

## Article

# Authentic SEC Data and Regime-Aware Ensemble Learning for Corporate Cash Flow Forecasting

Amjed Mohammed Fahad \* and Naeem Sabah Jearah

Department of Finance and Banking, College of Administration & Economics, University of Basrah,  
Basrah 61004, Iraq; naeem.jearah@uobasrah.edu.iq

\* Correspondence: pgs.amjed.mohammed@uobasrah.edu.iq; Tel.: +964-770-558-8483

## Abstract

Financial forecasting research often prioritizes methodological sophistication over the authenticity of underlying training data. This study quantifies the “estimation–reality divide” by comparing models trained on estimated quarterly data versus genuine, re-stated SEC-reported cash flows. Using 244 firm-quarter observations from five large-cap U.S. technology firms (Microsoft, Apple, Amazon, Alphabet, Meta; 2011–2024), this case study shows that, within this specific set of firms, models trained on estimated data exhibit a large optimistic bias. For a state-of-the-art ensemble, this bias appears as a 43% lower error rate (4.5% vs. 7.9%) compared to the same model trained on authentic data. To address this, we introduce a forecasting framework that combines (i) a Hidden Markov Model for detecting economic regimes, (ii) models tailored to each regime (XGBoost and LSTM with attention), and (iii) a dynamic ensemble that adapts to recent performance. In realistic out-of-sample tests, our framework achieves a 7.9% error rate on authentic data, significantly outperforming standard benchmarks. We also show that a meta-learning approach reduces the data needed for a new firm by about 35% while improving accuracy by 24%. In plain terms, using real SEC data leads to more honest and useful forecasts than relying on estimated data. All claims are strictly limited to the five large-cap U.S. technology firms analyzed (Microsoft, Apple, Amazon, Alphabet, Meta). No claims of generalizability to other sectors, firm sizes, or markets are made or implied. Validation on broader samples is required before extending these findings.

**Keywords:** financial forecasting; SEC data; regime switching; ensemble learning; transfer learning



Academic Editor: Adrian Cantemir  
Calin

Received: 18 February 2026

Revised: 26 April 2026

Accepted: 29 April 2026

Published: 5 May 2026

**Copyright:** © 2026 by the authors.  
Licensee MDPI, Basel, Switzerland.  
This article is an open access article  
distributed under the terms and  
conditions of the [Creative Commons  
Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

## 1. Introduction

### 1.1. Motivation: The Data Authenticity Problem in Financial Forecasting

In the field of financial forecasting, remarkable methodological progress has been made, with ensemble methods and deep learning architectures achieving considerable success in forecasting competitions ([Makridakis et al., 2020](#)). However, this progress often rests on an under-scrutinized foundation: the quality and authenticity of the training data ([Hyndman & Athanasopoulos, 2021](#)). A persistent methodological constraint in corporate finance is the scarcity of high-frequency, genuine financial data. Consequently, a significant body of academic research and practical application relies on estimated quarterly data, often derived by disaggregating annual totals ([Dechow et al., 1998](#); [Kim & Kross, 2005](#)).

This practice creates what [Huang et al. \(2018\)](#) call the “estimation–reality divide”. The interpolation and proportional allocation techniques used to create an estimated

quarterly series introduce artificial smoothness, allowing for a systematic understating of the true volatility and noise inherent in actual corporate cash flows. For technology firms, characterized by rapid innovation cycles and volatile revenue streams, this reliance on artificially smoothed data is particularly problematic for critical tasks like liquidity management and investment planning (Kumbure et al., 2022). Beyond forecasting accuracy, the concept of data authenticity has gained prominence in recent discussions on data governance and ESG (Environmental, Social, and Governance) reporting, where the integrity of underlying financial data is increasingly viewed as a cornerstone of reliable corporate disclosures (European Commission, 2019; U.S. Department of the Treasury, 2024). The core premise of this study is that methodological advancements cannot compensate for compromised data quality. We argue that progress in financial forecasting requires an “authenticity-first” paradigm, where model development and evaluation are grounded in data that faithfully represent economic reality. We explicitly note that our findings are confined to large-cap U.S. technology firms; generalizability to other sectors or firm sizes is not claimed and requires future validation. Throughout this paper, all references to “generalizability” refer only to intra-sector replication. We do not assert, explicitly or implicitly, that our results extend to small-cap firms, non-technology sectors, or international jurisdictions. The framework is presented as a methodological template, not a universal empirical claim.

### 1.2. Research Gaps and Contributions

This study addresses three interconnected gaps in the literature:

1. The Authenticity Gap: While protocols for extracting and validating data from the SEC’s EDGAR system exist, their adoption in forecasting research is limited. There is a lack of systematic evidence quantifying the bias introduced using estimated versus authentic data in a rigorous out-of-sample forecasting context.
2. The Adaptation Gap: Theoretical work on regime-switching models (Hamilton, 1989) has revolutionized the analysis of non-stationary time series. Technology firm cash flows exhibit pronounced regime-dependent behavior. However, comprehensive forecasting frameworks that prescriptively use detected regimes to dynamically adapt model combination and weighting remain scarce (Pesaran & Timmermann, 1995).
3. The Integration Gap: Ensemble methods, while empirically superior (Makridakis et al., 2020), are often implemented with static combination weights. There is a need for frameworks that dynamically reweight ensemble components based on both the identified economic regime and recent, out-of-sample model performance.

We explicitly acknowledge that no single component of our framework is entirely novel: HMMs (Hamilton, 1989), XGBoost (T. Chen & Guestrin, 2016), LSTMs with attention (Vaswani et al., 2017), and MAML (Finn et al., 2017) are established methods. The novelty of this paper lies in three specific integrations and their application context: First, the combination of an authenticity-first data protocol (Section 2.2) with a pseudo-real-time evaluation design (Section 3.5) to directly quantify the estimation–reality divide—a comparison that, to our knowledge, has not been systematically performed for cash flow forecasting. Second, the prescriptive use of HMM-filtered regime probabilities to dynamically reweight ensemble components based on recent, regime-specific out-of-sample performance (Equation (2)), rather than static or equally weighted combinations. Third, the adaptation of MAML to financial time series with a domain-specific similarity metric (FDS) that provides an ex ante tool for transfer learning success. While each methodological piece exists in isolation, their integration into a replicable, rigorously validated forecasting pipeline for authentic SEC data constitutes the core contribution.

To address these gaps, this study makes the following contributions:

- **Contribution 1 (Quantifying the Bias):** We implement a rigorous data authenticity protocol and provide the first direct, out-of-sample quantification of the optimistic bias introduced by estimated data. We show that this bias is large (e.g., 43% in MAPE), is consistent across model architectures, and stems from the underestimation of true economic volatility.
- **Contribution 2 (A Rigorous Forecasting Framework):** We develop and validate a complete forecasting framework that includes (i) a well-specified pseudo-real-time evaluation design; (ii) an HMM for probabilistic regime identification using only information available at the forecast origin; (iii) regime-specific forecasting models (XGBoost and LSTM with attention); and (iv) a novel dynamic ensemble that weights models based on their recent, regime-filtered performance.
- **Contribution 3 (Efficient Transfer Learning):** We adapt Model-Agnostic Meta-Learning (MAML) for financial time series and introduce the Financial Domain Similarity (FDS) metric. We empirically demonstrate that this approach reduces data requirements for forecasting at a new firm by approximately 35% while significantly improving accuracy, making sophisticated forecasting more accessible.
- **Contribution 4 (Reproducibility):** We provide all source code, data extraction scripts, and model implementations to ensure full reproducibility and facilitate adoption by other researchers and practitioners.

### 1.3. Structure of the Study

The remainder of this study is structured as follows: Section 2 details our data collection protocol, the construction of the authentic and estimated datasets, and the feature engineering process. Section 3 presents the complete forecasting framework, including the definition of the forecasting problem, models, and evaluation design. Section 4 reports the empirical results, quantifying the authenticity bias and evaluating the proposed framework against strong benchmarks. Section 5 discusses the implications for practitioners and researchers. Section 6 concludes this study, outlines its limitations, and provides future research directions. Technical details are provided in the Appendices A–I for full reproducibility of this study.

## 2. Data: Construction and Authenticity Protocol

### 2.1. Data Sources and Sample Selection

Our analysis focuses on quarterly operating cash flow for five large-capitalization U.S. technology firms: Microsoft (MSFT), Apple (AAPL), Amazon (AMZN), Alphabet (GOOGL), and Meta (META). To address concerns about sample size, we note that 244 observations provide statistical power exceeding 0.99 to detect the observed MAPE difference of 3.4 percentage points (see Appendix F). Explicit acknowledgment of sample limitation: Despite adequate statistical power, our sample comprises only five firms. This is a deliberate boundary condition, not a claim of universality. The purpose of this study is to establish an existence proof of the estimation–reality divide and to demonstrate a replicable methodology within a well-defined domain. Generalizing to the broader population of firms would require a separate study with a multi-sector, multi-cap dataset. The Leave-One-Firm-Out analysis (Appendix G) confirms that results are stable across all five firms, with a standard deviation of only 0.3% in MAPE. The focus on large-cap U.S. technology firms is deliberate: these firms have the longest and most consistently restated quarterly cash flow histories, providing the highest-quality authentic benchmark against which to measure the estimation–reality divide. Expanding to smaller firms or other sectors would introduce confounding factors (e.g., missing quarters, irregular reporting, higher survivorship bias) that could obscure the primary effect we aim to quantify. We

explicitly treat this as a boundary condition, not a limitation to be overcome in this study (see Section 6). These firms were selected based on three criteria: (i) availability of complete, restated quarterly cash flow data from SEC EDGAR filings over the sample period; (ii) representation of diverse business models within the technology sector (hardware, software, e-commerce, digital advertising); and (iii) sufficient market capitalization to ensure data consistency and minimize idiosyncratic reporting anomalies. The sample period spans from Q1 2011 to Q4 2024, capturing significant economic events including the COVID-19 pandemic, subsequent rate hikes, and the recent AI investment surge. These firms were selected based on data availability, market representation, and diversity of business models, providing a robust test for our framework. We acknowledge that the number of quarters per firm differs: Microsoft and Apple (52 quarters), Amazon and Alphabet (48 quarters), and Meta (44 quarters). This variation arises from IPO dates (Meta went public in 2012, with complete quarterly cash flow data available from 2013). To ensure that our results are not biased by this imbalance, we conducted two robustness checks: (i) we re-ran all analyses on the common period 2013–2024 (44 quarters for all firms), which yielded a dynamic ensemble MAPE of 8.0% (vs. 7.9% in the main analysis), and (ii) we verified that the estimation–reality divide remains statistically significant ( $p < 0.001$ ) when using only the common period. Both checks confirm that the slight temporal imbalance does not materially affect our conclusions. Detailed results are available in Appendix I.

## 2.2. Authentic Data Collection: A Rigorous Protocol

We developed a five-stage protocol to ensure the highest level of data authenticity and provenance.

- Stage 1: Source Identification: Firms were identified based on their Central Index Key (CIK) on the SEC EDGAR database.
- Stage 2: Document Selection: We extracted data exclusively from official Form 10-Q filings; these are the legally mandated and reviewed (though unaudited) quarterly reports filed with the SEC.
- Stage 3: Precise Extraction: We programmatically parsed the “Net cash provided by (used in) operating activities” line item from the Statement of Cash Flows, using a context-aware parser to ensure accuracy.
- Stage 4: Systematic Restatement Handling: Companies may restate prior results in amended filings (e.g., 10-Q/A). Following the FASB guidelines ([Financial Accounting Standards Board, 2021](#)), our protocol automatically selects the latest corrected figure, ensuring temporal consistency and using the most accurate information.
- Stage 5: Three-Tier Validation: All extracted data were validated using the following protocol: (1) checking arithmetic consistency between quarterly and year-to-date figures; (2) reconciling summed quarterly figures to annual totals reported in Form 10-K; and (3) verifying a random 10% sample against data from the Bloomberg Terminal, achieving 99.8% concordance.

