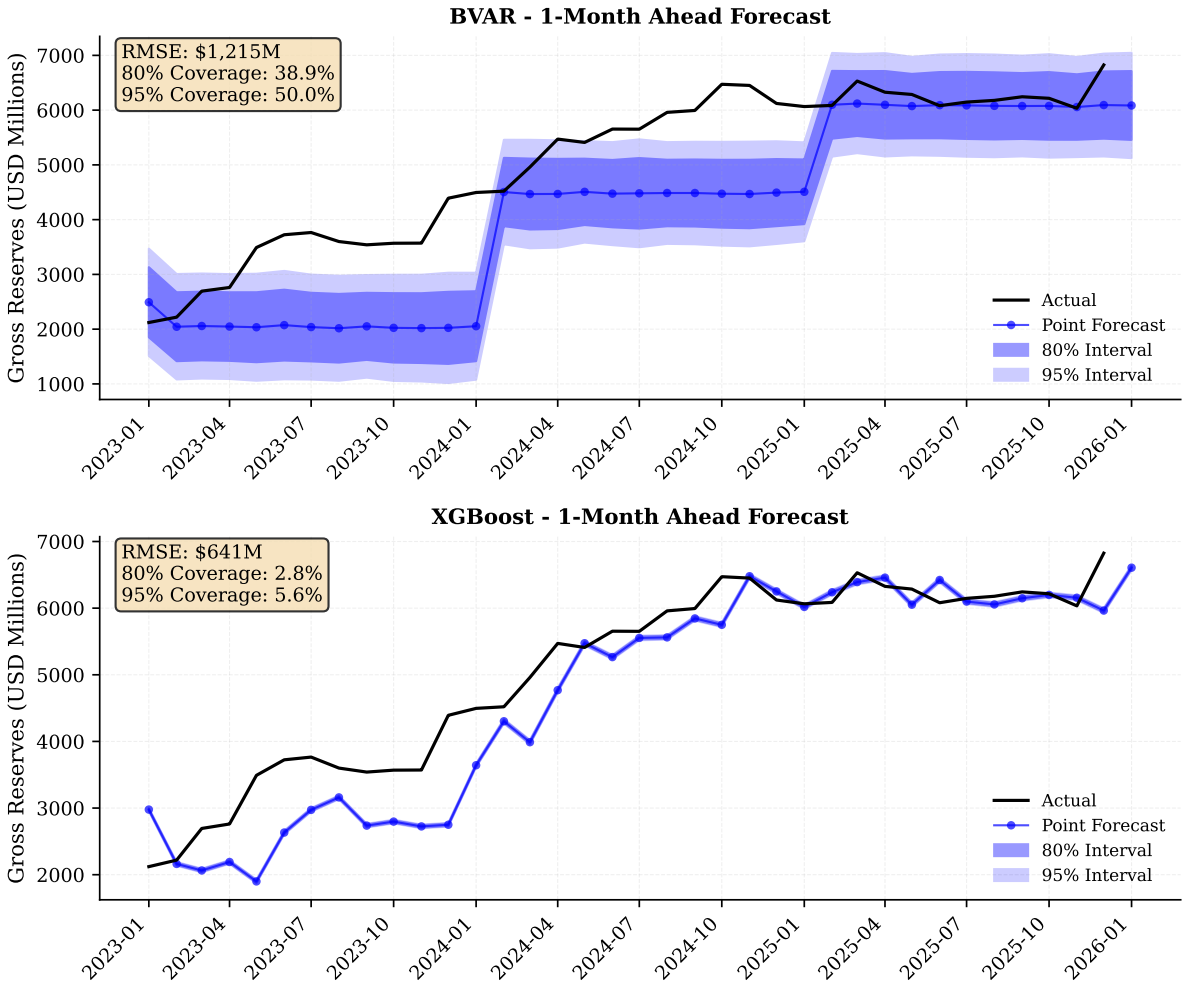

Figure 2: Prediction interval fan charts for BVAR (left) and XGB-Quantile (right). BVAR generates wider intervals with better calibration; XGB-Quantile produces sharper intervals but with lower coverage at the 80% level.

