Universiteit Leiden

# Master Computer Science

Comparing Classical and Foundation Models for Stock Price Forecasting


| | |
|---|---|
| Name: | Dimitrios Kourntidis |
| Student ID: | 3841863 |
| Date: | 13/05/2026 |
| Specialisation: | Artificial Intelligence |
| 1st supervisor: | Dr. Marc Hilbert |
| 2nd supervisor: | Dr. Yingjie Fan |


Master's Thesis in Computer Science


Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

# Acknowledgments

I would like to express my gratitude to my supervisor, Dr. Marc Hilbert, for his guidance, support, and encouragement throughout the course of this thesis. His perspective on research helped me remain motivated, particularly during moments when the results did not align with my initial expectations. He reminded me that scientific contribution is not solely about achieving state-of-the-art performance, but also about understanding, analyzing, and discussing why certain outcomes occur. This mindset greatly shaped both my approach to research and the way I view scientific progress.

I would also like to thank my friends and family for their continuous support during this journey. Whether through exchanging ideas, listening to my thoughts and concerns, or simply being present during difficult periods, their support meant more than they probably realize. Their patience and encouragement helped me maintain perspective and motivation throughout the completion of this work.

**Abstract**

This thesis investigates whether time-series foundation models improve stock price forecasting relative to traditional statistical baselines. Using Microsoft Corporation (MSFT) as a single-stock case study, it compares Naïve and ARIMA with TimesFM, Chronos, and Sundial across 2,400 out-of-sample rolling-forecast experiments. The evaluation spans daily and hourly frequencies, forecasting horizons from one day to 63 trading days, price-level and return targets, and three market disruption periods, using error metrics and directional accuracy. The results show that performance is regime-dependent rather than uniformly better for foundation models. Naïve and ARIMA remain strongest for short-horizon price forecasts and during the 2022 downturn, the least leakage-prone crisis window. Foundation models offer their clearest gains at longer horizons and under short training windows: TimesFM achieves the lowest 21-day MAPE (2.87% versus 3.06% for Naïve), while Chronos achieves the strongest 63-day result (5.67% versus 7.17%) and remains stable when ARIMA fails to converge under very limited data. Overall, foundation models complement rather than replace traditional baselines in this MSFT setting.

# Contents

# Chapter 1

# Introduction

The ability to accurately forecast stock prices remains a central objective in financial research, with direct implications for investment strategies, portfolio management, and risk assessment in volatile markets. The complexity of financial markets, characterized by non-linear dynamics, regime shifts, and sensitivity to exogenous events, has motivated the development of increasingly sophisticated forecasting methodologies over the past decades.

Traditional statistical approaches, such as ARIMA (Ariyo, Adewumi O. Adewumi, and Ayo, 2014), have been extensively applied to financial time-series forecasting and have proven effective at capturing linear dependencies and short-term patterns. However, their performance degrades when confronted with complex, non-linear relationships inherent in financial data (Sezer, Gudelek, and Ozbayoglu, 2020). To address these limitations, deep learning approaches such as Long Short-Term Memory Networks (LSTMs) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Units (GRUs) (Cho et al., 2014) have been adopted, demonstrating improved short-term predictive accuracy (Adebiyi, Aderemi O. Adewumi, and Ayo, 2014). Despite their promise, these methods frequently require substantial hyperparameter tuning, architectural design expertise, and large volumes of domain-specific training data, presenting barriers for practitioners and organizations without a strong machine learning background.

Recent advances in foundation models have introduced a new paradigm for time-series forecasting. Foundation models, large-scale models pre-trained on a lot and diverse data, have achieved remarkable success in natural language processing and computer vision through their capacity for generalization and few-shot (in-context) learning (Zhou et al., 2024). This success has motivated the development of foundation models specifically designed for time-series tasks, including Chronos (Ansari, Stella, et al., 2024), TimesFM (Das et al., 2024), and Sundial (Yong Liu, Qin, et al., 2025). These models offer the potential for zero-shot forecasting, promising faster deployment workflows and broader accessibility compared to traditional machine learning pipelines. However, while these models have demonstrated strong performance on general time-series benchmarks, their effectiveness for financial time-series forecasting, a domain characterized by low signal-to-noise ratios, non-stationarity, and extreme events, remains insufficiently explored.

## 1.1 Research Gap

Despite the growing body of literature on both traditional forecasting methods and time-series foundation models, rigorous evaluation of these models for financial applications remains limited. The benchmarks commonly used to evaluate foundation models, such as the Monash Time Series Forecasting Archive (Godahewa et al., 2021), predominantly feature datasets from energy, transport, and retail domains, with financial time series notably underrepresented (Xiangfei Liu et al., 2024).

Recent work has begun to address this gap. Rahimikia, Ni, and W. Wang (2025) present the first

comprehensive evaluation of time-series foundation models in global financial markets, finding that off-the-shelf models perform poorly compared to those pre-trained on financial data. However, their study focuses on daily excess returns with fine-tuning regimes, leaving open questions about zero-shot performance across different forecast horizons, temporal frequencies, and market conditions.

Specifically, no existing study systematically compares foundation models against traditional statistical baselines across the critical dimensions that define real-world financial forecasting scenarios: varying forecast horizons (from single-day to quarterly), different temporal frequencies (daily versus intraday), limited training data availability, and periods of major market disruption. Understanding these comparisons is essential for practitioners seeking to determine whether foundation models offer practical advantages over established methods for stock price prediction.

## 1.2 Research Questions

This thesis investigates the comparative performance of traditional statistical methods and time-series foundation models for stock price prediction. The study is guided by three research questions:

- **RQ1 – Forecast Horizon Length Effects:** *How does forecast horizon length affect the relative performance of time-series foundation models compared to traditional approaches for stock price forecasting?*
- **RQ2 – Temporal Frequency Effects:** *How does the temporal frequency of input data affect the relative performance of foundation models compared to traditional approaches for stock price forecasting?*
- **RQ3 – Market Disruption Resilience:** *How do foundation models perform relative to traditional approaches during major market disruptions?*

These research questions collectively examine whether the generalization capabilities of foundation models translate into measurable forecasting advantages under the specific conditions and challenges of financial time-series data.

## 1.3 Scope and Approach

This thesis uses a single-stock case study based on Microsoft Corporation (MSFT) historical price data, with the full data source and temporal splits detailed in Chapter 4. MSFT was selected because its trading history spans the key market regimes examined in this study, including the 2008 Global Financial Crisis, the 2020 COVID-19 crash, and extended periods of both bull and bear markets, enabling consistent evaluation across all experimental configurations. As a continuously traded S&P 500 constituent with high liquidity, it provides a clean testbed for comparing forecasting approaches without data quality concerns. Furthermore, as one of the most widely analyzed equities globally, MSFT is highly likely to appear in the diverse datasets used to pre-train time-series foundation models, making it an appropriate test case for evaluating whether foundation models can leverage such prior exposure to outperform traditional baselines. The experimental framework consists of 2,400 individual experiments spanning three training configurations, four forecast horizons (1, 5, 21, and 63 days ahead), two temporal frequencies (daily and hourly), two data representations (raw prices and fractional change), and three major market disruption periods (the 2008 Global Financial Crisis, the 2020 COVID-19 crash, and the 2022 market downturn). The study compares two traditional baselines (Naïve and ARIMA) against three foundation models (Chronos, TimesFM, and Sundial) using a rolling forecast methodology that ensures strictly out-of-sample evaluation.

The single-stock design is acknowledged as a limitation, however, it enables comprehensive investigation of the interaction effects between model type, horizon length, data availability, temporal frequency, and

market regime, providing methodological depth that would be difficult to achieve with a broader but shallower multi-stock analysis. Extension to multiple stocks and markets is positioned as future work.

## 1.4 Contributions

This thesis makes the following contributions:

1. A systematic empirical comparison of three time-series foundation models against traditional statistical baselines for stock price forecasting, evaluating performance across multiple dimensions.
2. An investigation of forecast horizon effects on model performance, including analysis of how training data availability interacts with horizon length through logarithmically and linearly spaced training period configurations.
3. An analysis of temporal frequency effects, comparing model performance on daily versus hourly data and assessing foundation models' zero-shot frequency adaptation capabilities.
4. An evaluation of model resilience during three major market disruptions, the 2008 Global Financial Crisis, the 2020 COVID-19 market crash, and the 2022 downturn, providing insights into foundation model robustness under extreme market conditions. (Note: Foundation models may have encountered these historical periods during pre-training, creating an asymmetric comparison with traditional models trained only on pre-crisis data. This limitation is explicitly acknowledged in the analysis.)
5. An assessment of the impact of data transformation (fractional change) on forecasting accuracy across both traditional and foundation model approaches.

## 1.5 Thesis Outline

The remainder of this thesis is organized as follows. Chapter 2 reviews the relevant literature on traditional time-series forecasting methods, deep learning approaches, and the emerging field of time-series foundation models. Chapter 3 presents the theoretical foundations and descriptions of the forecasting models and evaluation metrics employed in this study. Chapter 4 details the experimental design, including the dataset, rolling forecast methodology, and the three experimental configurations addressing each research question. Chapter 5 presents the experimental results organized by research question. Chapter 6 discusses the findings and their implications, including the study's limitations and directions for future work. Chapter 7 concludes the thesis with a summary of the main findings and final remarks.

# Chapter 2

# Literature Review

Stock price forecasting has attracted sustained academic interest due to its implications for investment decision-making and risk management. The field has evolved from classical statistical methods through deep learning approaches to the recent emergence of foundation models. This chapter reviews the literature needed for the comparison: market efficiency and predictability, ARIMA and deep learning approaches, time-series foundation models, the role of Naïve baselines, and the research gap addressed by this thesis.