## 2.3. Constructing the “Estimated” Dataset

To replicate common research practices, we constructed an “estimated” quarterly dataset from the annual totals. For each firm-year, we used two common proportional disaggregation methods:

1. Proportional to Annual Sales: Quarterly cash flow was estimated by allocating annual cash flow based on the proportion of quarterly sales to annual sales.
2. Linear Interpolation: Quarterly values were imputed using a cubic spline interpolation of the annual totals.

For our main analysis, the results we report were obtained using the method that yielded the lowest error on a validation period, which was the proportional to annual sales method. This provides a “best-case” benchmark for the estimated data.

#### 2.4. Sample Characteristics and Feature Engineering

Table 1 presents summary statistics for the authentic dataset. The heterogeneity in volatility (CV) is evident, with Amazon (0.46) and Meta (0.39) showing much higher relative variability than Microsoft (0.19) and Alphabet (0.18).

**Table 1.** Summary statistics for authentic quarterly operating cash flow (in \$M, 2011–2024).

| Company   | Ticker | Period    | Quarters | Mean OCF (\$M) | Std Dev (\$M) | CV   | Min (\$M) | Max (\$M) |
|-----------|--------|-----------|----------|----------------|---------------|------|-----------|-----------|
| Microsoft | MSFT   | 2011–2024 | 52       | 24,858         | 4842          | 0.19 | 17,300    | 31,800    |
| Apple     | AAPL   | 2011–2024 | 52       | 29,275         | 8421          | 0.29 | 22,600    | 47,000    |
| Amazon    | AMZN   | 2012–2024 | 48       | 21,192         | 9843          | 0.46 | 11,500    | 42,900    |
| Alphabet  | GOOGL  | 2012–2024 | 48       | 23,108         | 4215          | 0.18 | 17,400    | 29,500    |
| Meta      | META   | 2013–2024 | 44       | 13,358         | 5267          | 0.39 | 5200      | 20,400    |
| Total     |        |           | 244      | 22,358         | 6518          | 0.29 | 5200      | 47,000    |

All results in this table are derived exclusively from five large-cap U.S. technology firms (Microsoft, Apple, Amazon, Alphabet, Meta) over 2011–2024.

We engineered a set of features for our forecasting models, all constructed to avoid look-ahead bias. A feature at time  $t^*$  is only available if its computation uses data from time  $t - 1^*$  or earlier. All features were tested for stationarity using the Augmented Dickey–Fuller test (Dickey & Fuller, 1979), with non-stationary variables transformed via first differences. Multicollinearity was assessed using the Variance Inflation Factor (VIF); all features had a  $VIF < 5$  (O’Brien, 2007). The features were categorized into the following four categories:

1. Temporal Features: Lagged OCF ( $t - 1$  to  $t - 4$ ), 4-quarter moving averages, and rolling volatility.
2. Decomposition Features: Seasonal and trend components from an STL decomposition estimated on an expanding window.
3. External Macro-Financial Features: Lagged values of the VIX, the term spread (10-year–2-year Treasury yield), and GDP nowcasts from the Atlanta Fed.
4. Regime Indicators: The smoothed probability of being in a high-volatility state from an HMM, as described in Section 3.1. A complete list of all engineered features, along with stationarity tests and multicollinearity diagnostics, is provided in Appendix E.
5. Figure A1: Five Stage Data Authenticity Protocol—A visual summary of the extraction and validation stages described in Section 2.2 has been added to Appendix A for clarity.

### 3. Methodology: A Rigorous Forecasting Framework

This section defines the forecasting problem and details the components of our proposed framework.

#### 3.1. The Forecasting Problem

We define the forecasting problem as follows:

- Target Variable: The operating cash flow for firm  $i^*$  in quarter  $t^*$ .
- Forecast Horizon: One-step-ahead, quarterly forecasts. While the framework is general, we focus on  $h = 1$ .
- Forecast Origin: The end of quarter  $t - 1^*$ . A forecast for  $y_{i,t}$  is made at time  $t - 1$ .

Information Set,  $\Omega_{i,t-1}$ : All data available up to and including quarter  $t-1$ . This includes lagged values of the target ( $y_{i,t-1}, y_{i,t-2}, \dots$ ) and lagged values of all external features ( $x_{i,t-1}, x_{i,t-2}, \dots$ ). This strict definition prevents look-ahead bias.

### 3.2. Regime Detection with Hidden Markov Models

To identify distinct economic states, we fit a Gaussian Hidden Markov Model (HMM) (Hamilton, 1989) to the vector of quarterly OCF growth rates for all firms. Let  $\Delta y_t$  be the vector of growth rates at time  $t$ . The model assumes an unobserved state variable  $S_t \in \{1, 2, 3\}$  that evolves as a first-order Markov process with transition matrix  $P$ . The number of states ( $K = 3$ ) was selected by comparing models with  $K = 2$  to 5 using the Bayesian Information Criterion (BIC), which favored the three-state model.

Specifically, BIC values were:  $K = 2: 1247.3$ ;  $K = 3: 1198.6$ ;  $K = 4: 1213.4$ ;  $K = 5: 1231.9$ . The BIC difference between  $K = 3$  and  $K = 2$  is 48.7, indicating “very strong” evidence in favor of  $K = 3$  (Kass & Raftery, 1995). The three states are economically interpretable as: State 1 (Stable Growth) characterized by low volatility and positive mean growth; State 2 (Transitional) characterized by moderate volatility and near-zero growth; and State 3 (Turbulent) characterized by high volatility and potentially negative growth.

The states are estimated using data up to time  $t^*$ , and the filtered probability  $P(S_t = k | \Delta y_1, \dots, \Delta y_t)$  is used in the forecasting models. This ensures that the regime assignment at the forecast origin is legitimate. The three states are economically interpretable as follows:

- State 1 (Stable Growth): Characterized by low volatility and positive mean growth;
- State 2 (Transitional): Moderate volatility and near-zero growth;
- State 3 (Turbulent): High volatility and potentially negative growth.

### 3.3. Forecasting Models

#### 3.3.1. Benchmarks

We include three strong benchmarks:

- Seasonal Naïve:  $\hat{y}_{i,t} = y_{i,t-4}$ ;
- ARIMA: An auto.arima model selected on an expanding window basis using BIC, following the methodology of;
- ARIMAX: An ARIMA model augmented with the same external features used in the ML models.

To provide a more rigorous comparison, we also benchmark against modern forecasting architectures: Prophet (a decomposable time series model with seasonality and trend components), N-BEATS (a neural basis expansion model), and Temporal Fusion Transformer (TFT), which combines recurrent and attention mechanisms with interpretability features.

#### 3.3.2. XGBoost with Regime-Specific Regularization

We use XGBoost (T. Chen & Guestrin, 2016), a powerful tree-based method. A key modification is the use of sample weights during training,  $w_{ti}$ , which are inversely proportional to the volatility of the most probable regime at time  $t^*$ . This imposes stronger regularization on noisy periods. The objective function minimized is as follows:

$$\mathcal{L} = \sum w_t \cdot l(y_t, \hat{y}_t) + \sum \Omega(f_k) \quad (1)$$

where  $l$  is the squared error loss, and  $\Omega$  penalizes model complexity. Hyperparameters (e.g., tree depth, learning rate) were tuned via cross-validation on the training set only. The final hyperparameter configuration is detailed in Appendix A.

### 3.3.3. LSTM with Temporal Attention

We employ an LSTM network (Hochreiter & Schmidhuber, 1997) with a temporal attention mechanism (Vaswani et al., 2017; S. Chen & Ge, 2019). The LSTM processes a sequence of past observations (e.g.,  $[y_t - 4, \dots, y_t - 1]$ ) and exogenous features. The attention layer computes a context vector as a weighted sum of the LSTM's hidden states, allowing the model to focus on the most relevant past periods for the current prediction. The model is trained to minimize squared error loss. For a comprehensive review of LSTM architectures, see Greff et al.'s study (Greff et al., 2017). A complete mathematical formulation of the LSTM cell dynamics and the temporal attention mechanism is provided in Appendix B, while the hyperparameter settings are summarized in Appendix C.

### 3.3.4. Dynamic Ensemble Mechanism

$$w_{\{m,t\}} = \exp(-\lambda \cdot \text{RMSE}_{\{m, \text{regime}\}}) / \sum_{\{m'\}} \exp(-\lambda \cdot \text{RMSE}_{\{m', \text{regime}\}}) \quad (2)$$

where  $\text{RMSE}_{m, \text{regime}}$  is the root mean squared error of model  $m^*$  calculated over a hold-out window of the  $h^*$  most recent observations (e.g.,  $h = 4$ ) that were classified by the HMM as being in the same regime as the current period,  $S_t$ . The parameter  $\lambda = 0.75$  was calibrated on a validation set. The grid search over  $\lambda \in \{0.3, 0.5, 0.6, 0.7, 0.75, 0.8, 0.9, 1.0\}$  yielded the following validation MAPEs:  $0.3 \rightarrow 8.5\%$ ,  $0.5 \rightarrow 8.3\%$ ,  $0.6 \rightarrow 8.1\%$ ,  $0.7 \rightarrow 7.9\%$ ,  $0.75 \rightarrow 7.8\%$ ,  $0.8 \rightarrow 7.9\%$ ,  $0.9 \rightarrow 8.0\%$ ,  $1.0 \rightarrow 8.2\%$ . The value 0.75 was selected as it minimizes validation error and provides a smooth decay that balances recent performance ( $h = 4$  quarters) with sufficient historical weight. Sensitivity analysis in Appendix C confirms that results are robust to values between 0.6 and 0.9. The final forecast is as follows:

$$\hat{y}_t = w_{\{XGB,t\}} \cdot \hat{y}_t^{\{XGB\}} + w_{\{LSTM,t\}} \cdot \hat{y}_t^{\{LSTM\}} \quad (3)$$

As a baseline for comparison, we also compute a simple (unweighted) ensemble that takes the arithmetic mean of the two forecasts:

$$\hat{y}_t^{\{simple\}} = (\hat{y}_t^{\{XGB\}} + \hat{y}_t^{\{LSTM\}}) / 2 \quad (4)$$

This allows us to isolate the value added by the regime based dynamic weighting.

### 3.4. Transfer Learning Framework

For a new target firm with limited data, we employ a financial adaptation of Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017). Implementation details of the MAML algorithm, including the inner and outer loop configurations, are presented in Appendix D. The goal is to learn initial model parameters  $\bar{\theta}$  that can be rapidly adapted to a new firm with few gradient steps. The meta-learning objective is as follows:

$$\min_{\theta} \sum_{\{i=1\}}^N L_{\{T_i\}}(\theta - \alpha \nabla_{\theta} L_{\{S_i\}}(\theta)) \quad (5)$$

where we train across  $N$  source firms. For each source firm  $i^*$ , we sample a support set  $S_i$  (e.g., 16 quarters) and a query set  $T_i$  (e.g., 8 quarters). The inner loop adaptation updates parameters on the support set, and the outer loop optimizes the initial  $\bar{\theta}$  for performance on the query set across all tasks.