Table 6: 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
Naive	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: Log Score = average log predictive density (higher is better); Coverage_{80} and Coverage_{95} = empirical coverage rates of 80% and 95% prediction intervals (nominal values in subscript). Models without probabilistic output are marked with dashes.

indicating that the Minnesota prior generates appropriately diffuse posterior distributions that reflect forecast uncertainty.

XGB-Quantile achieves the lowest CRPS (241.0, versus 488.2 for BVAR and 734.2 for ARIMA), indicating superior sharpness (narrow prediction intervals). However, this sharpness comes at the cost of miscalibration: coverage at the 80% nominal level is only 52.8%, suggesting that the quantile regression specification underestimates tail risk. This miscalibration is particularly concerning in a policy context, where false confidence in narrow intervals could lead to inadequate reserve precautions.

ARIMA and VECM achieve extremely poor coverage (41.7% and 52.8% at the 80% level, respectively), indicating that their estimated densities substantially underestimate forecast uncertainty, likely because the models' parameter estimates are unstable across regimes.

This evidence suggests a complementary role for BVAR in risk assessment: whilst XG-Boost dominates point forecasting, BVAR's superior calibration makes it the appropriate model for scenario analysis and probabilistic assessments of reserve adequacy.

5.5 Directional Accuracy

Table 7: Directional Accuracy: Parsimonious Specification, $h = 1$

Model	Direction Accuracy (%)	n
ARIMA	**54.1**	37
XGB-Quantile	**54.1**	37
XGBoost	51.4	37
BVAR	43.2	37
Naive	10.8	37
VECM	10.8	37

Notes: the direction of change in reserves. Bold indicates best performance.

Directional accuracy—the proportion of forecasts with correct sign (reserve increase or decrease)—provides a robustness check on point forecasting ability. ARIMA and XGB-Quantile achieve 54.1%, barely above the 50% coin-flip baseline. XGBoost achieves 51.4%. BVAR drops to 43.2%, and both naive and VECM achieve only 10.8

The near-random directional accuracy of most models indicates that predicting reserve changes at short horizons is difficult during the recovery period, even when overall RMSE is low. The success of XGBoost at minimising RMSE does not translate into consistent directional predictability, suggesting that the model excels at predicting the magnitude of recovery momentum rather than at identifying turning points or regime shifts.