## 2.1 Market Efficiency and Stock Predictability

The question of whether stock prices can be predicted is fundamentally linked to the Efficient Market Hypothesis (EMH), formalised by Fama (1970) and popularised by Malkiel (1973). In its weak form, the EMH posits that historical price data cannot be used to predict future returns, as past information is already incorporated into current prices. This theoretical position implies that any forecasting model should, at best, perform no better than a naive baseline that simply predicts tomorrow's price equals today's price.

However, the empirical literature presents a more nuanced picture. Fama and French (1996) documented anomalies such as size and value effects that appear inconsistent with strict market efficiency, suggesting that exploitable patterns may exist under certain conditions. The existence of such anomalies does not invalidate the EMH as a working hypothesis but instead it reframes the question from *whether* prices are predictable in principle to *whether* a model's forecasts improve on the naive persistence forecast (tomorrow's price equals today's). Under weak-form efficiency, historical prices should not contain usable information for extrapolating future returns, which makes the naive forecast the central benchmark rather than a trivial comparison.

This framing supplies the methodological anchor for the present thesis: any forecasting model must demonstrate performance significantly better than a naive baseline to claim genuine predictive value. The reproducibility study of Buczyński et al. (2023), covering 15 deep learning approaches, found that most claimed improvements fail to hold under rigorous evaluation, providing empirical confirmation that the naive baseline is a load-bearing, not a ceremonial, comparator.

## 2.2 Traditional Statistical Approaches

The ARIMA model, formalized by Box and Jenkins (1976), remains the foundational approach for univariate time series forecasting (its mathematical formulation is detailed in Chapter 3). The application

of ARIMA to stock price forecasting has yielded mixed results in the literature. Ariyo, Adewumi O. Adewumi, and Ayo (2014) demonstrated that properly specified ARIMA models can achieve reasonable short-term forecasting accuracy on stock exchange data from both the New York Stock Exchange and Nigeria Stock Exchange. Similarly, Fattah et al. (2018) showed the effectiveness of ARIMA for demand forecasting applications.

However, comparative studies consistently identify limitations of ARIMA for stock forecasting. Adebiyi, Aderemi O. Adewumi, and Ayo (2014) found that while ARIMA performs reasonably well for short-term predictions, neural network approaches can achieve superior accuracy, particularly for capturing nonlinear patterns in financial data.

The literature also identifies fundamental theoretical constraints. Statistical models like ARIMA are effective under linear conditions but struggle with the nonlinear complexity of financial markets (Sezer, Gudelek, and Ozbayoglu, 2020). The assumption of constant variance is violated by the volatility clustering characteristic of financial time series, first formalised through the ARCH framework (Engle, 1982) and later generalised to GARCH (Bollerslev, 1986). The stationarity requirement likewise poses challenges for price series exhibiting trends and structural breaks (Yong Liu, Wu, et al., 2022).

Sezer, Gudelek, and Ozbayoglu (2020) conducted a systematic literature review of financial time series forecasting with deep learning from 2005–2019, documenting a significant increase in the use of deep learning approaches while noting that ARIMA models continue to serve as essential baseline benchmarks against which more complex methods are evaluated.

## 2.3 Deep Learning Approaches

### 2.3.1 Recurrent Neural Networks for Sequential Data

The limitations of linear statistical models motivated the application of neural networks to financial forecasting. Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) and the related Gated Recurrent Unit (GRU) (Cho et al., 2014) were designed to capture long-range dependencies in sequential data, making them natural candidates for time series forecasting tasks.

### 2.3.2 LSTM in Stock Prediction

Deep learning methods, particularly LSTMs, have been extensively applied to stock prediction with generally positive reported results, though as the reproducibility discussion below makes clear, the strength of these reports must be weighed against well-documented replication concerns. An influential early demonstration is Fischer and Krauss (2018), who applied LSTM networks to predict out-of-sample directional movements for S&P 500 constituents between 1992 and 2015, reporting daily returns that exceeded memory-free classification methods prior to transaction costs. Phuoc et al. (2024) applied LSTM models with technical indicators (SMA, MACD, RSI) to Vietnamese stock data, reporting high directional accuracy, though their evaluation was limited to a single market without comparison to naive baselines. S. Wang (2023) proposed a hybrid BiLSTM-Transformer architecture achieving improvements over baseline methods on Chinese stock data, but the baselines considered were themselves neural architectures rather than a naive or statistical benchmark, leaving the naive-forecast comparison unaddressed. These studies exemplify a broader pattern in the LSTM-for-stocks literature: single-market evaluations, neural-only baseline sets, and metrics that can mask a failure to beat the trivial predictor.

Al-Alawi and Alaali (2023) conducted a literature review of machine learning techniques for stock market prediction, examining 12 research papers across banking and healthcare sectors. Their findings indicate that LSTM and GRU techniques provide the best results in most of the selected papers regardless of the sector in which they are applied.

Beyond sequence models, Gu, Kelly, and Xiu (2020) conducted the most comprehensive empirical comparison of machine learning methods for asset pricing to date, evaluating ordinary least squares, generalised linear models, regression trees, random forests, and neural networks on a dataset of nearly 30,000 stocks over six decades. They found that machine learning methods, particularly shallow neural networks, delivered substantial gains in out-of-sample predictive $R^2$ over linear benchmarks for monthly return prediction. Their work established that the relative advantage of nonlinear methods is tied to the signal-to-noise regime and the feature set, a nuance often lost in single-market stock-prediction studies.

### 2.3.3 Reproducibility Concerns

Despite the positive results frequently reported, the literature reveals significant concerns about reproducibility and generalization. Buczyński et al. (2023) attempted to reproduce 15 deep learning approaches for financial time series forecasting, examining papers published over seven years. Their findings were revealing: "most of these approaches do not beat the naive approach, and only some barely beat it." The authors noted that most researchers did not provide sufficient data for full replication, and recommended rigorous evaluation including consistent test periods and comparison against appropriate baselines.

This reproducibility crisis underscores a critical point: the naive forecast, which simply predicts tomorrow's price equals today's price, represents a surprisingly difficult benchmark to beat in financial forecasting.

## 2.4 Time-Series Foundation Models

### 2.4.1 Foundation Model Paradigm

Before surveying foundation models specifically, it is useful to situate them within the broader evolution of deep learning for time series. Lim and Zohren (2021) provide a survey of this trajectory, tracing the progression from recurrent architectures through attention mechanisms and hybrid statistical–neural models. A pivotal architectural step in this lineage is PatchTST (Nie et al., 2023), which demonstrated that segmenting a univariate series into subseries-level patches and feeding them to a standard Transformer encoder yields state-of-the-art long-horizon accuracy while substantially reducing computational cost. The patch tokenisation idea introduced here is the direct architectural antecedent of the tokenisation strategies subsequently adopted by Chronos and TimesFM.

Foundation models represent a paradigm shift in machine learning, characterized by large-scale pretraining on diverse datasets followed by adaptation to specific tasks (Zhou et al., 2024). This approach, pioneered in natural language processing with models like GPT (Radford et al., 2018), has demonstrated remarkable capabilities for zero-shot and few-shot learning.

Ye et al. (2024) provide a comprehensive survey of time series foundation models, identifying two main research directions: pre-training foundation models from scratch on time series data, and adapting large language models for time series tasks. Liang et al. (2024) offer a complementary tutorial and survey examining model architectures, pre-training techniques, and adaptation methods, noting that while progress has been rapid, significant challenges remain in handling the heterogeneity of time series data across domains.

### 2.4.2 Chronos

Chronos (Ansari, Stella, et al., 2024) frames time series forecasting as a language modeling task by tokenizing continuous values into a discrete vocabulary (its architecture is detailed in Chapter 3). In the original benchmark comprising 42 datasets, Chronos demonstrated two key findings: (a) significant

outperformance on datasets included in the training corpus, and (b) comparable and occasionally superior zero-shot performance on new datasets relative to methods trained specifically on them. The authors concluded that "Chronos models can leverage time series data from diverse domains to improve zero-shot accuracy on unseen forecasting tasks" (Ansari, Stella, et al., 2024).

### 2.4.3 TimesFM

TimesFM (Das et al., 2024) is a decoder-only foundation model developed by Google, built on a patched-decoder Transformer that ingests input patches and autoregressively produces output patches of arbitrary length, allowing it to generalise across different forecasting history lengths, prediction lengths, and temporal granularities (its architecture is detailed in Chapter 3). Pre-trained on a corpus of 100 billion real-world timepoints combined with synthetic series, the model was evaluated across a broad set of public forecasting benchmarks. The authors report two central findings: (a) despite being considerably smaller than many language-model-based approaches, the 200M-parameter TimesFM attains near state-of-the-art zero-shot accuracy relative to supervised baselines specifically trained on each target dataset, and (b) performance scales favourably with model size and with the diversity of the pre-training corpus. They conclude that a pre-trained decoder-only model can "perform close to the state-of-the-art supervised forecasting models" (Das et al., 2024) across heterogeneous public datasets in a zero-shot setting.

### 2.4.4 Sundial

Sundial (Yong Liu, Qin, et al., 2025) represents a family of native time-series foundation models that, unlike Chronos and TimesFM, do not adapt language-model architectures but instead employ a generative approach based on flow matching, producing continuous-valued probabilistic forecasts without the discrete tokenisation step used by token-based models (its architecture is detailed in Chapter 3). Pre-trained on an aggregated corpus exceeding one trillion real-world timepoints, Sundial was evaluated across point and probabilistic forecasting benchmarks spanning multiple domains and frequencies. The authors report two key findings: (a) state-of-the-art zero-shot results on both point and probabilistic forecasting benchmarks relative to prior foundation models, and (b) training the same architecture from scratch on the aggregated datasets yields inferior performance compared to using the pre-trained checkpoints, implying that genuine knowledge transfer occurs through large-scale pre-training rather than simple memorisation of target distributions. They conclude that Sundial provides "highly capable foundation models" for general-purpose time-series forecasting (Yong Liu, Qin, et al., 2025).