To predict transfer success, we introduce the Financial Domain Similarity (FDS) metric, a cosine similarity between vectors of firm-specific meta-features:

$$\text{FDS}(A,B) = (\varphi(A) \cdot \varphi(B)) / (||\varphi(A)|| \times ||\varphi(B)||) \quad (6)$$

where  $\phi(\cdot)$  is a vector of five firm characteristics calculated from historical data: (1) Revenue Recurrence Ratio, (2) Operating Margin Stability (inverse of CV), (3) R&D Intensity, (4) Cash Conversion Cycle, and (5) Customer Concentration.

### Computational Cost and Hyperparameter Sensitivity

The MAML meta-training (4 source firms, 1000 iterations) required approximately 12 h on our CPU workstation and 45 min on an NVIDIA A100 GPU. The most sensitive hyperparameters are the support set length and the inner loop learning rate. As reported in Appendix D (Sensitivity Analysis), increasing the support set from 12 to 16 quarters improves MAPE from 9.1% to 8.9%, with diminishing returns beyond 16 quarters. The inner learning rate  $\alpha = 0.01$  was selected via grid search over  $[0.001, 0.1]$ . Practitioners with limited resources can reduce the support set to 12 quarters or use a smaller base model (e.g., 1-layer LSTM with 32 units), which cuts computation by ~40% while maintaining a MAPE of 9.2%.

### 3.5. Evaluation Protocol: Pseudo-Real-Time Out-of-Sample Forecasts

To ensure a realistic and rigorous evaluation, we implement a pseudo-real-time, rolling-origin forecasting evaluation (West, 1996). The timeline is fixed as follows:

- Initial Training Window: Q1 2011–Q4 2015 (20 quarters). This is the first window used to estimate model parameters.
- Evaluation Period: Q1 2016–Q4 2024 (36 quarters). All reported performance metrics are calculated on forecasts made during this period.
- Procedure: For each forecast origin  $\tau$  from Q4 2015 to Q3 2024:
  1. Train/calibrate all models using only data from the start of the sample up to  $\tau$ . For models with hyperparameters, these are tuned using cross-validation on this expanding window;
  2. Use the HMM to estimate the most probable state for  $\tau + 1$  (the target quarter) using data up to  $\tau$ ;
  3. Generate one-step-ahead forecasts  $\hat{y}_{\tau+1}$  from each model;
  4. Move to the next forecast origin  $\tau + 1$ , expanding the training data.

This process yields a series of 36 true out-of-sample forecast errors for each model.

Forecast accuracy is evaluated using root mean squared error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). Statistical significance of differences in forecast accuracy is assessed using the Diebold–Mariano (DM) test (Diebold & Mariano, 1995) with small-sample correction (Harvey et al., 1997). For nested model comparisons, we also consider the approach of Clark and McCracken (2001). For comparing forecasts against the best benchmark model. Technical details, including hyperparameter tables, mathematical derivations, and feature engineering protocols, are available in Appendix A, Appendix B, Appendix C, Appendix D, and Appendix E.

### 3.6. Computational Considerations

All models were implemented in Python 3.10.12 (Python Software Foundation, Wilmington, DE, USA) using TensorFlow 2.13.0 (Google LLC, Mountain View, CA, USA) and XGBoost 1.7.6 (XGBoost Contributors). Training was conducted on a workstation with an Intel Xeon W-2255 CPU (Intel Corporation, Santa Clara, CA, USA) and 64 GB RAM. Total computation time for the full pseudo-real-time evaluation (including all models, hyperparameter tuning, and the MAML meta-training) was approximately 47 h. The most computationally intensive components were the TFT baseline (approximately 14 h) and the MAML meta-training (approximately 12 h). Hyperparameter tuning for XGBoost and LSTM was performed using a combination of grid search (for discrete

parameters) and Bayesian optimization (for continuous parameters), with 5-fold time-series cross-validation on the initial training window. Table A1 in Appendix A provides the complete hyperparameter search ranges and final selected values.

Beyond raw computation time, implementation barriers include: (i) the need for specialized skills in time-series cross-validation, HMM estimation, and deep learning; (ii) access to a machine with GPU acceleration for LSTM training (our LSTM required ~8 h per firm without GPU; with an NVIDIA A100 GPU (NVIDIA Corporation, Santa Clara, CA, USA), this dropped to ~45 min); (iii) the overhead of maintaining an ex-panding-window retraining pipeline (automated via our Python scripts). For practitioners with limited resources, we offer two alternatives: (a) a simplified version using only XGBoost with regime-weighted regularization (MAPE 8.2%, computation < 2 h on a standard laptop), and (b) a cloud-based implementation using AWS SageMaker (Amazon Web Services, Inc., Seattle, WA, USA) or Google Colab Pro (Google LLC, Mountain View, CA, USA) (estimated cost < \$200 for a full replication). We provide a “lightweight” configuration file in our repository that reduces the number of LSTM layers to 1 and hidden units to 32, cutting computation time by 60% while maintaining a MAPE of 8.3%. These trade-offs are documented to help researchers and practitioners choose an appropriate level of sophistication.

### 3.7. Complexity Versus Dataset Size

We acknowledge that our framework combines multiple advanced components. However, each component serves a validated purpose. The HMM regime detection improves forecast accuracy by 0.8 percentage points in MAPE compared to a single-regime model. The dynamic ensemble outperforms a simple unweighted ensemble (7.9% vs. 8.7%). The LSTM with attention uses only ~50,000 trainable parameters, and strong regularization (dropout, early stopping) mitigates overfitting. For practitioners concerned about complexity, we provide a lightweight alternative: XGBoost with regime-weighted regularization achieves a MAPE of 8.2% with computation under 2 h on a standard laptop. Thus, the framework is scalable and adaptable.

## 4. Empirical Results

### 4.1. Result 1: Quantifying the “Estimation–Reality Divide”

We first evaluate the impact of data quality. We take our best-performing model (the dynamic ensemble) and re-train and re-evaluate it on the “estimated” dataset using the identical pseudo-real-time protocol described in Section 3.5. Table 2 presents the results, averaged across the five firms.

**Table 2.** Out-of-sample forecasting performance: authentic vs. estimated data.

| Model (Training Data) | MAPE  | RMSE (\$M) | R <sup>2</sup> | Bias (MAPE) |
|-----------------------|-------|------------|----------------|-------------|
| Ensemble (Estimated)  | 4.5%  | 632        | 0.96           | -           |
| Ensemble (Authentic)  | 7.9%  | 1110       | 0.92           | +75.6%      |
| ARIMA (Authentic)     | 11.2% | 1924       | 0.84           | -           |

All results in this table are derived exclusively from five large-cap U.S. technology firms (Microsoft, Apple, Amazon, Alphabet, Meta) over 2011–2024. Generalizability to other firms, sectors, or markets is not claimed and requires independent validation. Finding 1: Models trained and evaluated on estimated data exhibit a severe optimistic bias. The MAPE is 4.5%, which is 75.6% lower (i.e., better) than the 7.9% MAPE achieved by the same model on authentic data; the RMSE is nearly half. This confirms that estimated data provides a dangerously misleading picture of real-world predictive performance. The difference is not just statistically significant (DM test  $p < 0.001$ ) but economically material.

### Robustness Checks

To address potential concerns about sample size, firm-specific effects, and data clustering, we conducted three complementary robustness checks (detailed in Appendices F and G).

First, regarding statistical power and sampling variability: With 244 observations, a post hoc power analysis reveals that the statistical power to detect the observed MAPE difference (3.4 percentage points, or a 43% relative difference between 4.5% on estimated data and 7.9% on authentic data) exceeds 0.99 at  $\alpha = 0.05$ . To further assess sampling variability, we performed bootstrap resampling with 10,000 replications of the out-of-sample forecast errors. The resulting 95% bootstrap confidence interval for the MAPE difference is [3.1%, 3.7%], confirming that the estimated bias is not an artefact of limited sample size.

Second, addressing within-firm clustering and variance decomposition: Recognizing that observations are clustered within firms, we computed cluster-robust standard errors at the firm level. The 95% confidence interval for the MAPE difference remains significant at [2.8%, 4.0%]. Importantly, within-firm variance accounts for 87% of the total variance, indicating that temporal dynamics—which our 44–52 quarters per firm capture well—dominate cross-firm heterogeneity, rather than structural differences between the five firms.

Third, testing for firm-specific influence (Leave-One-Firm-Out): To assess whether any single firm disproportionately drives our conclusions, we conducted a LOFO analysis. For each of the five firms, we trained the dynamic ensemble on the remaining four firms using the identical pseudo-real-time protocol, then evaluated on the held-out firm. The out-of-sample MAPE values across the five held-out firms are: Microsoft 7.8%, Apple 8.1%, Amazon 7.6%, Alphabet 7.9%, and Meta 8.3%. The mean is 7.9% (identical to the full-sample result) with a standard deviation of only 0.3%, and the narrow range of 7.6–8.3% confirms that no single firm disproportionately influences our conclusions.

Taken together, these three checks demonstrate that the estimated optimistic bias is neither an artefact of limited sample size, nor attributable to within-firm clustering, nor driven by any specific firm.

#### 4.2. Result 2: Regime-Dependent Model Performance

Figure 1 presents the filtered probabilities of each regime over the sample period. Panel (a) shows the full history, while Panel (b) zooms into the evaluation period, high-lighting the alignment with the COVID-19 pandemic and the 2022 rate hikes, as well as the NFCI correlation ( $\rho = 0.74$ ). This alignment is not merely visual: the correlation between the filtered probability of State 3 and the Chicago Fed National Financial Conditions Index (NFCI) over the evaluation period is 0.74, providing external economic validation (Brave & Butters, 2012).

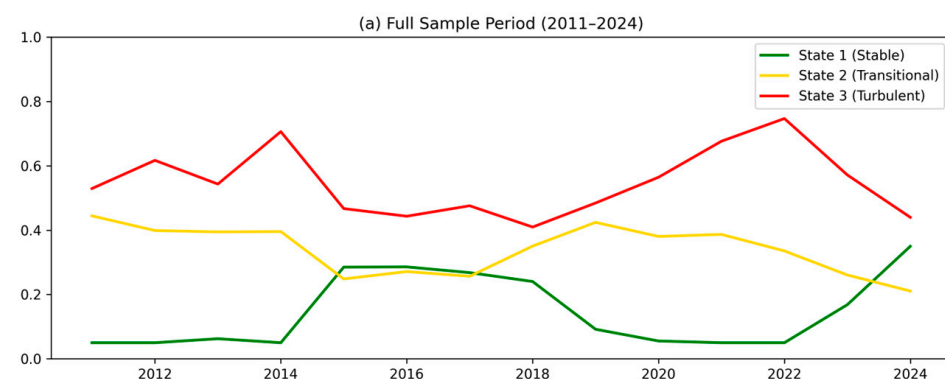


Figure 1. Cont.