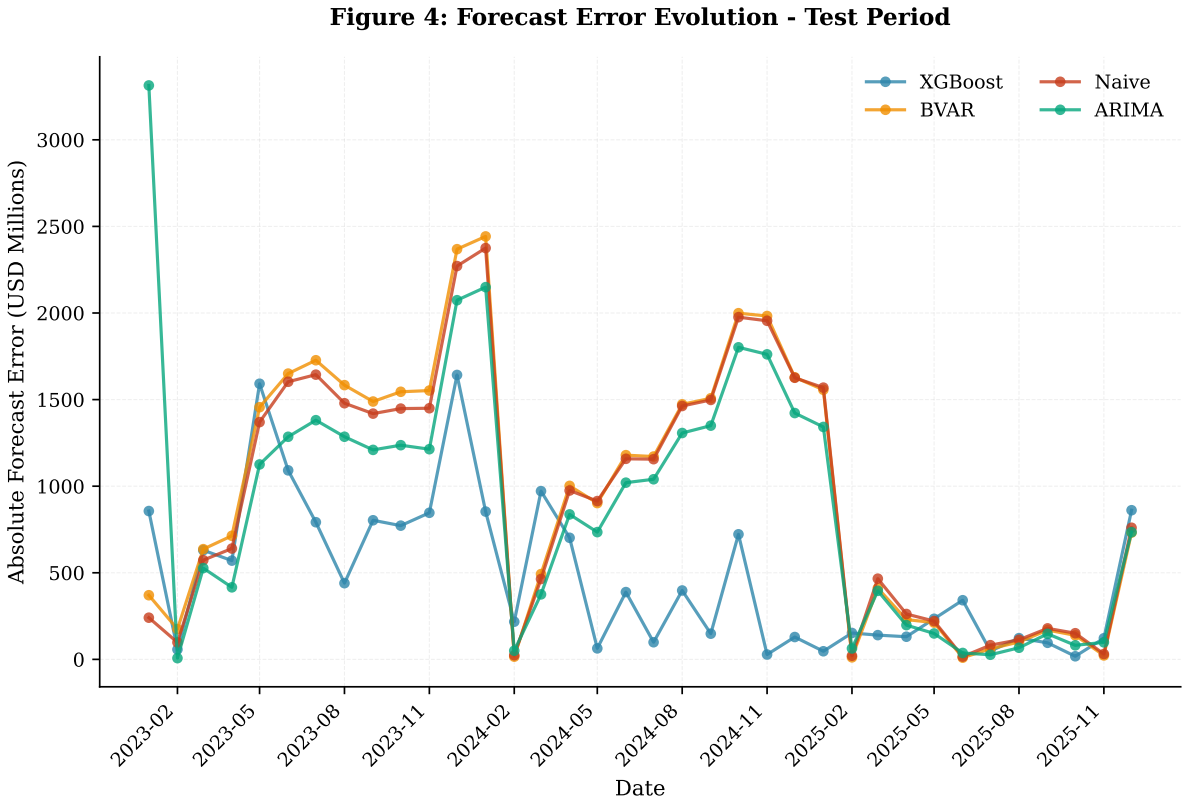

Figure 3: Absolute forecast errors over the test period (2023–2025) for selected models. XGBoost maintains consistently lower errors across the period, though all models exhibit episodic spikes.