### 2.4.5 Other Notable Models

The broader landscape of time series foundation models continues to expand rapidly. Lag-Llama (Rasul et al., 2024) demonstrated strong zero-shot generalization and state-of-the-art performance when fine-tuned on small fractions of target datasets. Moirai-MoE (Xu Liu et al., 2024) introduced sparse mixture of experts for cross-frequency learning. MOMENT (Goswami et al., 2024) contributed the Time Series Pile, a large-scale pre-training dataset, while Tiny Time Mixers (Ekambaram et al., 2024) showed that competitive performance is achievable with substantially fewer parameters, starting from just 1M. TimeGPT (Garza, Challu, and Mergenthaler-Canseco, 2024) further demonstrated that zero-shot inference can rival established statistical and deep learning methods on general benchmarks. Unified evaluations such as TSFM-Bench (Xiangfei Liu et al., 2024) place these models side-by-side on common protocols, providing broader empirical support for their general-domain competitiveness.

### 2.4.6 Foundation Models in Financial Applications

While foundation models have demonstrated impressive performance on general time series benchmarks, their application to financial data reveals important limitations. Rahimikia, Ni, and W. Wang (2025)

present the first comprehensive evaluation of time-series foundation models in global financial markets. Their findings are cautionary: "off-the-shelf pre-trained Time Series Foundation Models perform weakly in zero-shot forecasting of daily excess returns, underperforming strong ensemble models such as CatBoost and LightGBM."

Their study found that for both Chronos and TimesFM families, performance gradually improves with larger models and wider context windows, but even with fine-tuning on financial data, the models "fail to close the performance gap with benchmarks" (Rahimikia, Ni, and W. Wang, 2025). This suggests that the distinctive characteristics of financial time series (low signal-to-noise ratio, volatility clustering, regime changes) may limit transfer from general-domain pre-training.

## 2.5 Naïve Baselines

Sections 2.1 and 2.3 already flagged the difficulty of beating the naive forecast in financial settings. The same pattern appears in general-domain evidence, which reinforces that the concern is not peculiar to finance. The M4 competition (Makridakis, Spiliotis, and Assimakopoulos, 2020), comprising 100,000 series across multiple frequencies, found that pure machine learning methods systematically underperformed well-specified statistical models, with the winning submissions being hybrids that combined statistical decomposition with neural components. The subsequent M5 competition (Makridakis, Spiliotis, and Assimakopoulos, 2022) on hierarchical retail demand data reached a similar verdict, with simple exponential smoothing and gradient-boosted trees remaining competitive against substantially more complex architectures. Taken together with Buczyński et al. (2023), these results frame the standard that foundation-model claims in this thesis are held to: outperforming the naive persistence forecast, not merely other neural models.

## 2.6 Research Gaps

The literature reviewed above leaves several unresolved issues that motivate this thesis:

**Gap 1: Limited thorough evaluation of foundation models on financial data.** While foundation models have been extensively benchmarked on general time-series datasets (Godahewa et al., 2021; Xiangfei Liu et al., 2024), systematic evaluation on financial time series with proper baselines remains sparse. The findings of Rahimikia, Ni, and W. Wang (2025) suggest that zero-shot performance may be substantially weaker in financial domains, but their study focused on daily excess returns with fine-tuning regimes, leaving open questions about other configurations. This motivates a dedicated evaluation of time-series foundation models on raw stock-price data.

**Gap 2: Insufficient analysis across forecast horizons.** Although foundation models such as TimesFM claim generalisation across prediction lengths by design (Das et al., 2024), and the broader deep-learning forecasting literature has long flagged that relative model performance is horizon-dependent (Lim and Zohren, 2021), existing studies in finance typically evaluate a single or limited set of forecast horizons. Understanding how the relative performance of foundation models versus traditional approaches varies from short-term (1-day) to longer-term (quarterly) horizons is essential for practical deployment decisions.

**Gap 3: Limited investigation of temporal frequency effects.** General-domain benchmarks such as the M4 competition have shown that model rankings shift markedly across sampling frequencies (Makridakis, Spiliotis, and Assimakopoulos, 2020), and recent foundation models have begun explicitly targeting cross-frequency generalisation (Xu Liu et al., 2024). Yet most foundation-model evaluations in finance focus on daily data. Whether these models can effectively handle intraday financial series, and whether their relative advantage over traditional methods varies with frequency, remains an open question.

**Gap 4: Sparse evaluation during market disruptions.** Financial time series are well known to

exhibit volatility clustering and regime changes, yet model performance during extreme market events, precisely when accurate forecasting is most valuable, is rarely examined. The 2008 Global Financial Crisis, the 2020 COVID-19 crash, and the 2022 downturn represent critical test cases for understanding model robustness under stress conditions, and are additionally informative given that foundation-model pre-training corpora may overlap with the 2008 and 2020 episodes but not with 2022.

**Gap 5: Inconsistent baseline comparison.** Many studies compare foundation models only against other neural approaches without adequate comparison to naive baselines and well-specified statistical models. The Makridakis competitions (Makridakis, Spiliotis, and Assimakopoulos, 2020; Makridakis, Spiliotis, and Assimakopoulos, 2022) demonstrated in the general domain that simple statistical baselines remain difficult to beat, and the financial-specific reproducibility study of Buczyński et al. (2023) reached the same conclusion. Rigorous baseline comparison is therefore essential.

This thesis addresses these gaps through a systematic empirical comparison of three foundation models (Chronos, TimesFM, Sundial) against traditional baselines (Naïve, ARIMA) for stock price forecasting. The experimental design, detailed in Chapter 4, uses rolling forecast methodology with strictly out-of-sample evaluation across multiple forecast horizons (1, 5, 21, and 63 days), two temporal frequencies (daily and hourly), and explicit testing during the 2008, 2020 and 2022 market crises.

# Chapter 3

# Methodology

This chapter defines the five forecasting models used in the experiments: Naïve and ARIMA as traditional baselines, and TimesFM, Chronos, and Sundial as zero-shot foundation models. For the traditional methods, the focus is mathematical formulation. For the foundation models, the focus is architecture and inference behavior.

## 3.1 Naïve Forecasting

The naïve method is the simplest forecasting baseline: it assumes that the forecast for any future period equals the most recently observed value. This is formally expressed in Equation (3.1):

$$\hat{y}_{t+h|t} = y_t \tag{3.1}$$

Here, $\hat{y}_{t+h|t}$ represents the forecast for $h$ periods into the future, made at time $t$ based on information available at time $t$. The term $y_t$ denotes the actual observed value of the target variable at time $t$. The forecast horizon $h$ can be any positive integer ($h = 1, 2, \dots$).

Despite its simplicity, the naïve forecast is very important as a baseline model in time series analysis. More complex forecasting methods are often evaluated against the performance of the naïve forecast. A sophisticated model is generally considered useful only if it consistently demonstrates superior accuracy compared to this baseline, typically measured using standard error metrics.

The naïve method is particularly relevant for financial data, where the random walk hypothesis (Malkiel, 1973) suggests that future price changes are unpredictable based on historical prices alone, making the last observed price a theoretically grounded baseline estimate.

## 3.2 ARIMA (Autoregressive Integrated Moving Average)

Autoregressive Integrated Moving Average (ARIMA) models are a class of statistical models used for analyzing and forecasting time series data (Box and Jenkins, 1976). They are particularly effective when non-stationarity can be addressed through differencing, most commonly when the level or trend of the series changes over time. The ARIMA model combines three components, denoted ARIMA$(p, d, q)$:

- **Autoregressive (AR), order $p$:** Models linear dependence on the previous $p$ values of the series.

- **Integrated (I), degree $d$:** Differences the series $d$ times to achieve stationarity (e.g., $Y'_t = Y_t - Y_{t-1}$ for $d = 1$).
- **Moving Average (MA), order $q$:** Models dependence on the previous $q$ forecast errors, capturing short-run shocks.

When $d > 0$, the AR and MA components are applied to the differenced series. For instance, an ARIMA$(p, 1, q)$ model means that $Y'_t = Y_t - Y_{t-1}$ follows an ARMA$(p, q)$ process (Box and Jenkins, 1976):

$$Y'_t = c + \phi_1 Y'_{t-1} + \cdots + \phi_p Y'_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} \tag{3.2}$$

## 3.3 TimesFM

Beyond traditional statistical methods, this study uses TimesFM, a time series foundation model developed by Google Research (Das et al., 2024). As a pre-trained foundation model, TimesFM has been trained on an extensive corpus of 100 billion real-world time points from diverse domains (Das et al., 2024), enabling zero-shot forecasting on unseen datasets.

### 3.3.1 Architecture and Inference

The architecture of TimesFM is a decoder-only transformer, conceptually similar to those used in Large Language Models (LLMs) like GPT (Radford et al., 2018), but specifically adapted for time series data through input patching. As illustrated in Figure 3.1, the model does not process individual time points. Instead, it partitions the input time series of length $L$ into non-overlapping patches of a fixed input patch length ($p_{\text{in}} = 32$). These patches are treated as "tokens" and processed by transformer layers with causal self-attention, ensuring predictions depend only on past information.