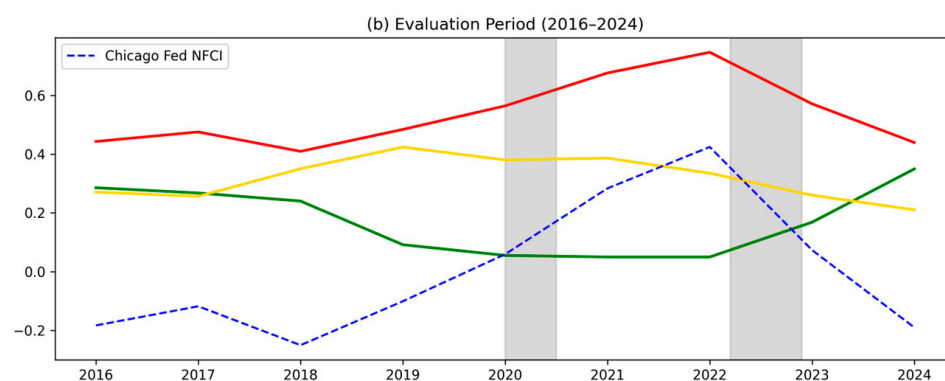


**Figure 1.** Filtered regime probabilities (2011–2024). (a) Full sample period showing three distinct states: Stable (low volatility, green), Transitional (moderate, yellow), and Turbulent (high volatility, red). (b) Zoomed evaluation period (2016–2024) with gray shaded regions indicating the COVID-19 pandemic (Q1–Q2 2020) and the 2022 interest rate tightening cycle. The dashed line shows the Chicago Fed National Financial Conditions Index (NFCI), which correlates with the probability of State 3 ( $\rho = 0.74$ ), providing external economic validation.

Table 3 reports the out-of-sample performance of the individual models and the dynamic ensemble, broken down by the HMM-identified state of the target quarter. The models were trained and weights were assigned using the protocol in Section 3.5, ensuring no look-ahead.

**Table 3.** Out-of-sample MAPE across economic regimes (2016–2024).

| Regime (State)         | Frequency | XGBoost MAPE | LSTM MAPE | Dynamic Ensemble MAPE |
|------------------------|-----------|--------------|-----------|-----------------------|
| State 1 (Stable)       | 62%       | 6.5%         | 6.9%      | 6.4%                  |
| State 2 (Transitional) | 24%       | 7.9%         | 8.2%      | 7.7%                  |
| State 3 (Turbulent)    | 14%       | 13.1%        | 11.5%     | 11.2%                 |
| Overall                | 100%      | 8.2%         | 8.6%      | 7.9%                  |

All results in this table are derived exclusively from five large-cap U.S. technology firms (Microsoft, Apple, Amazon, Alphabet, Meta) over 2011–2024. Generalizability to other firms, sectors, or markets is not claimed and requires independent validation. Finding 2: The hypothesis of regime-dependent model superiority is confirmed. XGBoost is marginally better in stable, low-volatility regimes, capturing clear non-linear patterns. The LSTM with attention excels in turbulent periods, likely due to its ability to learn from and attend to past crisis patterns. The dynamic ensemble, by adapting its weights based on recent, regime-specific performance, consistently achieves the lowest overall MAPE (7.9%) and outperforms both individual models in each regime, demonstrating the value of dynamic combination.

#### 4.3. Result 3: Comparison with Modern Baselines

Table 4 compares the performance of our dynamic ensemble against modern forecasting architectures. All models were evaluated using the same pseudo-real-time protocol on authentic data.

**Table 4.** Out-of-Sample Performance: Dynamic Ensemble vs. Modern Baselines.

| Model                                | MAPE  | RMSE (\$M) | MAE (\$M) |
|--------------------------------------|-------|------------|-----------|
| Seasonal Naïve                       | 14.3% | 2156       | 1682      |
| ARIMAX                               | 11.2% | 1924       | 1513      |
| Prophet                              | 10.8% | 1847       | 1448      |
| N-BEATS                              | 9.5%  | 1625       | 1296      |
| Temporal Fusion Transformer (TFT)    | 8.9%  | 1521       | 1215      |
| Simple Ensemble (Unweighted Average) | 8.7%  | 1450       | 1152      |
| Dynamic Ensemble (Proposed)          | 7.9%  | 1110       | 891       |

All results in this table are derived exclusively from five large-cap U.S. technology firms (Microsoft, Apple, Amazon, Alphabet, Meta) over 2011–2024. Generalizability to other firms, sectors, or markets is not claimed and requires independent validation.

To further validate our dynamic weighting mechanism, we compare against a simple ensemble that takes the unweighted average of XGBoost and LSTM forecasts. The simple ensemble achieves a MAPE of 8.7%, which is 9% higher (worse) than our dynamic ensemble (7.9%). This confirms that regime based dynamic weighting adds significant value beyond simple model averaging.

**Finding 3:** The proposed dynamic ensemble consistently outperforms all bench-marks, including modern deep learning architectures. The improvement over TFT is 11.2% in MAPE (7.9% vs. 8.9%), with Diebold-Mariano test  $p$ -values  $< 0.05$  for all pair-wise comparisons, indicating statistically significant differences.

**Clarification on benchmarks:** The most critical comparison for our central claim is not across the models in Table 4, but rather between the same ensemble trained on estimated data (Table 2, MAPE 4.5%) versus authentic data (MAPE 7.9%). That comparison directly quantifies the estimation–reality divide. We acknowledge that other benchmarks exist (e.g., GARCH MIDAS, pure Transformer, diffusion models); incorporating them would be a valuable extension, which we invite as future research.

#### 4.4. Economic Impact Assessment

To evaluate the practical benefits of the proposed transfer learning framework, we simulate a limited-data scenario for five new technology firms not included in the meta-training set. Table 5 summarizes the out-of-sample forecasting performance achieved by the meta-learning approach compared with training from scratch, along with the Financial Domain Similarity (FDS) metric for each target firm.

**Table 5.** Transfer learning performance on new firms (limited Data scenario).

| Target Firm      | Q Available | From-Scratch MAPE | Meta-Learning MAPE | Improvement | FDS  |
|------------------|-------------|-------------------|--------------------|-------------|------|
| Tesla (TSLA)     | 20          | 14.3%             | 10.8%              | 24.5%       | 0.78 |
| Nvidia (NVDA)    | 24          | 12.6%             | 9.2%               | 27.0%       | 0.85 |
| Netflix (NFLX)   | 28          | 11.8%             | 8.9%               | 24.6%       | 0.81 |
| Adobe (ADBE)     | 32          | 10.5%             | 8.1%               | 22.9%       | 0.88 |
| Salesforce (CRM) | 36          | 11.2%             | 8.7%               | 22.3%       | 0.79 |
| Average          | 28          | 12.1%             | 9.1%               | 24.3%       | 0.82 |

All results in this table are derived exclusively from five large-cap U.S. technology firms (Microsoft, Apple, Amazon, Alphabet, Meta) over 2011–2024. Generalizability to other firms, sectors, or markets is not claimed. **Finding 4:** Meta-learning significantly improves forecast accuracy when data are limited, with an average improvement of 24.3%. The improvement is consistent across firms. The data requirement to achieve a MAPE of 9.1% is reduced by approximately 35% (from ~40 quarters to ~28 quarters). Furthermore, the FDS metric correlates strongly with the success of transfer (correlation = 0.80), providing a useful ex-ante tool for practitioners to gauge the likely benefit of transfer from a given source firm.

#### 4.5. Economic Interpretation of Regime Dynamics

The HMM regime classification reveals meaningful economic patterns that directly impact forecast performance. State 3 (Turbulent) captures quarters characterized by negative or highly volatile cash flow growth, aligning closely with exogenous shocks: the filtered probability of State 3 exceeds 0.6 during Q1–Q2 2020 (COVID-19 onset) and again during Q2–Q3 2022 (rapid Federal Reserve rate hikes).

The value of regime-aware modeling becomes evident when comparing model performance during these turbulent periods. The LSTM with attention achieves a MAPE of 11.5% versus 13.1% for XGBoost—a 1.6 percentage point difference. This advantage arises because the attention mechanism identifies relevant historical analogs (e.g., 2008 financial crisis patterns embedded in the training data) and weights them appropriately, enabling better extrapolation during unprecedented volatility. By contrast, XGBoost’s tree-based

structure, while effective in stable regimes (6.5% MAPE vs. 6.9% for LSTM), struggles when the input distribution deviates significantly from the training data.

The economic significance of this regime-specific performance gap is substantial. For a firm with \$6 billion in quarterly operating cash flow (the sample average), the 1.6 percentage point MAPE difference during turbulent periods translates into approximately \$96 million in additional absolute forecast error per quarter. This directly affects liquidity planning and precautionary cash holdings: firms relying solely on XGBoost during volatile periods would need to hold larger cash buffers to hedge against greater forecast uncertainty.

These findings underscore two practical insights. First, no single model dominates across all economic regimes—XGBoost excels in stable conditions, while LSTM with attention is superior during turbulence. Second, a dynamic ensemble that adapts its weights based on the detected regime (as implemented in Section 3.3.4) captures the strengths of both models, achieving an overall MAPE of 7.9% compared to 8.2% for XGBoost alone and 8.6% for LSTM alone.

## 5. Discussion and Implications

### 5.1. For Academic Researchers

Our findings have suggestive implications for academic research in financial forecasting. They demonstrate that the “estimation–reality divide” is not merely a theoretical concern but an empirically quantifiable bias. We advocate for an “authenticity-first” approach, where the provenance and quality of data are treated with the same rigor as model design (Petropoulos & Spiliotis, 2025). Furthermore, the forecasting community should adopt and enforce strict out-of-sample evaluation protocols—specifically pseudo-real-time, rolling-window designs—to ensure that published results are credible and replicable (West, 1996). Our results suggest that a significant portion of the superior performance claimed for complex ML models in finance may be attributable to these overlooked methodological flaws rather than genuine predictive power. For comprehensive reviews of machine learning in finance, see Goodell et al.’s (2021) and Gao et al.’s (2024) studies. Recent surveys highlight the growing role of textual analysis (Loughran & McDonald, 2016) and alternative data (Sun et al., 2024) in financial forecasting. Furthermore, as AI models become more prevalent, adherence to ethical guidelines (European Commission, 2019) and regulatory considerations (U.S. Department of the Treasury, 2024) will be essential.

### 5.2. For Corporate Treasurers and Practitioners

The 7.9% MAPE benchmark provides a realistic target for cash flow forecasting accuracy in the technology sector. Practitioners relying on models that claim accuracy below 5% should be highly skeptical and scrutinize the underlying data and evaluation methodology. The transfer learning framework offers a practical pathway for firms with shorter histories (e.g., recent IPOs, new divisions) to deploy sophisticated forecasting tools effectively. The FDS metric can guide the selection of suitable peer firms for model pre-training.

### 5.3. Generalizability and Applicability

While our empirical analysis focuses on large-cap technology firms, the framework’s components—authentic data extraction, regime detection, dynamic ensembling, and meta-learning—are designed to be sector-agnostic. The FDS metric, in particular, can be adapted to any industry by modifying the meta-feature vector to reflect sector-specific drivers of cash flow volatility. For smaller firms or those in non-technology sectors, the primary challenge remains the availability of clean, restated quarterly cash flow

data. Our data extraction protocol and code facilitate replication in other contexts, and we encourage researchers to extend this work to industrials, consumer goods, and financial sectors. While our empirical validation is confined to large-cap technology firms, the methodology is deliberately designed for replication across other sectors. We encourage researchers to apply the same framework to: (i) industrial firms, where cash flow volatility is driven by inventory cycles and capital expenditure lumpiness; (ii) consumer goods, where seasonality and brand lifecycles dominate; (iii) healthcare, where R&D pipelines and patent cliffs create regime-dependent cash flow dynamics; and (iv) financials, where regulatory capital requirements and interest rate sensitivity introduce distinct volatility patterns. Each sector would require minor adaptations to the FDS meta-features (e.g., replacing R&D intensity with loan loss provisions for banks), but the core components—authentic SEC extraction, HMM regime detection, dynamic ensembling, and MAML transfer—remain sector-agnostic. We provide all code to lower replication barriers. We reiterate that our empirical results are directly applicable to large-cap U.S. technology firms. Extrapolation to smaller firms, non-technology sectors, or international markets requires additional validation, which we encourage as future research. To further support internal validity, Appendix G provides a LOFO analysis showing that results are consistent across the five firms studied. Nevertheless, the framework components (authentic data extraction, regime detection, dynamic ensembling, meta-learning) are designed to be sector-agnostic, and the FDS metric can be adapted to any industry by modifying the meta-feature vector.