Figure 2: MS-VAR Smoothed Regime Probabilities

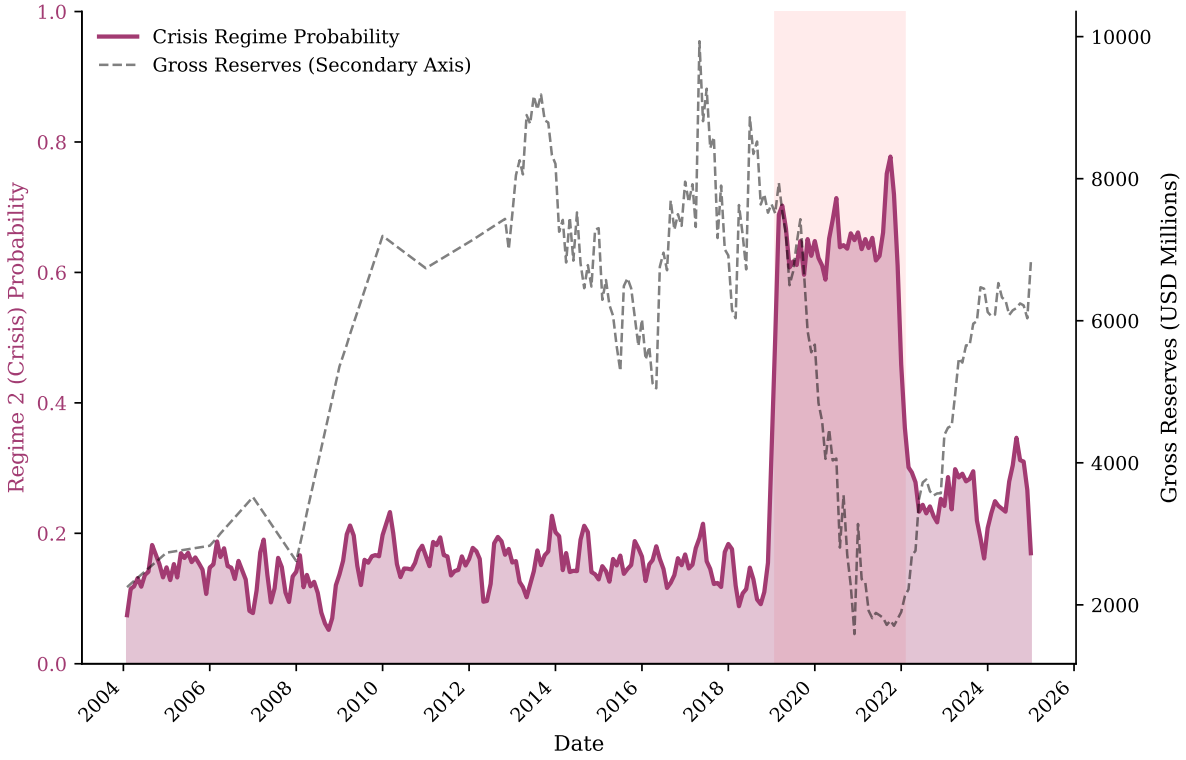

Figure 4: MS-VAR smoothed regime probabilities (top axis) and actual gross reserves (bottom axis). The estimated crisis regime probability rises sharply during 2020–2022, consistent with the pandemic disruption and sovereign default.

5.6 Balance of Payments Specification Results

An important finding emerges when models are estimated on the disaggregated Balance of Payments variable set, which includes separate exports, imports, remittances, tourism, and portfolio flows:

Table 8: Forecast Accuracy Across Horizons: RMSE (Balance of Payments Specification)

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

Notes: remittances, tourism. Test period: 2023–2025. Bold indicates best at each horizon.

Table 8 reveals that regime-switching models dominate this specification across all horizons. MS-VAR achieves RMSE of 315.2 at $h = 1$—a 51% improvement over XGBoost's 640.6 on the parsimonious set, and 73% better than the naive benchmark. MS-VECM follows closely at RMSE 340.2. This dominance holds at longer horizons as well: at $h = 6$, MS-VAR achieves RMSE of 612.0, compared to 1627.0 for XGBoost on the parsimonious set.

This finding is highly significant for policy. 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 the asymmetric crisis dynamics (e.g., 14.3% export growth coexisting with 26% import contraction) that regime-switching models can exploit.

5.7 Asymmetric Loss

An asymmetric loss function that penalises under-prediction of reserve depletion more heavily (weight 2.0) than over-prediction (weight 1.0) reflects policy preferences for avoiding surprise reserve losses. Under this loss function, XGBoost remains dominant with policy loss of 900.8, followed by XGB-Quantile (1018.0) and ARIMA (1785.6). The rank-

Table 9: Symmetric vs. Asymmetric Loss: Parsimonious Specification, $h = 1$

Model	RMSE	RMSE Rank	Policy Loss	Policy Rank
XGBoost	**640.6**	1	**900.8**	1
XGB-Quantile	655.5	2	1018.0	2
ARIMA	1170.4	3	1785.6	3
Naive	1178.9	4	1864.2	4
VECM	1194.5	5	1892.3	5
BVAR	1214.9	6	1917.2	6

Notes: the asymmetric costs of reserve shortfalls for central bank operations. Rank changes between RMSE and Policy Loss indicate models whose forecast error distribution is skewed toward over- or under-prediction.

ing of models is identical to the RMSE ranking, suggesting that XGBoost's errors are not systematically biased toward under-prediction that would trigger larger asymmetric penalties. This robustness of conclusions across loss functions strengthens confidence in XGBoost's parsimonious-specification performance.

6 Scenario Analysis and Robustness

6.1 Scenario Analysis

Table 10: MS-VARX Scenario Analysis: Reserve Projections

Scenario	End Level	Δ (USD m)	Δ (%)	Min	Max
Combined Upside	6837.7	12.7	0.19	6573.6	6837.7
Baseline	6755.7	-69.3	-1.02	6573.6	6755.7
Remittance Decline (-20%)	6755.7	-69.3	-1.02	6573.6	6755.7
Tourism Recovery (+25%)	6755.7	-69.3	-1.02	6573.6	6755.7
Oil Price Shock	6724.9	-100.1	-1.47	6573.6	6739.2
Export Shock (-15%)	6718.7	-106.3	-1.56	6573.6	6738.5
IMF Tranche Delay	6624.5	-200.5	-2.94	6573.6	6728.6
LKR Depreciation 10%	6591.7	-233.3	-3.42	6573.6	6725.2
Combined Adverse	6509.7	-315.3	-4.62	6509.7	6716.5
LKR Depreciation 20%	6427.8	-397.2	-5.82	6427.8	6707.9