A key design choice is that the output patch length ($p_{\text{out}} = 128$) exceeds the input patch length, enabling the model to predict multiple future timesteps simultaneously. The model is trained to minimize the Mean Squared Error (MSE) between predicted and actual future patches (Das et al., 2024). While TimesFM 2.0 provides experimental quantile heads, its released checkpoint is primarily designed for point forecasting. In this study, we therefore use the model's **point forecasts** for direct comparison with the other models.

For inference, the model takes an input context of length $L$ to generate a forecast for a specified horizon $H$:

$$f : (y_{1:L}) \longrightarrow \hat{y}_{L+1:L+H} \tag{3.3}$$

For horizons longer than the output patch length, the model operates auto-regressively, appending predicted values to its input context and repeating until the full forecast horizon is covered (Das et al., 2024).

## 3.4 Chronos

Chronos is a framework for pretrained probabilistic time series models developed by Amazon (Ansari, Stella, et al., 2024). As illustrated in Figure 3.2, the framework introduces a distinctive approach by tokenizing time series values through scaling and quantization, effectively transforming continuous temporal data into a discrete vocabulary. This transformation enables the training of existing transformer-based language model architectures on tokenized time series using a cross-entropy loss, fundamentally treating time series forecasting as a language modeling problem with minimal modifications to established architectures.

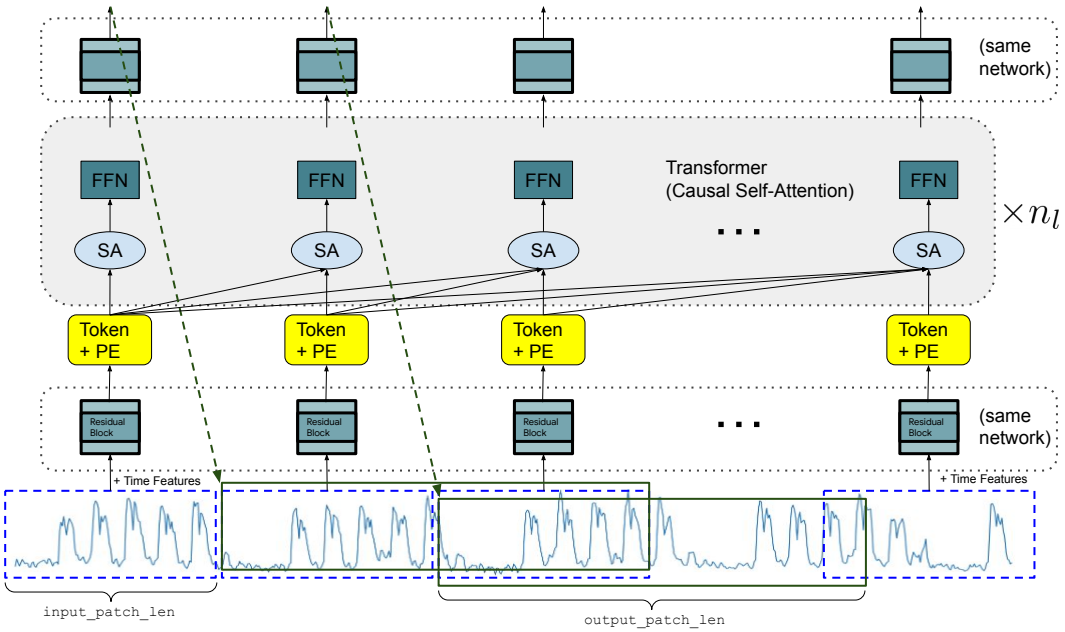


Figure 3.1: Illustration of the TimesFM model architecture. The input time-series is divided into patches, which are processed by residual blocks, combined with positional encoding (PE), and fed through stacked transformer layers (utilizing causal self-attention and feed-forward networks). The output is mapped via another set of residual blocks to predict future patches. Source: (Das et al., 2024).

### 3.4.1 Tokenization and Architecture

The core innovation of Chronos lies in treating time series forecasting as a language modeling problem. Real-valued observations are first normalized via mean absolute scaling and then quantized into discrete tokens drawn from a vocabulary of 4096 bins uniformly spaced across a fixed interval (Ansari, Stella, et al., 2024). This transforms continuous temporal data into a discrete sequence that can be processed by standard language model architectures. Notably, Chronos intentionally disregards explicit time and frequency information, relying on the model to implicitly learn temporal patterns from the sequential structure alone (Ansari, Stella, et al., 2024).

The original Chronos models utilize variants of the T5 encoder-decoder architecture (Raffel et al., 2020), ranging from 20M parameters (Mini) to 710M parameters (Large), trained with a cross-entropy loss over the token vocabulary. A subsequent release, Chronos-Bolt, retains the T5 encoder-decoder architecture but replaces autoregressive token-by-token decoding with direct multi-step forecasting. Instead of quantizing individual observations into discrete tokens, Chronos-Bolt chunks the historical time series into patches of multiple observations that are fed to the encoder. The decoder then generates quantile forecasts across multiple future steps in a single forward pass, making Chronos-Bolt up to 250× faster and 20× more memory-efficient than the original Chronos models (Ansari, Stella, et al., 2024; AWS Machine Learning Blog, 2025). To address the relative scarcity of public time series data, the Chronos framework augments its training corpus with synthetic time series generation (Ansari, Stella, et al., 2024). In this study, we use the Chronos-Bolt Base variant (205M parameters) (Ansari, Stella, et al., 2024).

### 3.4.2 Probabilistic Inference

During inference, Chronos-Bolt generates probabilistic forecasts by passing patched historical observations through the encoder and using the decoder to directly predict multiple quantiles of the forecast distribution in a single forward pass (Ansari, Stella, et al., 2024; AWS Machine Learning Blog, 2025). By predicting at multiple quantile levels simultaneously, the model constructs a predictive distribution that

captures forecast uncertainty, enabling the computation of prediction intervals.

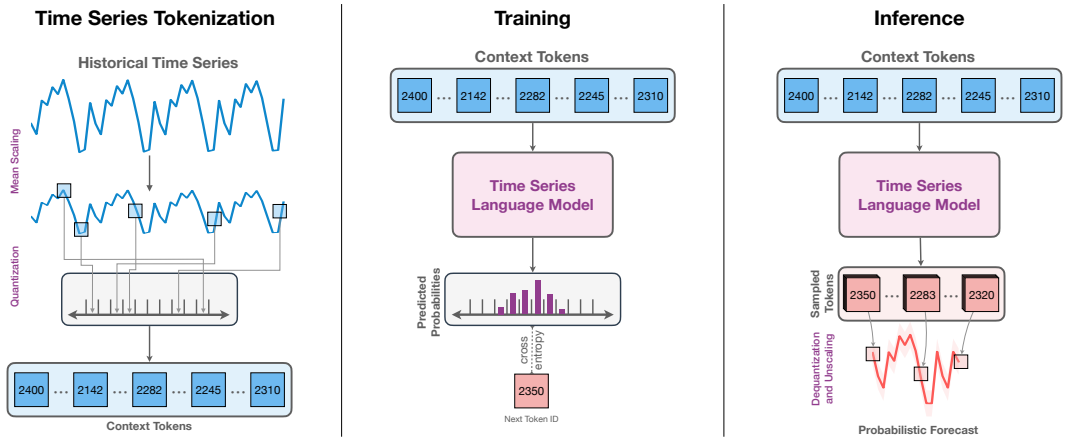


Figure 3.2: Illustration of the original Chronos framework. Real-valued time series are scaled and quantized into discrete tokens, which are processed by a transformer-based language model architecture (e.g., T5). The model outputs a categorical distribution over the token vocabulary, from which samples are dequantized and unscaled to produce probabilistic forecasts. Chronos-Bolt, used in this study, modifies this framework through patching and direct multi-step quantile forecasting. Source: (Ansari, Stella, et al., 2024).

## 3.5 Sundial

Sundial is a generative time series foundation model (Yong Liu, Qin, et al., 2025) that preserves original continuous values through patch-level tokenization, unlike the original Chronos model's discrete quantization (Ansari, Stella, et al., 2024). As illustrated in Figure 3.3, the architecture combines a patch-based tokenizer, a decoder-only Transformer backbone, and a flow-matching generative head. The model was pre-trained on TimeBench, a large-scale corpus comprising over one trillion time points from diverse domains.

### 3.5.1 Architecture

Sundial uses a two-stage tokenization process. First, each time series is normalized using a non-parametric instance normalization scheme (Yong Liu, Wu, et al., 2022) to handle non-stationarity and diverse scales. The normalized series is then divided into patches of length $P = 16$, which are projected into embeddings via a shared MLP. This patch-level approach reduces the effective sequence length for the Transformer.

The core backbone is a decoder-only Transformer with causal self-attention and Rotary Positional Embeddings (RoPE) (Su et al., 2024). The Sundial family is described in three configurations: Small (32M parameters), Base (128M parameters), and Large (444M parameters). In this study, we use the Base checkpoint.

### 3.5.2 Generative Inference

Sundial uses flow matching (Lipman et al., 2023) as its generative mechanism, learning a smooth transformation that gradually converts random Gaussian noise into realistic forecast values. During inference, the model starts from a sample of random noise and iteratively refines it through this learned transformation to produce a forecast. Since each run starts from different random noise, each produces a different plausible forecast trajectory, yielding **probabilistic forecasts**. Generating multiple such trajectories enables the construction of predictive distributions and uncertainty quantification through empirical quantiles.