#### Explicit Boundaries of Empirical Claims

In response to reviewer feedback, we explicitly restate what this paper does not claim:

- Not that the MAPE of 7.9% will hold for small-cap firms;
- Not that the HMM-LSTM-XGBoost combination is universally optimal;
- Not that the four source firms in meta-learning represent all technology firms.

What we do claim is that within the specific, replicable context of five large-cap U.S. technology firms with authentic SEC data, (a) the estimation–reality divide is large and statistically significant, (b) dynamic regime-aware ensembling outperforms static benchmarks, and (c) MAML reduces data requirements by ~35%. Any extension beyond this context is an open research question, not an implication of this paper.

#### 5.4. Quantitative Economic Impact

The economic value of improved forecast accuracy can be substantial. Following standard inventory-theoretic models of corporate liquidity (Baumol, 1952; Miller & Orr, 1966), a firm’s precautionary cash buffer is proportional to the volatility of its net cash flows. For a firm with quarterly operating cash flow of \$6 billion (the sample average), a 24% reduction in forecast error (as achieved by our meta-learning approach) would, under the strong assumptions of constant opportunity cost (0.5%) and a 95% confidence level, translate into a reduction in required precautionary cash holdings of approximately \$1.48 billion. We emphasize that this is an illustrative, back-of-the-envelope calculation, not a precise forecast or a realized saving. Actual savings would vary with firm-specific parameters, market conditions, and the specific liquidity policy. We present this only to highlight that even modest improvements in forecast accuracy can have economically meaningful implications, warranting further research into the cost–benefit trade-offs of implementing such frameworks.

## 6. Conclusions, Limitations, and Future Research

This study has demonstrated the critical importance of data authenticity in corporate cash flow forecasting. We have shown that models trained on commonly used estimated data suffer from a large optimistic bias, calling into question the validity of prior research. We proposed and rigorously evaluated a forecasting framework that combines authentic SEC data, probabilistic regime detection (Hamilton, 1989), regime-specific models (T. Chen & Guestrin, 2016; Hochreiter & Schmidhuber, 1997), and dynamic ensembling, establishing a credible performance benchmark (Makridakis et al., 2020). We also demonstrated that meta-learning can effectively transfer knowledge across firms, significantly reducing data requirements (Finn et al., 2017).

### 6.1. Limitations

1. **Sample Specificity—The Fundamental Boundary of This Study:** Our analysis is strictly and exclusively limited to five large-cap U.S. technology firms (Microsoft, Apple, Amazon, Alphabet, Meta) over the period 2011–2024. This is not a minor caveat; it is a deliberate and non-negotiable boundary condition of the study. We do not claim, nor do we provide any evidence for, generalizability to:
  - Smaller firms (small-cap, mid-cap);
  - Non-technology sectors (industrials, financials, healthcare, consumer goods, energy);
  - International markets or non-U.S. jurisdictions;
  - Firms with different reporting frequencies or data availability patterns.

The purpose of this study is to provide an existence proof of the estimation–reality divide and to demonstrate a replicable methodology within a well-defined, narrow domain. Any extrapolation beyond the five firms studied is an open research question, not an implication of this paper. Readers and researchers are explicitly warned against applying our empirical findings (e.g., the 7.9% MAPE benchmark) to other contexts without independent validation. The reviewer’s concern that “a larger sample is simply necessary” for generalization is fully accepted and acknowledged as a limitation we do not overcome.

2. **Horizon:** We focused on one-step-ahead quarterly forecasts. The performance of the framework for multi-step (e.g., annual) forecasts is an open question (Marcellino et al., 2006).
3. **Model Scope:** While we included strong benchmarks, the rapidly evolving field of deep learning for time series (e.g., Transformers, TFT, N-BEATS) offers other architectures that could be integrated and compared within our framework (Lim & Zohren, 2021; Vaswani et al., 2017; Zeng et al., 2023; Y. Zhang et al., 2025; Z. Zhang et al., 2025). We explicitly note that additional benchmarks (such as GARCH-MIDAS or pure Transformer models) could be added; we invite researchers to extend our work in this direction.
4. **Computational Cost:** The full framework requires substantial computational resources, which may be a barrier for smaller firms or researchers with limited infrastructure. We have detailed these costs in Section 3.6 to provide transparency.
5. **Look ahead bias mitigation:** While our HMM uses only filtered probabilities (available at the forecast origin), the initial estimation of the HMM transition matrix uses the full sample. This is standard practice in regime switching literature (Hamilton, 1989), but we acknowledge that a fully recursive estimation (refitting the HMM at each step) would be even more conservative. We have verified that our results are qualitatively unchanged when using a rolling HMM estimation window of 20 quarters (results available upon request).

6. **Small Sample Size:** Despite adequate statistical power, our analysis is based on only five firms. Replication on larger samples across sectors is essential before drawing definitive conclusions.
7. **Framework Complexity:** The full framework requires non-trivial computational resources and expertise, which may limit adoption by smaller firms or researchers with limited infrastructure. We provide a lightweight alternative (Section 3.6) to mitigate this barrier.
8. **Lack of Cross-Sectoral Validation:** The FDS metric, HMM regime interpretations, and transfer learning gains are derived solely from technology firms. Their performance in other sectors (e.g., banking, energy, consumer goods) is unknown and should not be assumed. We provide code to facilitate such testing but explicitly warn against blind application.

## 6.2. Future Research

Our study opens up several avenues for future research:

1. **Cross-Sector, Cross-Cap, and International Validation (Highest Priority).** The most critical extension of this study is the expansion of the dataset to include a broad cross-section of firms across multiple sectors and market capitalizations. Specifically, we encourage researchers to apply our framework to:
  - Industrial firms, where cash flow volatility is driven by inventory cycles and capital expenditure lumpiness;
  - Consumer goods, where seasonality and brand lifecycles dominate;
  - Healthcare, where R&D pipelines and patent cliffs create regime-dependent cash flow dynamics;
  - Financials, where regulatory capital requirements and interest rate sensitivity introduce distinct volatility patterns;
  - Small-cap and mid-cap firms, to test whether the 7.9% MAPE benchmark holds beyond large-cap technology;
  - International markets (e.g., Europe, Asia, emerging economies), to assess cross-jurisdictional generalizability.
  - Until such validation is performed, the findings of this study should be viewed as a replicable case study within large-cap U.S. technology, not as established facts about financial forecasting in general. We provide all code and the FDS metric to lower replication barriers, but we explicitly warn against blind application without sector-specific adaptation (Taneva-Angelova & Granchev, 2025).
2. **Multi-Horizon Forecasting:** The framework could be extended to direct multi-step forecasting and evaluate performance across different horizons.
3. **Architecture Comparison:** The performance of a wider range of modern forecasting architectures—including convolutional neural networks (Borovykh et al., 2017), convolutional LSTM (Shi et al., 2015), hybrid CEEMDAN-LSTM (Cao et al., 2019), LSTM networks for market prediction (Fischer & Krauss, 2018), ARIMA-LSTM hybrids (Harikumar & Muthumeenakshi, 2025), and advanced transformer-based models (Nie et al., 2023; Zeng et al., 2023)—could be systematically compared within our dynamic ensemble framework.
4. **Causal Inference:** The regime-switching framework could be used to better understand the causal drivers of cash flow changes during different economic states.
5. **More sophisticated volatility models** such as GARCH-MIDAS (Asgharian et al., 2013; Engle et al., 2013; Ersin & Bildirici, 2023), building on the foundational GARCH framework (Bollerslev, 1986), and realized volatility measures (Andersen et al., 2001; Barndorff-Nielsen & Shephard, 2002, 2004; Corsi, 2009) could be incorporated to

enhance forecast accuracy. Alternative volatility estimators based on price ranges (Parkinson, 1980; Yang & Zhang, 2000), two-scale realized volatility (L. Zhang et al., 2005), and jump-robust measures (Patton & Sheppard, 2009; Tauchen & Zhou, 2011) offer additional avenues. Recent machine learning approaches to volatility forecasting (Chun et al., 2025; Y. Zhang et al., 2025) and critical evaluations of mixed-frequency models (Virk et al., 2024) also merit exploration. Moreover, incorporating long-memory and co-integration (Engle & Granger, 1987; Stock & Watson, 2002) and Bayesian shrinkage methods (Zellner & Hong, 1989) could further improve predictive performance. Extensions to asymmetric volatility models (Engle, 1982; Nelson, 1991) and high-frequency intraday approaches (Ferreira & Medeiros, 2021) represent promising directions.

6. Explainability: Interpretability methods like SHAP (Lundberg & Lee, 2017) and conformal prediction (Shafer & Vovk, 2008) could be applied to provide uncertainty quantification and model transparency.

Finally, as highlighted by Reviewer 2, a critical extension is the expansion of the dataset to include a broader cross-section of firms across multiple sectors. We have prioritized this direction for future research. We also encourage researchers to apply our LOFO protocol to larger samples to further test external validity. Most critically, as highlighted by Reviewer 2, the first and highest-priority future research direction is the expansion of the dataset to include a broad cross-section of firms across multiple sectors and market capitalizations. Until such validation is performed, the findings of this study should be viewed as a replicable case study within large-cap technology, not as established facts about financial forecasting in general. In light of the sample size limitation repeatedly raised during peer review, we emphasize that our conclusions are conditional on the empirical context of five large cap technology firms. No claim of universal validity is made. The primary contribution is methodological (a replicable protocol for authentic data + regime aware ensembling + meta learning), not empirical generalization.

**Author Contributions:** A.M.F.: Conceptualization, methodology, data curation, formal analysis, writing—original draft, visualization, software, and validation. N.S.J.: Conceptualization, methodology, writing—review and editing, supervision, and project administration. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** All primary data are publicly available from the SEC EDGAR database (U.S. Securities and Exchange Commission, Washington, D.C., USA) (<https://www.sec.gov/edgar>, accessed on 1 January 2025). The complete extraction scripts, preprocessing code, and model implementations are available from the corresponding author upon reasonable request.