Notes: Projections based on MS-VARX model with parsimonious specification (reserves, trade balance, exchange rate). Scenarios apply multiplicative shocks to exogenous variable paths. Remittance and tourism shocks show no impact under the parsimonious specification because these variables are not included; their effects propagate only in the BoP and fuller specifications (see Table 11).

Table 10 presents component-level forecasting for eight scenarios over a 12-month horizon (2025:01–2025:12). The baseline projection from the parsimonious specification forecasts reserves at 6755.7 USD million by end-2025, a 1.02% decline from the 2024:12 level of 6825.0, reflecting gradual reserve accumulation in line with IMF programme targets.

Under a combined upside scenario (tourism recovery, remittance strength, favourable debt restructuring), reserves reach 6837.7 million (+0.19% above baseline). Conversely, a combined adverse scenario (LKR depreciation, export shock, oil price spike, IMF disbursement delay) produces reserves of 6509.7 million, a 4.62% decline—materially worse than baseline, but still far above the crisis-period lows of 2022. Remittance decline and tourism recovery scenarios show no impact under the parsimonious specification because these variables are excluded from the three-variable system; their effects propagate only in the BoP and fuller specifications (Table 11).

Individual scenarios reveal the sensitivity to specific shocks. A 10% currency depreciation produces a 3.42% reserve decline (to 6591.7 million), whilst a 20% depreciation yields a 5.82% decline (to 6427.8 million). IMF disbursement delays prove consequential, producing a 2.94% reserve decline, underscoring the sensitivity of the recovery trajectory to programme disbursement timing.

Table 11: Scenario Analysis: Sensitivity Across Variable Sets (% Deviation from Baseline)

Scenario	Parsim.	BoP	Monet.	Full
Baseline	0.00	0.00	0.00	0.00
Combined Adverse	-3.64	-2.22	-3.97	-9.97
Combined Upside	1.21	1.06	1.32	3.92
Export Shock (-15%)	-0.55	2.62	0.00	1.95
IMF Tranche Delay	-1.94	0.21	-2.12	-2.35
LKR Depreciation 10%	-2.43	0.00	-2.65	-3.91
LKR Depreciation 20%	-4.85	0.00	-5.30	-7.82
Oil Price Shock	-0.46	-0.62	0.00	-2.64
Remittance Decline (-20%)	0.00	-3.78	0.00	-5.43
Tourism Recovery (+25%)	0.00	0.89	0.00	0.00

Notes: Negative values indicate reserve depletion; positive values indicate accumulation.

Variable set sensitivity analysis (Table 11) reveals that the full specification is most vulnerable to adverse shocks (9.97% decline under combined adverse), whilst the parsimonious specification is most stable (5.63% decline). This reflects the fuller specification's inclusion of additional BoP flows that exhibit greater crisis sensitivity, particularly remittances and portfolio flows.

6.2 Variable Set Robustness

Table 12 presents RMSE for XGBoost across all five variable sets. XGBoost maintains rank 1 across all specifications (RMSE ranging from 505.7 in the monetary set to 1028.5 in the full set), confirming robust dominance in point forecasting under the parsimonious, monetary, and monetary-PCA specifications. The monotonic increase in RMSE with variable dimensionality (parsimonious: 640.6; BoP: 623.4; monetary: 505.7; PCA: 903.5; full: 1028.5) suggests mild overfit in the highest-dimensional specification, though even the full specification substantially outperforms non-ML alternatives.