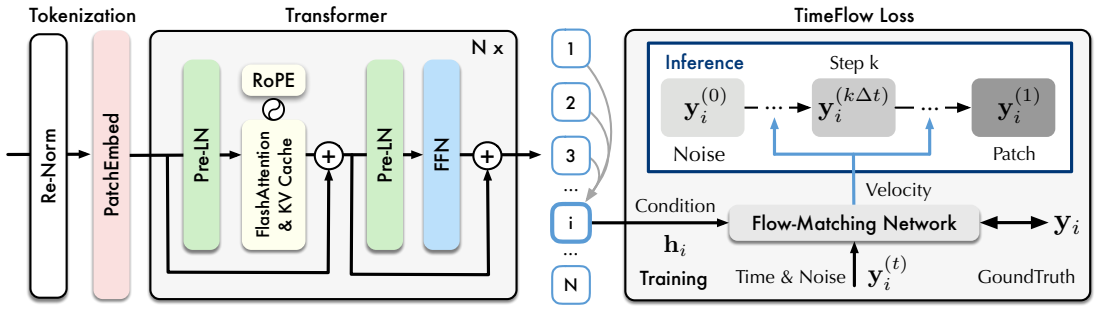


Figure 3.3: Illustration of the Sundial model architecture. The input time series is normalized and divided into patches, which are embedded and processed by a decoder-only Transformer with causal self-attention. The generative head uses flow matching to transform Gaussian noise into probabilistic forecasts through learned velocity fields. Source: (Yong Liu, Qin, et al., 2025).

## 3.6 Foundation Models: Summary

Table 3.1 provides a comparative overview of the three foundation models used in this study, highlighting their key architectural and methodological differences.

Table 3.1: Comparison of time series foundation model checkpoints used in this study.

| Aspect | TimesFM | Chronos-Bolt | Sundial |
|---|---|---|---|
| Developer | Google Research | Amazon | Tsinghua University |
| Architecture | Decoder-only | Encoder-decoder | Decoder-only |
| Training Data | $\mathcal{O}$(100B) time points + synthetic (Das et al., 2024) | 84B observations + synthetic (Ansari, Stella, et al., 2024) | 1T+ time points (Yong Liu, Qin, et al., 2025) |
| Output Type | Point forecast | Probabilistic quantiles | Probabilistic samples |
| Training Objective | MSE | Quantile loss | TimeFlow loss |
| Checkpoint Used | `google/timesfm-2.0-500m-pytorch` | `amazon/chronos-bolt-base` | `thuml/sundial-base-128m` |

The models represent diverse architectural philosophies: TimesFM adapts the decoder-only paradigm with patch-based inputs, Chronos-Bolt repurposes the T5 encoder-decoder architecture with patching and direct multi-step quantile generation, and Sundial uses generative flow matching for probabilistic forecasting. This diversity enables comprehensive evaluation of different foundation model approaches for financial time series forecasting.

# Chapter 4

# Experimental Design

This chapter defines the dataset, temporal splits, transformations, rolling forecast protocol, and experimental configurations used to answer the three research questions. It builds on the model definitions in Chapter 3 and specifies how forecast horizon length, temporal frequency, and market disruption resilience are evaluated.

## 4.1 Dataset and Temporal Split

### 4.1.1 Stock Selection Rationale

This study adopts a single-stock experimental design, using Microsoft Corporation (MSFT) as the sole subject of analysis. This deliberate depth-over-breadth approach enables comprehensive evaluation across 2,400 individual experiments spanning multiple forecast horizons, training lengths, temporal frequencies, data transformations, and market regimes. A multi-stock design of equivalent depth would be extremely time-consuming given the number of experimental configurations involved.

MSFT was selected because it provides a suitable and well-documented case study for the experimental design. First, it is a highly liquid large-cap equity, which reduces the possibility that the results are driven by sparse trading, missing observations, or other data-quality problems. Second, its long trading history provides enough observations to evaluate the same models across multiple market conditions, including the three disruption periods examined in this study. Third, the stock's behaviour and price changes over the years, from a weaker growth period in the 2000s to stronger performance after the mid-2010s.

It is important to acknowledge that findings from a single-stock study may not generalize to all equities. Stocks with different characteristics, such as lower liquidity, higher volatility, or different sector dynamics, may yield different relative model performance. Multi-stock validation across diverse market capitalizations, sectors, and geographies remains an important direction for future work.

### 4.1.2 Data Source and Characteristics

The historical stock price data for MSFT was obtained using the XM MetaTrader 5 platform[1]. The downloaded source data span approximately 22 years of trading history, from September 10, 2003 to November 12, 2025. For the empirical analysis, however, only observations through 2024 are used. Post-2024 observations are retained only as part of the downloaded source file and are not included in the reported experiments.

[1] https://www.xm.com/mt5

The dataset is analyzed at two temporal granularities:

- **Daily (1d) frequency**: 5,571 trading day records
- **Hourly (1h) frequency**: 37,530 hourly records during market hours

The target variable for all forecasting experiments is the closing price (`Close`), representing the last traded price for each time period.

Figure 4.1 presents the full daily closing price series, illustrating the train-test split and the three crisis periods examined in this study. The plot highlights the substantial price regime changes over the 22-year period: a prolonged low-price phase from 2003 to approximately 2014, followed by sustained growth that accelerated after 2019, with prices reaching above $400 by the test period.

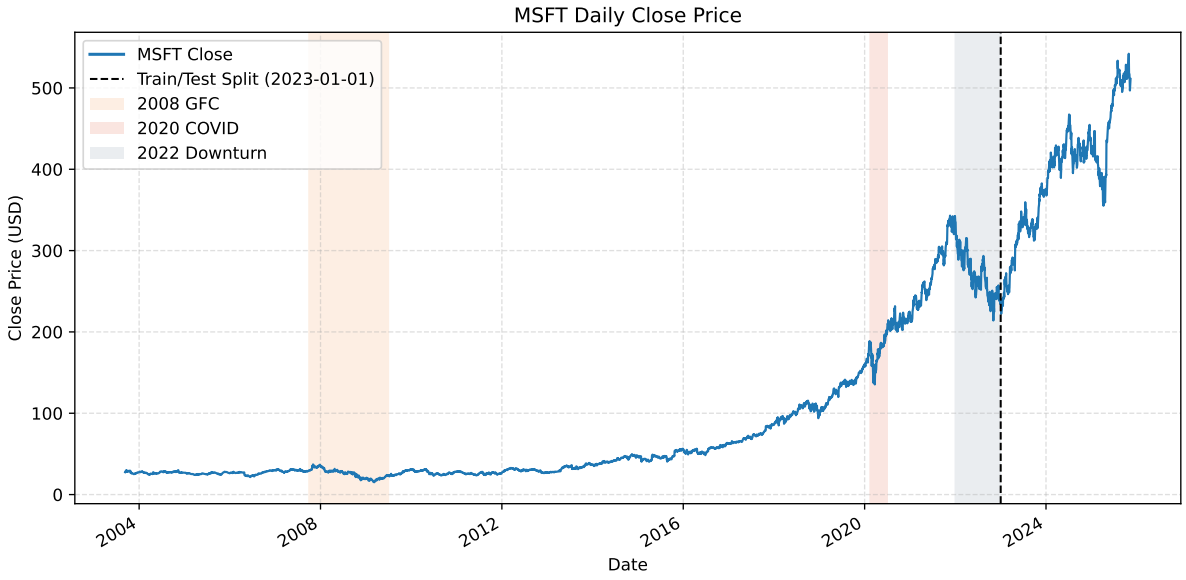


Figure 4.1: MSFT daily closing price from September 2003 to November 2025. The dashed vertical line marks the train-test split at January 1, 2023. Shaded regions indicate the three market disruption periods: the 2008 Global Financial Crisis, the 2020 COVID-19 crash, and the 2022 market downturn.

### 4.1.3 Train-Test Temporal Split

Each experimental configuration uses a clearly defined test period. For RQ1 and RQ2, a fixed temporal split is set at **January 1, 2023**, yielding:

- **Training period**: Variable length historical data before January 1, 2023
- **Test period**: One calendar year, from January 1, 2023 to December 31, 2023

For RQ3, each crisis event defines its own split: models are trained on pre-crisis data and tested over the crisis window itself, which spans approximately two months per event (exact boundaries specified in Section 4.3.3). While shorter than the one-year test period used for RQ1 and RQ2, these crisis windows are sufficient to capture the acute market dislocation and early recovery phases that are the focus of the resilience analysis.

Across all configurations, the variable training period length (detailed in Section 4.3) allows systematic investigation of how data availability impacts model performance.

### 4.1.4 Data Transformations

To assess model robustness under different data representations, the following transformation is applied as an alternative data representation:

**Fractional Change Transformation** Converts absolute price values to relative returns using the formula:

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}} \tag{4.1}$$

where $P_t$ is the closing price at time $t$ and $r_t$ is the fractional change. This transformation normalizes the data relative to previous values, making it scale-invariant. This transformation is particularly relevant for foundation models, as it may align better with the scale-invariant patterns learned during pretraining on diverse time series datasets. The transformation is applied to the entire time series before the train-test split to maintain temporal consistency.

Beyond scale invariance, the fractional change transformation also avoids data leakage, since each return depends only on two consecutive prices rather than on future information or global statistics such as the series mean or standard deviation.

A key motivation for applying the fractional change transformation is the Autocorrelation Function (ACF) structure of raw price data. As shown in Figures 4.2 and 4.4, raw closing prices exhibit near-unity autocorrelation across all lags at both daily and hourly frequencies. The Partial Autocorrelation Function (PACF) plots confirm that this structure is dominated by a single lag-1 dependency: each price is almost entirely explained by the immediately preceding value. In such a setting, even a naïve model that simply repeats the last observed price can achieve low error, making raw price forecasting a weak test of a model's ability to capture meaningful temporal patterns.