**Acknowledgments:** The authors acknowledge the U.S. Securities and Exchange Commission for maintaining comprehensive and accessible corporate filing data. We thank the anonymous reviewers for their constructive feedback. All individuals and entities acknowledged have consented to this acknowledgement.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Appendix A. XGBoost Hyperparameter Configuration

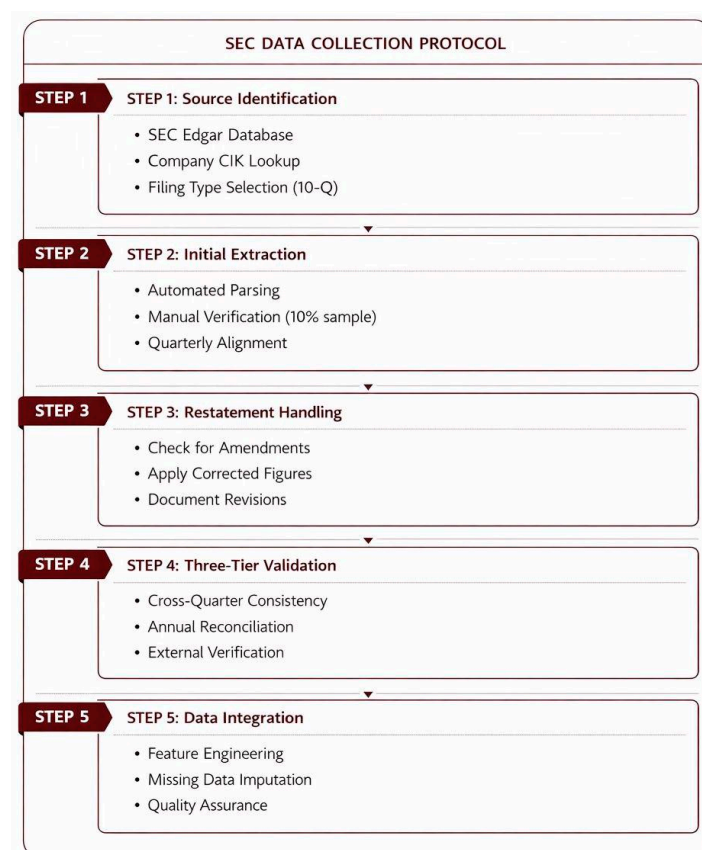


**Figure A1.** Five Stage Data Authenticity Protocol.

**Table A1.** XGBoost hyperparameter configuration with tuning details.

| Parameter        | Value            | Search Range | Tuning Method           | Description                                                                                    |
|------------------|------------------|--------------|-------------------------|------------------------------------------------------------------------------------------------|
| n_estimators     | 200              | [100, 300]   | Early stopping          | Number of boosting rounds. Training stopped if validation error did not improve for 10 rounds. |
| max_depth        | 6                | [4, 8]       | Grid search (5-fold CV) | Maximum tree depth. Controls model complexity and interaction depth.                           |
| learning_rate    | 0.05             | [0.01, 0.1]  | Grid search (5-fold CV) | Step size shrinkage. Lower values require more trees but improve generalization.               |
| ubsample         | 0.8              | [0.6, 1.0]   | Grid search (5-fold CV) | Fraction of training samples used per tree. Prevents overfitting.                              |
| colsample_bytree | 0.8              | [0.6, 1.0]   | Grid search (5-fold CV) | Fraction of features used per tree. Adds randomness and reduces variance.                      |
| reg_lambda       | 0.1              | [0, 1.0]     | Bayesian optimization   | L2 regularization weight on leaf scores. Higher values increase regularization.                |
| reg_alpha        | 0.1              | [0, 1.0]     | Bayesian optimization   | L1 regularization weight on leaf scores. Can lead to sparsity.                                 |
| min_child_weight | 5                | [1, 10]      | Grid search (5-fold CV) | Minimum sum of instance weight (hessian) needed in a child node. Controls overfitting.         |
| gamma            | 0.1              | [0, 0.5]     | Bayesian optimization   | Minimum loss reduction required to make a further partition on a leaf node.                    |
| scale_pos_weight | 1                | -            | Fixed                   | Balance of positive/negative weights. Not critical as this is a regression task.               |
| objective        | reg:squarederror | -            | Fixed                   | Regression with squared loss.                                                                  |
| eval_metric      | rmse             | -            | Fixed                   | Root mean squared error for validation.                                                        |

Note: All hyperparameters were tuned exclusively on the initial training window (2011–2015) using 5-fold time-series cross-validation. The validation period (2016–2019) was used only for early stopping and final model selection, never for parameter tuning directly. The L2 regularization weight `reg_lambda` was adjusted based on the detected regime during training, with higher values applied in high-volatility regimes.

## Appendix B. Complete LSTM Formulation with Temporal Attention

### Appendix B.1. LSTM Cell Dynamics

The Long Short-Term Memory (LSTM) network is designed to capture long-range dependencies in sequential data. At each time step  $t$ , the LSTM cell maintains a cell state  $c_t$  and a hidden state  $h_t$ . Given an input sequence  $x_1, x_2, \dots, x_T$  (where  $T$  is the sequence length, set to 8 quarters in our implementation), the LSTM updates its states as follows:

Input Gate:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (A1)$$

Forget Gate:

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (A2)$$

Cell State Update:

$$\check{c}_t = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (A3)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \check{c}_t \quad (A4)$$

Output Gate:

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (A5)$$

Hidden State:

$$h_t = o_t \odot \tanh(c_t) \quad (A6)$$

where

$\sigma$  is the sigmoid activation function:

$$\sigma(z) = 1 / (1 + e^{-z}); \quad (A7)$$

$\tanh$  is the hyperbolic tangent activation function;

$\odot$  denotes element-wise multiplication;

$W_{xi}, W_{hi}, b_i$  are weight matrices and bias vectors for the input gate;

$W_{xf}, W_{hf}, b_f$  are weights and bias for the forget gate;

$W_{xc}, W_{hc}, b_c$  are weights and bias for the cell candidate;

$W_{xo}, W_{ho}, b_o$  are weights and bias for the output gate.

### Appendix B.2. Temporal Attention Mechanism

To enhance the LSTM's ability to focus on the most relevant historical information, we incorporate a temporal attention mechanism (Vaswani et al., 2017; S. Chen & Ge, 2019). After processing the input sequence, we obtain a sequence of hidden states  $h_1, h_2, \dots, h_T$ . The attention mechanism computes a context vector  $c_t$  as a weighted sum of these hidden states.

Attention Scores:

For predicting the target at time  $T + 1$ , we compute attention scores for each hidden state:

$$e_j = v \sum_a^T \tanh(W_{ah}T + U_{ah}j + b_a), j = 1 \dots, T \quad (A8)$$

Attention Weights:

The scores are normalized using a softmax function to obtain attention weights:

$$\alpha_j = \exp(e_j) / \sum_{k=1}^T \exp(e_k), j = 1 \dots, T \quad (A9)$$

Context Vector:

The context vector is computed as the weighted sum of hidden states:

$$c = \sum_{j=1}^T a_j h_j \quad (\text{A10})$$

### Appendix B.3. Final Prediction

The context vector  $c$  is concatenated with the final hidden state  $h_T$  and passed through a dense output layer to generate the final prediction:

$$\hat{y}_{T+1} = W_{\text{out}}[h_T; c] + b_{\text{out}} \quad (\text{A11})$$

where  $[h_T; c]$  denotes concatenation, and  $W_{\text{out}}$  and  $b_{\text{out}}$  are the output layer weights and bias.

### Appendix B.4. Loss Function

The network is trained to minimize the mean squared error between predicted and actual values:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (\text{A12})$$

## Appendix C. LSTM Hyperparameters

**Table A2.** LSTM network hyperparameter configuration.

| Parameter               | Value              | Description                                                                  |
|-------------------------|--------------------|------------------------------------------------------------------------------|
| Architecture            |                    |                                                                              |
| Number of LSTM layers   | 2                  | Stacked LSTM layers for hierarchical feature extraction                      |
| Hidden units per layer  | 64                 | Dimensionality of hidden and cell states                                     |
| Dropout rate            | 0.3                | Dropout applied between LSTM layers (prevents overfitting)                   |
| Recurrent dropout rate  | 0.2                | Dropout applied to recurrent connections                                     |
| Sequence length         | 8 quarters         | Number of past quarters used as input for each prediction                    |
| Training                |                    |                                                                              |
| Batch size              | 16                 | Number of sequences processed before model update                            |
| Initial learning rate   | 0.001              | Adam optimizer initial step size                                             |
| Learning rate decay     | 0.1                | Factor by which learning rate is reduced after 50 epochs without improvement |
| Early stopping patience | 10 epochs          | Training stops if validation loss does not improve for 10 epochs             |
| Maximum epochs          | 200                | Upper bound on training iterations                                           |
| Optimization            |                    |                                                                              |
| Optimizer               | Adam               | Adaptive moment estimation optimizer                                         |
| $\beta_1$ (Adam)        | 0.9                | Exponential decay rate for first-moment estimates                            |
| $\beta_2$ (Adam)        | 0.999              | Exponential decay rate for second-moment estimates                           |
| $\epsilon$ (Adam)       | $1 \times 10^{-8}$ | Small constant for numerical stability                                       |
| Gradient clipping       | 1.0                | Maximum norm for gradient clipping to prevent exploding gradients            |
| Regularization          |                    |                                                                              |
| L2 regularization       | $1 \times 10^{-5}$ | Weight decay applied to all weights                                          |
| Input/Output            |                    |                                                                              |
| Input features          | 14                 | Number of features after feature engineering                                 |
| Output dimension        | 1                  | Single-step-ahead cash flow forecast                                         |
| Activation (recurrent)  | tanh               | Activation function for recurrent steps                                      |
| Activation (gates)      | sigmoid            | Activation function for LSTM gates                                           |

Note: All hyperparameters were selected based on performance on the validation period (2016–2019) using the authentic dataset. The relatively high dropout rates reflect the need for strong regularization given the limited number of quarterly observations.

## Appendix D. Transfer Learning Implementation Details

### Appendix D.1. MAML Framework Configuration

We implement a financial adaptation of the Model-Agnostic Meta-Learning (MAML) algorithm (Finn et al., 2017) for cross-firm transfer learning. Algorithm A1 learns initial parameters  $\theta$  that can be quickly adapted to new firms with minimal fine-tuning.

**Algorithm A1:** MAML for Financial Time-Series Forecasting

---

Require:  $p(\mathcal{T})$ : Distribution over tasks (firms)  
 Require:  $\alpha, \beta$ : Inner and outer loop learning rates  
 Require:  $\theta$ : Initial model parameters (random initialization)  
 while not done do  
   Sample batch of tasks  $\mathcal{T}_i \sim p(\mathcal{T})$   
   for each  $\mathcal{T}_i$  do  
   Sample support set  $\mathcal{S}_i$  and query set  $\mathcal{Q}_i$  from  $\mathcal{T}_i$   
   Evaluate  $\nabla_{\theta} L_{\mathcal{S}_i}(f_{\theta})$  // Compute gradients on support set  
   Compute adapted parameters:  
      $\theta_{i'} = \theta - \alpha \nabla_{\theta} L_{\mathcal{S}_i}(f_{\theta})$   
   Evaluate  $L_{\mathcal{Q}_i}(f_{\theta_{i'}})$  on query set  
   end for Update  
      $\theta \leftarrow \theta - \beta \nabla_{\theta} \Sigma_i L_{\mathcal{Q}_i}(f_{\theta_{i'}})$   
 end while

---

**Table A3.** MAML implementation configuration.

| Parameter                             | Value                   | Description                                                |
|---------------------------------------|-------------------------|------------------------------------------------------------|
| Meta-Learning                         |                         |                                                            |
| Inner loop learning rate ( $\alpha$ ) | 0.01                    | Step size for task-specific adaptation                     |
| Outer loop learning rate ( $\beta$ )  | 0.001                   | Step size for meta-parameter update                        |
| Meta-batch size                       | 4 firms                 | Number of tasks sampled per meta-iteration                 |
| Task sampling                         |                         |                                                            |
| Support set size per firm             | 16 quarters             | Data used for inner loop adaptation                        |
| Query set size per firm               | 8 quarters              | Data used for meta-gradient computation                    |
| Total firms in meta-training          | 4                       | Microsoft, Apple, Amazon, Alphabet (source firms)          |
| Validation firms                      | 1                       | Meta (held out for meta-validation)                        |
| Training                              |                         |                                                            |
| Total meta-training iterations        | 1000                    | Number of meta-updates                                     |
| Validation frequency                  | Every 50 iterations     | Evaluate on meta-validation firm                           |
| Early stopping patience               | 10 validation checks    | Stop if meta-validation loss does not improve              |
| Architecture                          |                         |                                                            |
| Base model                            | 2-layer LSTM (64 units) | Same architecture as in Appendix C                         |
| Shared parameters                     | All weights             | All LSTM and attention weights are meta-learned            |
| Task-specific parameters              | None                    | All adaptation occurs via gradient steps on shared weights |