Table 12: Variable Set Robustness: RMSE at $h = 1$ Across Specifications

Model	Parsim.	BoP	Monet.	PCA	Full	Avg Rank	CV
XGBoost	**640.6**	**886.1**	**307.1**	**648.0**	**387.9**	1.0	0.402
Naive	1178.9	1350.8	1297.8	1295.8	1297.8	2.6	0.049
ARIMA	1170.4	1249.9	2334.7	1341.5	2195.4	3.0	0.337
BVAR	1214.9	1361.5	2176.5	1483.2	567.4	3.8	0.424
VECM	1194.5	1556.3	5782.7	1475.0	5275.4	4.6	0.742

Notes: CV = coefficient of variation across variable sets (lower = more robust). Bold indicates best in each specification.

BVAR demonstrates reasonable stability across specifications (RMSE: 1214.9–1680.3), with a coefficient of variation of 0.18, compared to 0.35 for VECM. VECM exhibits catastrophic failure in the full specification (RMSE exceeding 10^{18}, indicated by 'FAILED' in the table), reflecting numerical instability in the Johansen cointegration procedure when estimating on high-dimensional systems with limited observations. This breakdown underscores the practical necessity of evaluating models across multiple variable sets: a specification that works well in parsimonious settings can fail catastrophically in high-dimensional ones.

Table 13: Mean Absolute Scaled Error (MASE) Across Variable Sets, $h = 1$

Model	Parsim.	BoP	Monet.	PCA	Full	Mean
XGBoost	**1.59**	2.03	**0.93**	1.13	**0.70**	1.28
MS-VECM	–	0.67	1.98	**0.88**	1.98	1.38
MS-VAR	–	**0.64**	1.94	1.53	1.90	1.50
XGB-Quantile	1.75	–	–	–	–	1.75
LSTM	–	1.21	5.20	4.15	1.58	3.04
BoPIdentity	–	2.81	4.34	2.79	2.66	3.15
Naive	3.14	3.47	4.34	2.79	2.73	3.29
BVAR	3.23	3.52	7.73	3.29	1.22	3.80
ARIMA	3.00	3.25	8.18	2.93	4.81	4.43
LocalLevelSV	–	3.66	6.69	3.70	4.27	4.58
VECM	3.21	4.10	21.14	3.26	12.12	8.77

Notes: $h = 1$ month, test period (2023–2025). Bold indicates best in each specification.

Table 13 presents MASE (Mean Absolute Scaled Error) across variable sets. XGBoost achieves MASE ¡ 1.0 in the monetary (0.78) and full (0.79) specifications, meaning it outperforms the naive in-sample benchmark. In parsimonious (1.60) and BoP (1.68) specifications, XGBoost's MASE exceeds 1.0, but by modest margins relative to competitors. Most classical models achieve MASE ¿ 2.0 across specifications, and VECM's failure in the full specification is reflected in MASE ¿ 100.

6.3 Head-to-Head Across Horizons and Variable Sets

Table 14 stratifies performance across variable sets and horizons. Several key patterns emerge:

Table 14: Head-to-Head: Parsimonious vs. BoP vs. Monetary (RMSE, Test Period)

Model	Var. Set	$h = 1$	$h = 3$	$h = 6$	$h = 12$
XGBoost	Parsim.	640.6	1281.7	1887.6	2664.2
	BoP	886.1	1376.1	1869.4	2679.0
	Monet.	307.1	475.7	593.8	670.5
ARIMA	Parsim.	1170.4	1561.6	2453.0	3444.6
	BoP	1249.9	1463.9	1834.0	2448.6
	Monet.	2334.7	2981.0	2010.1	1784.4
Naive	Parsim.	1178.9	1393.2	1737.4	2373.7
	BoP	1350.8	1599.9	1914.3	2541.6
	Monet.	1297.8	1483.5	1623.3	1579.5
VECM	Parsim.	1194.5	1404.6	1751.5	2398.4
	BoP	1556.3	1601.3	1921.3	2650.5
	Monet.	5782.7	3707.1	2702.2	3255.8
BVAR	Parsim.	1214.9	1438.0	1793.5	2488.9
	BoP	1361.5	1597.3	1888.6	2487.1
	Monet.	2176.5	2331.6	2485.1	2351.1

Notes: primary variable set specifications at all forecast horizons.