After applying the fractional change transformation, the autocorrelation structure changes dramatically. As shown in Figures 4.3 and 4.5, the ACF of returns drops to near zero at all lags beyond lag 0, and the PACF shows no significant partial correlations. The transformed series thus resembles white noise, removing the trivial persistence that inflates raw-price accuracy. Forecasting returns is therefore a substantially more challenging task that better discriminates between models: a model that performs well on returns must identify implicit patterns in the data rather than relying on simple price persistence. This is particularly relevant for evaluating foundation models, as it tests whether their pre-trained representations can capture subtle dependencies that are invisible in the autocorrelation structure.

Each experimental configuration (described in Section 4.3) is executed both with and without this transformation, enabling analysis of its impact on forecasting accuracy across different model types.

## 4.2 Rolling Forecast Methodology

This study uses a rolling forecast origin methodology (Tashman, 2000) to evaluate model performance. This approach simulates realistic forecasting scenarios where models are periodically updated with new observations as they become available, ensuring all forecasts are strictly out-of-sample.

The forecasting procedure follows these steps iteratively:

1. An initial model or context is established using the training set, $D_{\text{train}}$.
2. A forecast for the next $H$ time steps, $\hat{y}_{t+1:t+H}$, is generated.
3. The model or its input context is updated by incorporating the *actual* observed values for those $H$ periods from the test set, $y_{t+1:t+H}$.

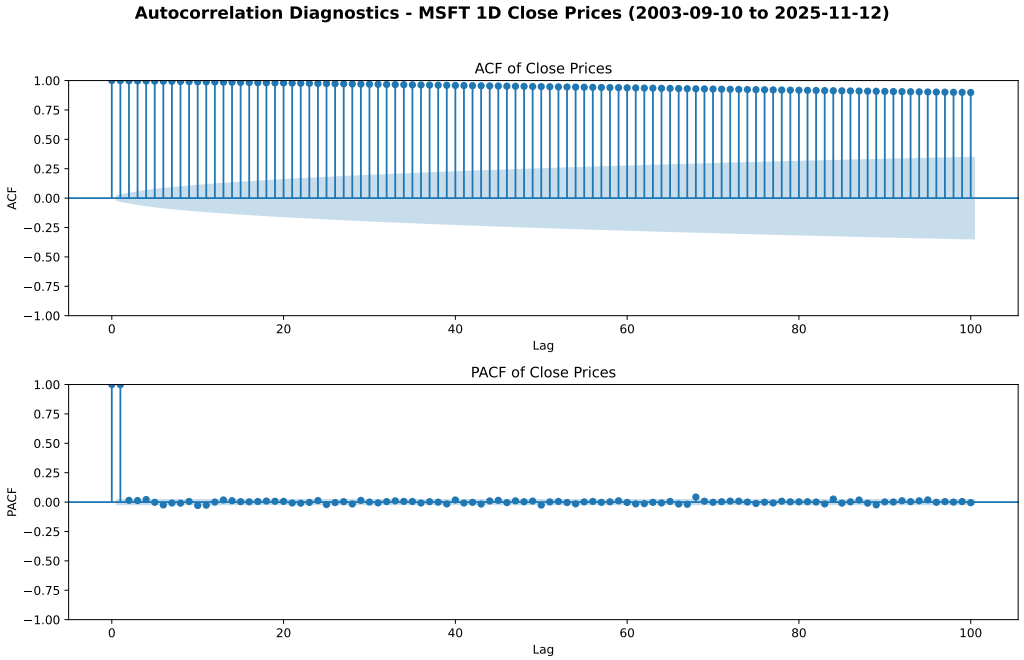


Figure 4.2: ACF and PACF of MSFT daily raw closing prices. The ACF remains near unity across all 100 lags, while the PACF shows a single dominant spike at lag 1, confirming that raw prices follow a near-random-walk process where each value is almost entirely determined by its immediate predecessor.

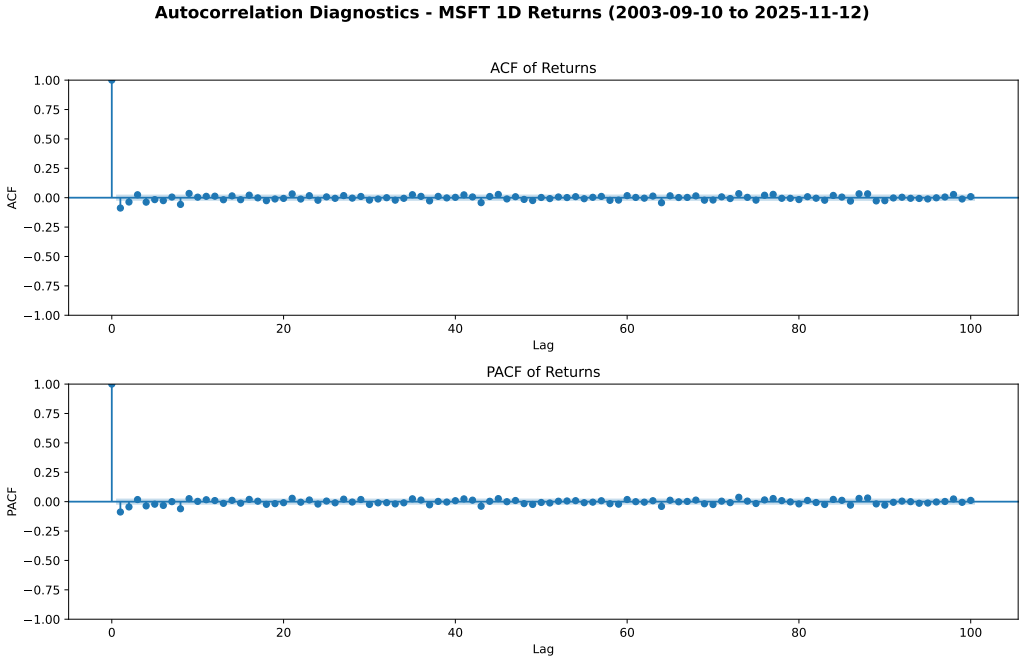


Figure 4.3: ACF and PACF of MSFT daily fractional returns. After the transformation, autocorrelation drops to near zero at all lags beyond lag 0, and no significant partial correlations remain. The series resembles white noise, making forecasting substantially more challenging.

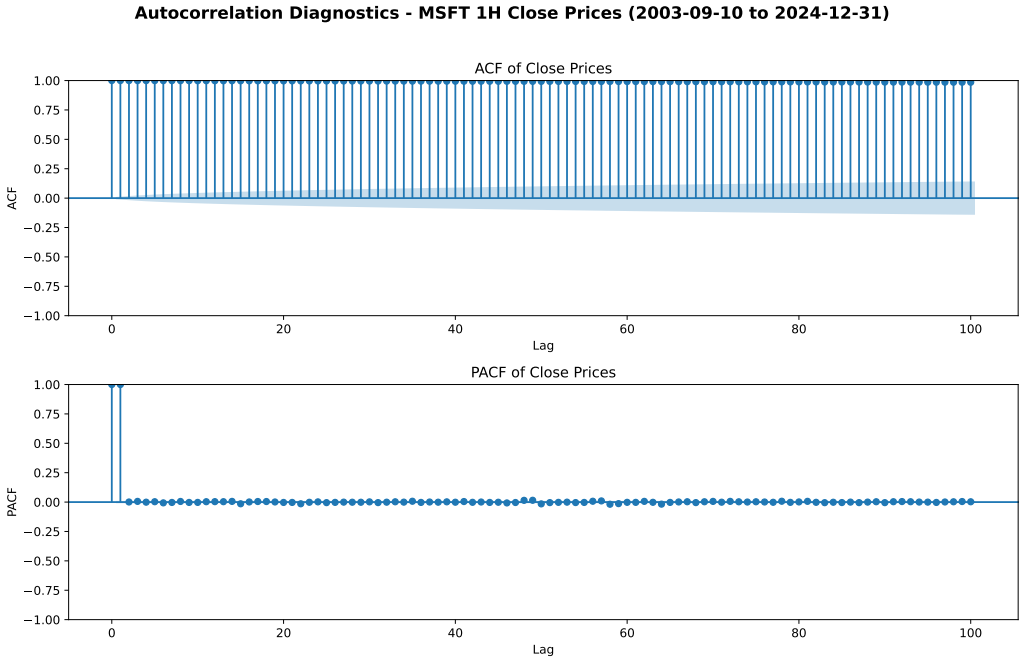


Figure 4.4: ACF and PACF of MSFT hourly raw closing prices. The autocorrelation structure mirrors the daily frequency, with persistent near-unity ACF values across all lags and a dominant lag-1 PACF spike.