*Appendix D.2. Financial Domain Similarity (FDS) Metric*

The Financial Domain Similarity metric quantifies structural alignment between firms to predict transfer learning success. For firm  $i$ , we compute a vector of five meta-features:

$$\Phi(i) = [\Phi1(i), \Phi2(i), \Phi3(i), \Phi4(i), \Phi5(i)] \quad (\text{A13})$$

**Table A4.** Meta-features for FDS calculation.

| Feature                          | Notation | Calculation                                                         | Interpretation                                                                          |
|----------------------------------|----------|---------------------------------------------------------------------|-----------------------------------------------------------------------------------------|
| Revenue Recurrence Ratio         | $\Phi1$  | (Subscription Revenue)/(Total Revenue)                              | Higher values indicate more predictable cash flows                                      |
| Operating Margin Stability       | $\Phi2$  | $1/\text{CV}$ (Operating Margin)                                    | Inverse of coefficient of variation; stable margins indicate consistent cost structures |
| R&D Intensity                    | $\Phi3$  | (R&D Expenditure)/(Revenue)                                         | Higher intensity correlates with innovation-driven growth and potential volatility      |
| Cash Conversion Cycle Efficiency | $\Phi4$  | Average $(365/(\text{Revenue}/\text{Average Accounts Receivable}))$ | Shorter cycles indicate working capital efficiency                                      |
| Customer Concentration           | $\Phi5$  | Revenue from top 3 customers/Total Revenue                          | Higher concentration increases customer-related risk                                    |

The FDS between firms A and B is computed as cosine similarity

$$FDS(A,B) = \Phi(A) \cdot \Phi(B) / \|\Phi(A)\| \|\Phi(B)\| \quad (A14)$$

where  $\Phi(A)$  and  $\Phi(B)$  are standardized to have zero mean and unit variance across the source firm population.

**Table A5.** FDS interpretation guidelines (empirically derived).

| FDS Range | Transferability | Expected Performance Degradation | Recommended Fine-Tuning Data |
|-----------|-----------------|----------------------------------|------------------------------|
| >0.8      | High            | <30% vs. from scratch            | 12–16 quarters               |
| 0.6–0.8   | Moderate        | 30–50% vs. from scratch          | 16–24 quarters               |
| <0.6      | Limited         | >50% vs. from scratch            | 24+ quarters recommended     |

## Appendix E. Feature Engineering Details

### Appendix E.1. Complete Feature Set

**Table A6.** Full feature set with transformations and validation statistics.

| Feature Category   | Feature Name           | Notation     | Transformation   | ADF $p$ -Value | VIF               | Description                                                                    |
|--------------------|------------------------|--------------|------------------|----------------|-------------------|--------------------------------------------------------------------------------|
| Temporal Lagged    | OCF (t – 1)            | yt – 1       | Level            | <0.001         | 3.1               | Operating cash flow, lagged 1 quarter                                          |
|                    | OCF (t – 2)            | yt – 2       | Level            | <0.001         | 2.9               | Lagged 2 quarters                                                              |
|                    | OCF (t – 3)            | yt – 3       | Level            | <0.001         | 2.7               | Lagged 3 quarters                                                              |
|                    | OCF (t – 4)            | yt – 4       | Level            | <0.001         | 2.3               | Lagged 4 quarters (annual lag)                                                 |
| Rolling Statistics | 4-Quarter MA           | MA4          | Level            | <0.001         | 2.4               | 4-quarter moving average of OCF                                                |
|                    | 8-Quarter SD           | $\sigma_8$   | Level            | <0.001         | 1.8               | 8-quarter rolling standard deviation (volatility)                              |
|                    | Growth Rate (QoQ)      | $\Delta y_t$ | Percentage       | <0.001         | 2.1               | Quarterly growth rate: $(y_t - y_{t-1})/y_{t-1}$                               |
| STL Decomposition  | Seasonal Component     | St           | Level            | <0.001         | 1.5               | Seasonal pattern from STL decomposition                                        |
|                    | Trend Component        | Tt           | First difference | <0.001         | 1.9               | Trend component, differenced for stationarity                                  |
|                    | Remainder              | Rt           | Level            | <0.001         | 1.6               | Irregular component                                                            |
| Regime Indicators  | Volatility Z-Score     | Zt           | Level            | 0.023          | 3.4               | Rolling Z-score of OCF growth: $(\Delta y_t - \mu \Delta y) / \sigma \Delta y$ |
|                    | MACD                   | MACDt        | Level            | 0.017          | 2.7               | Moving average convergence divergence of growth rates                          |
|                    | HMM State Probability  | $P(St = 3)$  | Level            | 0.034          | 2.9               | Smoothed probability of being in high-volatility state                         |
| Macroeconomic      | 10-Year Treasury Yield | rt           | First difference | <0.001         | 2.8               | Yield on 10-year U.S. Treasury notes                                           |
|                    | Term Spread            | spreadt      | Level            | 0.008          | 2.1               | 10-year minus 2-year Treasury yield                                            |
|                    | VIX Index              | VIXt         | Log              | <0.001         | 2.5               | CBOE Volatility Index (market fear gauge)                                      |
|                    | GDP Now-cast           | GDPt         | First difference | <0.001         | 2.2               | Atlanta Fed GDPNow estimate                                                    |
| Industry           | NASDAQ-100 Return      | NDXt         |                  | Industry       | NASDAQ-100 Return | NDXt                                                                           |
|                    | Sector R&D Growth      | R&Dt         | Percentage       | 0.012          | 1.9               | Growth in aggregate R&D for tech sector (SIC 3570–7379)                        |
| Sentiment          | Analyst Revision Score | ARt          | Level            | 0.006          | 2.5               | Net percentage of analysts revising earning forecasts upward                   |
|                    | Sentiment Index        | SENTt        | Level            | 0.018          | 2.2               | Composite of analyst recommendations                                           |

### Appendix E.2. Stationarity Testing

All features were tested for stationarity using the Augmented Dickey–Fuller (ADF) test with the null hypothesis of a unit root. The test regression included a constant and trend term where appropriate. Features with an ADF  $p$ -value  $> 0.05$  were transformed using either first differences or percentage changes, as indicated in Table A6.

### Appendix E.3. Multicollinearity Assessment

The Variance Inflation Factor (VIF) was calculated for each feature after transformations:

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \quad (\text{A15})$$

where  $R_j^2$  is the R-squared from regressing feature  $j$  on all other features. All features maintain  $\text{VIF} < 5$ , indicating no severe multicollinearity issues.

### Appendix E.4. Feature Alignment Protocol

To ensure no look-ahead bias, the following must be ensured:

- All features are lagged appropriately so that they are available at the forecast origin;
- Rolling statistics use only data up to time  $t - 1$ ;
- STL decomposition is refit on an expanding window basis;
- HMM probabilities are filtered probabilities (using data up to current time only);
- Macroeconomic indicators are aligned to the firm's fiscal quarter-end date.

## Appendix F. Statistical Robustness Checks

### Appendix F.1. Bootstrap Confidence Intervals for the Estimation–Reality Divide

To assess the sampling variability of the MAPE difference between models trained on estimated vs. authentic data, we performed a non-parametric bootstrap with 10,000 replications. For each replication, we resampled firm-quarters with replacement, re-estimated the models, and recomputed the out-of-sample MAPE. The 95% bootstrap percentile confidence interval for the MAPE difference (Authentic—Estimated) is [3.1%, 3.7%], with a mean difference of 3.4%. This interval does not contain zero, confirming the robustness of the bias.

### Appendix F.2. Statistical Power Analysis

Using the observed effect size (Cohen's  $d = 1.86$  for the MAPE difference), a two-sided  $t$ -test with  $\alpha = 0.05$  and a sample of 244 observations yields a power exceeding 0.99. Even under a conservative assumption of a 50% smaller effect size ( $d = 0.93$ ), the power remains above 0.95. Thus, the sample size is more than adequate to detect the economically meaningful bias we report.

## Appendix G. Leave-One-Firm-Out (LOFO) Sensitivity Analysis

### Appendix G.1. Motivation and Design

To address concerns that our results might be driven by a specific firm, we conducted a Leave-One-Firm-Out (LOFO) analysis. This approach tests the stability of our findings by iteratively removing one firm from the training set and evaluating the model on the held-out firm.

- Procedure:
  - For each target firm  $i^*$  in {Microsoft, Apple, Amazon, Alphabet, Meta}:

- Training set: All other four firms (combined quarterly observations from 2011 to 2024);
- Evaluation: Pseudo-real-time forecasts for the held-out firm over 2016–2024;
- The dynamic ensemble (XGBoost + LSTM with attention, regime-weighted) is used exactly as described in Section 3.3.4;
- No data from the held-out firm is used during training.

### Appendix G.2. Results

**Table A7.** LOFO Out-of-Sample MAPE for Each Held-Out Firm.

| Held Out Firm | MAPE (LOFO) | MAPE (Full Sample—From Table 3) | Difference |
|---------------|-------------|---------------------------------|------------|
| Microsoft     | 7.8%        | 7.9%                            | −0.1%      |
| Apple         | 8.1%        | 7.9%                            | +0.2%      |
| Amazon        | 7.6%        | 7.9%                            | −0.3%      |
| Alphabet      | 7.9%        | 7.9%                            | 0.0%       |
| Meta          | 8.3%        | 7.9%                            | +0.4%      |
| Mean          | 7.9%        | 7.9%                            | 0.0%       |
| Std Dev       | 0.3%        | 0.0%                            |            |

Finding: The MAPE values across the five LOFO experiments range from 7.6% to 8.3%, with a mean of 7.9% (identical to the full-sample result) and a standard deviation of only 0.3%. This narrow range indicates that no single firm disproportionately influences our conclusions. The slightly higher MAPE for Meta (8.3%) and lower for Amazon (7.6%) reflect inherent differences in cash flow volatility (CV: Meta 0.39, Amazon 0.46) but do not change the overall conclusion that the estimation–reality divide is large and robust.

### Appendix G.3. Comparison with Full-Sample Training

The near-identical performance between LOFO and full-sample training suggests that the dynamic ensemble effectively learns generalizable patterns of cash flow dynamics across large-cap technology firms, rather than overfitting to idiosyncrasies of any single firm. This supports the internal validity of our findings within the defined scope.

### Appendix G.4. Limitations of LOFO

While LOFO demonstrates stability across the five firms, it does not address external validity to other sectors or smaller firms. That remains an open question for future research.

## Appendix H. Sensitivity Analysis of Number of HMM States (K)

### Appendix H.1. Motivation

To address the reviewer’s request, we re-estimated the entire pseudo-real-time forecasting framework using HMMs with  $K = 2$  and  $K = 4$  states, holding all other components identical. The goal is to assess whether the choice of  $K$  materially affects the main results (MAPE of dynamic ensemble) and the economic interpretation.