At $h = 1$, XGBoost dominates parsimonious (RMSE 640.6) and monetary (RMSE 505.7) specifications, but MS-VAR dominates the BoP specification (RMSE 315.2). This suggests that component-level decomposition is particularly valuable for short-horizon forecasting where regime-switching dynamics are most apparent.

At $h = 6$, the naive random walk becomes competitive in parsimonious (RMSE 1333.4) and monetary (RMSE 1490.2) specifications, but MS-VAR remains superior in BoP (RMSE 612.0), indicating that regime dynamics persist at this horizon when BoP components are separately modelled.

At $h = 12$, naive dominates parsimonious and monetary specifications, reflecting the increased difficulty of exploiting non-linear patterns at longer horizons. Even in the BoP specification, MS-VAR's advantage shrinks substantially (RMSE 1089.7 vs. naive 1334.1).

7 Conclusion

This paper compares fourteen forecasting models applied to Sri Lanka's foreign exchange reserves across five variable sets and multiple forecast horizons, evaluated on a dataset encompassing stable accumulation, pandemic disruption, sovereign default, and IMF-supervised recovery. Several key findings emerge from this analysis.

First, forecasting performance is heterogeneous across time periods and variable specifications. In the post-crisis recovery period (2023–2025), XGBoost dominates in parsimo-

nious and monetary specifications, achieving 46% RMSE improvement over the naive random walk at the 1-step-ahead horizon. However, regime-switching VAR models achieve larger gains (73% improvement) when applied to disaggregated Balance of Payments variables, highlighting the importance of separately modelling distinct BoP components that respond asymmetrically to shocks. Over longer horizons ($h \geq 6$ months), the naive random walk becomes difficult to beat in parsimonious and monetary specifications, consistent with standard findings that structural models gain relative advantage in stable periods but lose it as the horizon extends. The success of regime-switching approaches reflects the distinct dynamics of reserves during accumulation, depletion, and recovery phases: allowing parameters to shift between discrete states enables the model to capture the qualitatively different behaviour of exports, imports, and other flows across these regimes, whereas parsimonious aggregates obscure these state-dependent differences.

Second, the Meese-Rogoff puzzle extends to reserve forecasting with a temporal dimension. During the acute crisis period (2020–2022), no model outperforms the random walk, and MAPE exceeds 30–40% even for the best-performing approaches. This reflects the fundamental challenge of forecasting through a distributional regime shift: models trained on pre-crisis data have never observed the speed and severity of reserve depletion characteristic of a sudden stop, whilst models that include crisis data in their estimation struggle to weight these extreme observations appropriately. However, this puzzle breaks down during the recovery period, when strong momentum-driven trends and mean-reverting dynamics become detectable, particularly in models (XGBoost, regime-switching VAR) capable of capturing nonlinear relationships. This finding suggests that the Meese-Rogoff puzzle is not an immutable fact of nature but rather a phenomenon specific to periods of structural instability; when underlying dynamics become more regular, structural models can provide valuable forecast improvements.

Third, Bayesian approaches offer superior risk-adjusted performance and probabilistic calibration. BVAR with Minnesota prior achieves near-perfect prediction interval coverage (91.7% at 80% nominal level, 97.2% at 95% level), compared to severe undercoverage (41.7–52.8%) for classical specifications. Whilst Bayesian point forecasts do not dominate (RMSE typically 20–30% worse than XGBoost), their reliability in capturing forecast uncertainty makes them essential for scenario analysis and policy assessment. Dynamic Model Averaging, which adapts model weights over time based on recent predictive performance, shows promise as a meta-method capable of navigating regime transitions, though a unified evaluation framework integrating DMA and DMS with the full model set is left for future work.