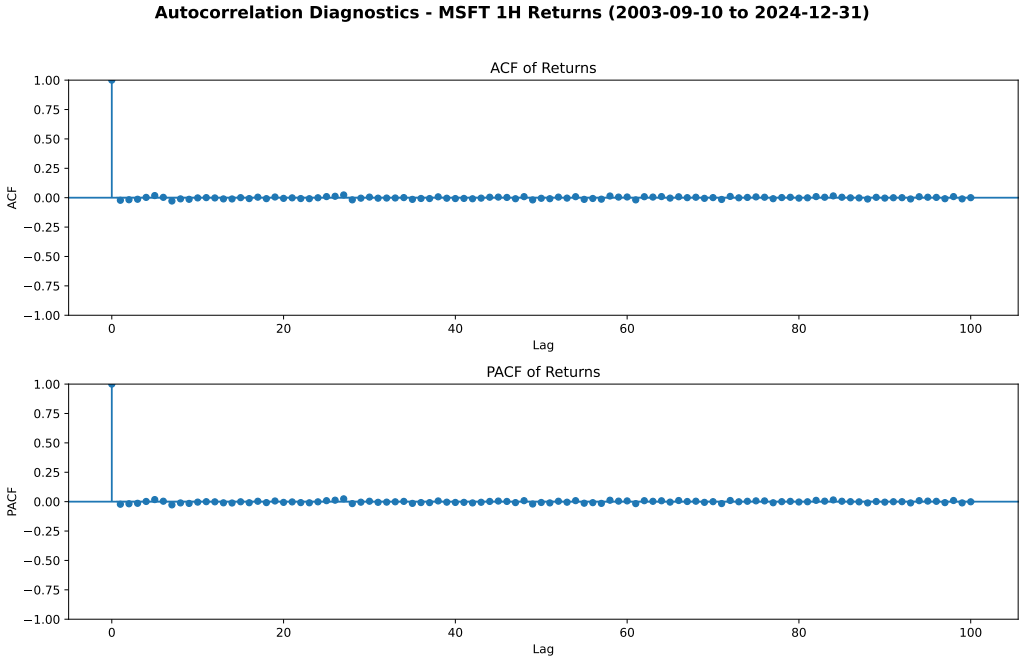


Figure 4.5: ACF and PACF of MSFT hourly fractional returns. As with daily returns, the transformation removes all significant autocorrelation, confirming that the fractional change transformation eliminates trivial price persistence at both temporal frequencies.

4. Steps 2 and 3 are repeated, advancing the forecast origin by $H$ steps, until forecasts cover the entire test period $D_{\text{test}}$.

This iterative updating mechanism ensures that each forecast uses only information that would have been available at the time of prediction, preventing look-ahead bias. The specific implementation of this procedure varies between ARIMA (which updates model coefficients while maintaining fixed order parameters determined via auto-ARIMA from the `pmdarima` Python package (Smith, 2023)) and foundation models (which expand their input context window). For foundation models that produce probabilistic forecasts (Chronos and Sundial), the median of the predictive distribution is used as the point forecast at each step.

## 4.3 Experimental Configurations

The experimental design is structured around three research questions that systematically compare foundation models and traditional forecasting approaches across different forecasting scenarios. Each configuration isolates a specific factor (such as forecast horizon length, temporal data frequency, or market regime).

### 4.3.1 RQ1: Forecast Horizon Length Effects

This configuration addresses RQ1: *How does forecast horizon length affect the relative performance of time-series foundation models compared to traditional approaches for stock price forecasting?* The experiment tests whether foundation models maintain more consistent performance across varying horizons and demonstrate performance advantages at longer forecast distances.

All models are evaluated across four forecast horizons using daily frequency data:

- $H = 1$, short-term
- $H = 5$, approximately one trading week
- $H = 21$, approximately one trading month
- $H = 63$, approximately one financial quarter

To comprehensively investigate how training data availability interacts with horizon length effects, this configuration uses three complementary training period strategies:

**Year-Increment** Training data is varied from 1 to 10 years in annual increments, providing a systematic investigation of how performance changes with increasing amounts of historical context. This configuration spans a wide range of data availability while maintaining consistent annual granularity in the training period variation.

**Logarithmic Spacing** Models are evaluated using 10 logarithmically-spaced training periods between 25 and 250 trading days (approximately 1-12 months). Logarithmic spacing allocates more experimental samples at the lower end of the data range (e.g., 25, 32, 41, 53 days) where marginal increases in data availability may yield larger performance differences. This higher resolution at shorter training periods enables precise characterization of how foundation models and traditional approaches differ in their minimum viable data requirements across varying forecast horizons.

**Linear Spacing** Models are evaluated using 10 linearly-spaced training periods between 25 and 250 trading days (25, 50, 75, ..., 250 days), providing uniform sampling across the limited-data regime. This ensures unbiased coverage and helps validate that observed performance trends are not artifacts of the logarithmic sampling scheme.

### 4.3.2 RQ2: Temporal Frequency Effects

This configuration addresses RQ2: *How does the temporal frequency of input data affect the relative performance of foundation models compared to traditional approaches for stock price forecasting?* The experiment tests whether foundation models exhibit greater robustness when transitioning between frequencies, benefit differently from higher temporal resolution, and achieve competitive zero-shot performance across frequencies without task-specific retraining.

All models perform 1-step-ahead predictions ($H = 1$) on both temporal frequencies:

- **Daily frequency**: 1-day ahead predictions using daily closing prices
- **Hourly frequency**: 1-hour ahead predictions during market hours

For each frequency, models receive equivalent temporal spans of training data. ARIMA is retrained from scratch at each frequency with optimized parameters. Foundation models (Chronos, Sundial, TimesFM) operate in zero-shot mode without frequency-specific fine-tuning, relying on their pre-trained capabilities to adapt to different temporal resolutions.

### 4.3.3 RQ3: Market Disruption Resilience

This configuration addresses RQ3: *Do foundation models' extensive pre-training on diverse market data provide practical advantages over traditional approaches during major market disruptions?* The experiment tests whether foundation models exhibit smaller performance degradation during crisis periods and maintain more consistent performance rankings compared to traditional models trained only on pre-crisis data.

Models are evaluated during three major historical market disruption events using hourly data with linearly- and log-spaced training lengths (25 to 250 days):

For each event, ARIMA is trained exclusively on pre-crisis data, ensuring that no crisis observations leak into the training set. Table 4.1 summarises the three crisis windows and corresponding training cutoffs.

Table 4.1: Crisis periods evaluated and training data cutoffs for the market disruption resilience experiments.

| **Crisis Event** | **Crisis Period Evaluated** | **Training Cutoff** |
|---|---|---|
| Global Financial Crisis | Jan 22, 2008 – Mar 19, 2008 | Jan 21, 2008 |
| COVID-19 Crash | Feb 28, 2020 – Apr 27, 2020 | Feb 27, 2020 |
| 2022 Market Downturn | Mar 31, 2022 – May 27, 2022 | Mar 30, 2022 |

Figure 4.6 visualizes the three selected crisis periods in the context of MSFT's full price history. The upper panel shows the hourly price series with crash periods shaded by severity (moderate, major, and severe drawdowns), while the lower panel displays the corresponding drawdown from the running maximum. The 2008 crisis produced the most severe drawdown (exceeding 60%), followed by the 2020 COVID crash and the 2022 downturn, each exhibiting distinct crash-and-recovery dynamics.

For all three crisis periods, ARIMA is trained only on pre-crisis data, simulating realistic scenarios where recent training data does not include extreme market conditions. Foundation models leverage their pre-

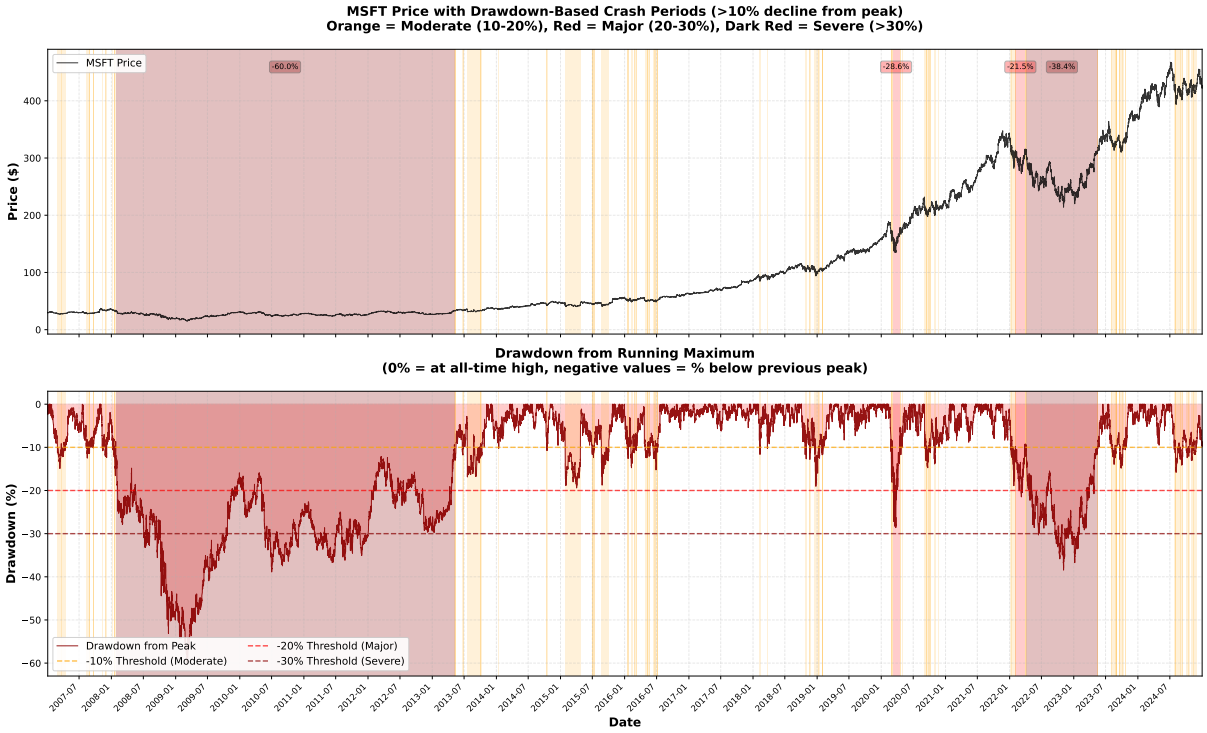


Figure 4.6: MSFT hourly price and drawdown-based crash detection. **Top:** price series with shaded crash periods classified by drawdown severity (orange: 10–20% moderate; red: 20–30% major; dark red: >30% severe). **Bottom:** percentage drawdown from the running all-time high. The three crisis windows used in the RQ3 configuration correspond to the major shaded regions around 2008, 2020, and 2022.

existing knowledge from pre-training on large-scale diverse datasets, which inherently includes exposure to various historical market regimes and crisis periods globally. All models generate 1-hour-ahead predictions ($H = 1$) on hourly data throughout the crisis periods.