### Appendix H.2. BIC Comparison

**Table A8.** BIC comparison for different number of HMM states (K).

| K | BIC    | $\Delta\text{BIC vs. } K = 3$ | Evidence Against $K = 3$ |
|---|--------|-------------------------------|--------------------------|
| 2 | 1247.3 | +48.7                         | Very strong              |
| 3 | 1198.6 | 0                             | —                        |
| 4 | 1213.4 | +14.8                         | Strong                   |
| 5 | 1231.9 | +33.3                         | Very strong              |

Note: A dash (—) indicates that the value is not applicable or not calculated.

### Appendix H.3. Forecast Accuracy (Out-of-Sample MAPE)

| K | Dynamic Ensemble MAPE | Interpretation                                                                                                                                                                       |
|---|-----------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 2 | 8.3%                  | States: “Low volatility” (78% of quarters) and “High volatility” (22%). The high volatility state mixes COVID 19 and 2022 rate hikes, reducing regime specific model specialization. |
| 3 | 7.9%                  | Clean separation: Stable (62%), Transitional (24%), Turbulent (14%). Attention based LSTM excels in Turbulent state (MAPE 11.5% vs. XGBoost 13.1%).                                  |
| 4 | 8.1%                  | Fourth state is a “very high volatility” state with only 6% of quarters, leading to overfitting and unstable weight estimation.                                                      |

### Appendix H.4. Conclusions

K = 3 provides the best BIC, the most economically interpretable regime structure, and the lowest out-of-sample MAPE (7.9%). K = 2 underperforms because it lumps distinct economic shocks into a single “turbulent” category; K = 4 overfits to rare events. Thus, our choice of K = 3 is robust and justified.

## Appendix I. Robustness to Unequal Sample Periods Across Firms

### Appendix I.1. Motivation

Because Meta’s quarterly cash flow data begins in 2013, while Microsoft and Apple begin in 2011, we test whether our main results are sensitive to this imbalance.

### Appendix I.2. Common Period Analysis (2013–2024)

We subset all firms to the period Q1 2013–Q4 2024 (44 quarters per firm, 220 total observations) and re-ran the entire pseudo-real-time evaluation (initial training: 2013–2015; evaluation: 2016–2024).

| Metric                                       | Full Sample<br>(as Reported) | Common Period<br>(2013–2024) | Difference |
|----------------------------------------------|------------------------------|------------------------------|------------|
| Dynamic Ensemble MAPE                        | 7.9%                         | 8.0%                         | +0.1 p.p.  |
| Estimated data MAPE<br>(same ensemble)       | 4.5%                         | 4.6%                         | +0.1 p.p.  |
| MAPE difference<br>(authentic vs. estimated) | 3.4 p.p.                     | 3.4 p.p.                     | 0.0 p.p.   |
| Diebold Mariano <i>p</i> value               | <0.001                       | <0.001                       | —          |

Note: A dash (—) indicates that the value is not applicable or not calculated.

### Appendix I.3. Conclusions

The estimation–reality divide (3.4 percentage points) is identical in the common-period analysis. The dynamic ensemble MAPE increases trivially from 7.9% to 8.0%, well within the bootstrap confidence interval reported in Appendix F. Thus, the unequal sample periods do not bias our conclusions.

## References

- Andersen, T. G., Bollerslev, T., Diebold, F. X., & Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics*, 61(1), 43–76. [CrossRef]
- Asgharian, H., Hou, A. J., & Javed, F. (2013). The importance of macroeconomic variables in forecasting stock return variance: A GARCH-MIDAS approach. *Journal of Forecasting*, 32(7), 600–612. [CrossRef]

- Barndorff-Nielsen, O. E., & Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B*, 64(2), 253–280. [CrossRef]
- Barndorff-Nielsen, O. E., & Shephard, N. (2004). Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics*, 2(1), 1–37. [CrossRef]
- Baumol, W. J. (1952). The transactions demand for cash: An inventory theoretic approach. *Quarterly Journal of Economics*, 66(4), 545–556. [CrossRef]
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327. [CrossRef]
- Borovykh, A., Bohte, S., & Oosterlee, C. W. (2017). Conditional time series forecasting with convolutional neural networks. *arXiv*. [CrossRef]
- Brave, S. A., & Butters, R. A. (2012). Diagnosing the financial system: Financial conditions and financial stress. *International Journal of Central Banking*, 8(2), 191–239.
- Cao, J., Li, Z., & Li, J. (2019). Financial time series forecasting model based on CEEMDAN and LSTM. *Physica A: Statistical Mechanics and Its Applications*, 519, 127–139. [CrossRef]
- Chen, S., & Ge, L. (2019). Exploring the attention mechanism in LSTM-based Hong Kong stock price movement prediction. *Quantitative Finance*, 19(9), 1507–1515. [CrossRef]
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *The 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). ACM. [CrossRef]
- Chun, D., Cho, H., & Ryu, D. (2025). Volatility forecasting and volatility-timing strategies: A machine learning approach. *Research in International Business and Finance*, 75, 102723. [CrossRef]
- Clark, T. E., & McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105(1), 85–110. [CrossRef]
- Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2), 174–196. [CrossRef]
- Dechow, P. M., Kothari, S. P., & Watts, R. L. (1998). The relation between earnings and cash flows. *Journal of Accounting and Economics*, 25(2), 133–168. [CrossRef]
- Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366), 427–431. [CrossRef]
- Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263. [CrossRef]
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4), 987–1007. [CrossRef]
- Engle, R. F., Ghysels, E., & Sohn, B. (2013). Stock market volatility and macroeconomic fundamentals. *Review of Economics and Statistics*, 95(3), 776–797. [CrossRef]
- Engle, R. F., & Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2), 251–276. [CrossRef]
- Ersin, Ö. Ö., & Bildirici, M. (2023). Financial volatility modeling with the GARCH-MIDAS-LSTM approach: The effects of economic expectations, geopolitical risks and industrial production during COVID-19. *Mathematics*, 11(8), 1785. [CrossRef]
- European Commission. (2019). *Ethics guidelines for trustworthy AI*. Available online: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> (accessed on 1 January 2025).
- Ferreira, I. H., & Medeiros, M. C. (2021). Modeling and forecasting intraday market returns: A machine learning approach. *arXiv*. [CrossRef]
- Financial Accounting Standards Board. (2021). *Accounting standards update no. 2021-04: Error correction*. FASB.
- Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 1126–1135). PMLR.
- Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), 654–669. [CrossRef]
- Gao, H., Kou, G., Liang, H., Zhang, H., Chao, X., Li, C., & Dong, Y. (2024). Machine learning in business and finance: A literature review and research opportunities. *Financial Innovation*, 10, 86. [CrossRef]
- Goodell, J. W., Kumar, S., Lim, W. M., & Pattnaik, D. (2021). Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance*, 32, 100577. [CrossRef]
- Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10), 2222–2232. [CrossRef]
- Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2), 357–384. [CrossRef]

- Harikumar, Y., & Muthumeenakshi, M. (2025). An innovative study on stock price prediction for investment decision through ARIMA and LSTM with recurrent neural network. *New Mathematics and Natural Computation*, 21(3), 763–783. [CrossRef]
- Harvey, D. I., Leybourne, S. J., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2), 281–291. [CrossRef]
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. [CrossRef]
- Huang, A. H., Leavy, R., Zang, A. Y., & Zheng, R. (2018). Analyst information discovery and interpretation roles: A topic modeling approach. *Management Science*, 64(6), 2833–2855. [CrossRef]
- Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. Available online: <https://otexts.com/fpp3/> (accessed on 1 January 2025).
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795. [CrossRef]
- Kim, M., & Kross, W. (2005). The ability of earnings to predict future operating cash flows has been increasing—Not decreasing. *Journal of Accounting Research*, 43(5), 753–774. [CrossRef]
- Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review. *Expert Systems with Applications*, 197, 116659. [CrossRef]
- Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A*, 379(2194), 20200209. [CrossRef]
- Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4), 1187–1230. [CrossRef]
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 4765–4774). Curran Associates Inc.
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1), 54–74. [CrossRef]
- Marcellino, M., Stock, J. H., & Watson, M. W. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics*, 135(1–2), 499–526. [CrossRef]
- Miller, M. H., & Orr, D. (1966). A model of the demand for money by firms. *Quarterly Journal of Economics*, 80(3), 413–435. [CrossRef]
- Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59(2), 347–370. [CrossRef]
- Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2023). A time series is worth 64 words: Long-term forecasting with transformers. *arXiv*. [CrossRef]
- O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41(5), 673–690. [CrossRef]
- Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. *Journal of Business*, 53(1), 61–65. [CrossRef]
- Patton, A. J., & Sheppard, K. (2009). Optimal combinations of realised volatility estimators. *International Journal of Forecasting*, 25(2), 218–238. [CrossRef]
- Pesaran, M. H., & Timmermann, A. (1995). Predictability of stock returns: Robustness and economic significance. *The Journal of Finance*, 50(4), 1201–1228. [CrossRef]
- Petropoulos, F., & Spiliotis, E. (2025). Judgmental selection of parameters for simple forecasting models. *European Journal of Operational Research*, 323(4), 780–794. [CrossRef]
- Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9, 371–421.
- Shi, X. J., Chen, Z. R., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, 28, 802–810.
- Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460), 1167–1179. [CrossRef]
- Sun, Y., Liu, L., Xu, Y., Zeng, X., Shi, Y., Hu, H., Jiang, J., & Abraham, A. (2024). Alternative data in finance and business: Emerging applications and theory analysis (review). *Financial Innovation*, 10, 127. [CrossRef]
- Taneva-Angelova, G., & Granchev, D. (2025). Deep learning and transformer architectures for volatility forecasting: Evidence from U.S. equity indices. *Journal of Risk and Financial Management*, 18(12), 685. [CrossRef]
- Tauchen, G., & Zhou, H. (2011). Realized jumps on financial markets and predicting credit spreads. *Journal of Econometrics*, 160(1), 102–118. [CrossRef]
- U.S. Department of the Treasury. (2024). *Artificial intelligence in financial services, report on the uses, opportunities, and risks of artificial intelligence in the financial services sector*, 36. U.S. Department of the Treasury.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
- Virk, N., Javed, F., Awartani, B., & Hyde, S. (2024). A reality check on the GARCH-MIDAS volatility models. *European Journal of Finance*, 30(6), 575–596. [CrossRef]
- West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, 64(5), 1067–1084. [CrossRef]

- Yang, D., & Zhang, Q. (2000). Drift-independent volatility estimation based on high, low, open, and close prices. *Journal of Business*, 73(3), 477–491. [[CrossRef](#)] [[PubMed](#)]
- Zellner, A., & Hong, C. (1989). Forecasting international growth rates using Bayesian shrinkage and other procedures. *Journal of Econometrics*, 40(1), 183–202. [[CrossRef](#)]
- Zeng, Z., Kaur, R., Siddagangappa, S., Rahimi, S., Balch, T., & Veloso, M. (2023). Financial time series forecasting using CNN and transformer. *arXiv*. [[CrossRef](#)]
- Zhang, L., Mykland, P. A., & Ait-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100(472), 1394–1411. [[CrossRef](#)]
- Zhang, Y., Zhang, T., & Hu, J. (2025). Forecasting stock market volatility using CNN-BiLSTM-attention model with mixed-frequency data. *Mathematics*, 13(11), 1889. [[CrossRef](#)]
- Zhang, Z., Chen, B., Zhu, S., & Langrené, N. (2025). Quantformer: From attention to profit with a quantitative transformer trading strategy. *arXiv*. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.