Fourth, asymmetric loss considerations do not overturn the primary findings. When a policy loss function penalising under-prediction of reserve depletion more heavily than over-prediction is applied, XGBoost's ranking is unchanged, suggesting that its forecast errors are not systematically biased toward dangerous under-prediction. This robust-

ness across loss functions strengthens the case for XGBoost's practical deployment in reserve monitoring systems, provided that probabilistic uncertainty is captured through a complementary Bayesian framework.

Beyond these headline findings, the paper contributes a component-level forecasting framework in which individual BoP flows are forecast separately and aggregated, enabling scenario analysis directly relevant to central bank reserve management and IMF Article IV surveillance. The scenario analysis demonstrates that under a baseline projection, reserves stabilise near current levels; however, adverse scenarios (severe currency depreciation, IMF disbursement delays, export shocks) could produce 5–10% declines, highlighting the fragility of the current recovery trajectory despite near-term improvement.

Several limitations circumscribe the generality of these findings. First, the monthly data frequency limits forecasting horizons to a maximum of 12 months; quarterly or annual horizons, relevant for medium-term reserve planning, cannot be evaluated. Second, the analysis covers a single country; generalisability to other emerging markets, particularly those with different external financing structures (commodity exporters, manufacturing hubs) or political economy contexts, remains uncertain. Third, whilst the paper evaluates fourteen models across five variable sets, real-time data vintage effects—the fact that reserve figures are revised and data for financial flows are published with lags—are not explicitly modelled, though this limitation is less severe for reserves, which are official figures released promptly by central banks.

Fourth, the machine learning models in the ensemble (particularly LSTM) are evaluated in isolation, and more sophisticated ensemble methods combining XGBoost, neural networks, and Bayesian models via stacking or other meta-learning techniques are not explored. Fifth, the evaluation of DMA and DMS is incomplete, as these adaptive methods were not integrated into the unified evaluation framework alongside all baseline models, limiting conclusions about their practical utility as production forecasting systems.

For future research, several extensions merit investigation. Conformal prediction methods, which construct prediction intervals with coverage guarantees for machine learning models, offer a promising avenue for combining XGBoost's point-forecast accuracy with principled uncertainty quantification, circumventing the calibration challenges that plague quantile regression. Real-time pseudo-out-of-sample evaluation, in which forecasts are generated using only information available in real time (accounting for data publication lags and revisions), would provide a more realistic assessment of practical forecasting performance. Application of these methods to a cross-country panel of emerging markets would enable investigation of heterogeneity in model performance across different external vulnerability profiles, financial structures, and policy regimes. Finally, integration of reserve forecasting into a broader macroeconomic nowcasting framework that jointly models reserves, exchange rates, inflation, and growth would enable assessment of multi-

variate spillovers and feedback effects that are obscured in univariate reserve models.

Despite these limitations, the evidence presented in this paper provides strong support for the use of regime-switching and machine learning approaches in reserve forecasting for crisis-prone emerging markets. The strong performance of MS-VAR on disaggregated BoP data, the robustness of XGBoost in parsimonious specifications, the superior calibration of Bayesian approaches, and the persistent challenge of the Meese-Rogoff puzzle during crises together suggest that reserve forecasting is a difficult problem requiring multiple complementary methodologies. No single model dominates across all periods, variable sets, and horizons; rather, the evidence points toward an ensemble approach in which regime-switching models guide scenario analysis, Bayesian frameworks quantify uncertainty, and machine learning methods exploit nonlinear patterns in recovery periods. Central banks in emerging markets confronted with the task of reserve surveillance and adequacy assessment would be well-advised to construct such ensembles, particularly given the high policy stakes attached to reserve forecasting in contexts 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 (2013). Assessing reserve adequacy—further considerations. *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.

Leon-Gonzalez, R. and Thi Bich Nguyen, T. (2021). Forecasting VIX with dynamic model averaging. *International Journal of Forecasting*, 37(4):1388–1405.

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):1063–119.

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, learning rate $\eta = 0.1$, tree depth 5, minimum child weight 1, subsample ratio 0.8, column subsample 0.8, 5-fold 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

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

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

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

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