**Data leakage consideration.** An important caveat of this experimental configuration is that the foundation models evaluated in this study were pre-trained on large-scale datasets that very likely include financial data from the 2008 and 2020 crisis periods. Consequently, any performance advantage observed during these periods may reflect pattern recognition from pre-training data rather than genuine generalization to unseen market disruptions. This creates an asymmetric comparison: traditional models (ARIMA, Naïve) are trained exclusively on pre-crisis data, while foundation models may have already "seen" the crisis patterns during pre-training. The 2022 market downturn, being more recent, is less likely to be included in the pre-training corpora and therefore provides a more reliable assessment of true crisis resilience. Results from all three periods should be interpreted with this limitation in mind.

## 4.4 Evaluation Metrics

This study uses four evaluation metrics, chosen to capture complementary aspects of forecast quality as discussed by Hyndman and Koehler (2006): Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Directional Accuracy (MDA). Together, they cover scale-independent relative error (MAPE), absolute error with large-deviation penalization (RMSE), absolute error with equal weighting (MAE), and directional correctness (MDA).

### 4.4.1 Mean Absolute Percentage Error (MAPE)

MAPE expresses forecast error as a scale-independent percentage, enabling comparison across series with different magnitudes. However, it is undefined when actual values are zero and can yield extremely large values when $y_i$ is close to zero:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\,, \tag{4.2}$$

where $y_i$ denotes the actual value, $\hat{y}_i$ the predicted value, and $n$ the number of predictions.

### 4.4.2 Root Mean Square Error (RMSE)

RMSE is expressed in the same unit as the target variable and penalizes larger errors disproportionately due to the squaring step:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.3}$$

### 4.4.3 Mean Directional Accuracy (MDA)

MDA measures the percentage of predictions that correctly forecast the direction of price change. It is particularly relevant for financial forecasting where directional accuracy often matters more than magnitude, and is especially valuable for fractional change data where error-based metrics produce very small absolute values that are difficult to interpret.

For raw price forecasts, directional accuracy is computed by comparing the sign of the realised price change with the sign of the predicted price change:

$$\text{MDA} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{(\text{sign}(y_i - y_{i-1}) = \text{sign}(\hat{y}_i - y_{i-1}))} \tag{4.4}$$

In this thesis, MDA is used only for fractional-change forecasts. Since the fractional-change target is already a signed return, the operative metric for the MDA tables is computed directly on the signs of the realised and predicted returns:

$$\text{MDA}_r = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{(\text{sign}(r_i) = \text{sign}(\hat{r}_i))} \tag{4.5}$$

where $\mathbf{1}_{(\cdot)}$ is the indicator function (1 when the condition is true, 0 otherwise). In the implementation, this is calculated as the mean of the equality comparison between `np.sign(y_true)` and `np.sign(y_pred)`. The calculation therefore follows NumPy's sign convention, where zero has sign 0 rather than a positive or negative direction. Consequently, a zero return forecast is counted as directionally correct only when the realised return is also exactly zero.

### 4.4.4 Mean Absolute Error (MAE)

MAE calculates the average absolute difference between predicted and actual values, expressed in the same unit as the target variable. Unlike RMSE, MAE treats all deviations equally without penalizing larger errors:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.6}$$

## 4.5 Implementation and Execution

### 4.5.1 Computational Infrastructure

All experiments are implemented in Python 3.11 using a modular pipeline architecture. Due to conflicting dependency requirements among foundation models (particularly `transformers` library versions: Chronos requires $\geq$4.48.0 while Sundial requires exactly 4.40.1), separate virtual environments are used for each model.

### 4.5.2 Experimental Pipeline

Experiments are executed through a unified pipeline script(`pipeline.py`) that performs the following steps:

1. Data loading from pre-downloaded historical stock data
2. Application of optional preprocessing transformation
3. Train-test splitting
4. Model initialization and rolling forecast execution
5. Metric calculation across all evaluation measures
6. Automated result logging and configuration archival

Each experiment's configuration parameters, including the resulting evaluation metrics, are automatically serialized to YAML files with timestamps, while forecast results and actual values are saved as CSV files.

# Chapter 5

# Results

This chapter reports the rolling forecast results for Naïve, ARIMA, Chronos, Sundial, and TimesFM on MSFT. The sections follow the three research questions and use four evaluation metrics: MAPE, RMSE, MAE, and MDA. MDA is reported for fractional-change forecasts but omitted for raw-price forecasts because directional accuracy is consistently 100% and therefore uninformative.

## 5.1 RQ1: Forecast Horizon Length Effects

RQ1 examines how increasing forecast horizon length affects model performance. Experiments use a fixed test period of January–December 2023 with variable training data preceding that date. Two training configurations are evaluated: a year-increment regime spanning 1–10 years and a limited-data regime using sub-year training lengths sampled on a logarithmic scale (25–249 days). A linear-spaced counterpart of the limited-data regime (25–250 days) was also evaluated and yields consistent findings, those tables are reported in Appendix A.1.

### 5.1.1 Raw Price Forecasting

#### Year-Increment Configuration

Tables 5.1–5.3 present MAPE, RMSE, and MAE for raw closing price forecasting across four horizons (1d, 5d, 21d, 63d) and ten training lengths (1–10 years).

At short horizons (1d, 5d), Naïve and ARIMA dominate across the error metrics. ARIMA matches Naïve with 1–3 years of training and marginally improves from 6 years onward, while the foundation models generally trail both baselines. At longer horizons, the gap narrows and occasionally reverses: TimesFM achieves the lowest MAPE at 21d with 6 years of training (2.87% vs. Naïve's fixed 3.06%), and Chronos leads at 63d with 1 year of training (5.67% vs. Naïve's 7.17%). Sundial is the weakest model overall at the longer horizons, especially at 21d and most 63d settings. RMSE rankings largely mirror MAPE, with one notable exception: at the 63d horizon, Chronos achieves the lowest RMSE at nine out of ten training lengths (all except 8 years, where ARIMA edges it by a narrow margin), a dominance not apparent from the MAPE rankings alone.

Table 5.1: MAPE (%) for raw price forecasting, year-increment configuration (1–10 years of daily training data). Bold indicates best performance; underline indicates second best.

| Horizon | Method | 1y | 2y | 3y | 4y | 5y | 6y | 7y | 8y | 9y | 10y |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1d | Naive | **1.2265** | **1.2265** | **1.2265** | 1.2265 | 1.2265 | 1.2265 | 1.2265 | 1.2265 | 1.2265 | 1.2265 |
| | ARIMA | **1.2265** | **1.2265** | **1.2265** | **1.2218** | **1.2218** | **1.2174** | **1.2179** | **1.2094** | **1.2094** | **1.2096** |
| | TimesFM | 1.2760 | 1.2594 | 1.3468 | 1.3859 | 1.3796 | 1.3716 | 1.3236 | 1.3271 | 1.3100 | 1.3100 |
| | Chronos | 1.3066 | 1.3553 | 1.4162 | 1.4087 | 1.3918 | 1.4565 | 1.4573 | 1.4454 | 1.4478 | 1.4478 |
| | Sundial | 1.2896 | 1.2793 | 1.2807 | 1.3204 | 1.3342 | 1.3637 | 1.3575 | 1.3555 | 1.3571 | 1.3721 |
| 5d | Naive | **1.9748** | **1.9748** | **1.9748** | **1.9748** | **1.9748** | 1.9748 | 1.9748 | 1.9748 | 1.9748 | 1.9748 |
| | ARIMA | **1.9748** | **1.9748** | **1.9748** | 1.9843 | 1.9842 | **1.9521** | **1.9562** | **1.9384** | **1.9398** | **1.9411** |
| | TimesFM | 2.0028 | 1.9972 | 1.9796 | 2.0351 | 2.0400 | 2.1250 | 2.0076 | 2.1145 | 2.1248 | 2.1248 |
| | Chronos | 2.1037 | 2.1727 | 2.2491 | 2.2851 | 2.2953 | 2.3557 | 2.2827 | 2.2811 | 2.2810 | 2.2810 |
| | Sundial | 2.1710 | 2.0258 | 2.1635 | 2.1682 | 2.2137 | 2.1879 | 2.1509 | 2.0760 | 2.1485 | 2.1234 |
| 21d | Naive | **3.0611** | **3.0611** | **3.0611** | **3.0611** | **3.0611** | 3.0611 | 3.0611 | 3.0611 | 3.0611 | 3.0611 |
| | ARIMA | **3.0611** | **3.0611** | **3.0611** | 3.1722 | 3.1736 | 3.0426 | **3.0537** | **3.0296** | **3.0351** | **3.0382** |
| | TimesFM | 3.2359 | 3.1178 | 3.1905 | 3.0837 | 3.1279 | **2.8674** | 3.1594 | 3.1121 | 3.1292 | 3.1292 |
| | Chronos | 3.3894 | 3.3920 | 3.2775 | 3.3475 | 3.3364 | 3.3333 | 3.2751 | 3.2698 | 3.2670 | 3.2670 |
| | Sundial | 4.0683 | 3.6599 | 3.8193 | 4.2684 | 3.9659 | 3.8185 | 3.4312 | 3.5583 | 3.6799 | 3.8132 |
| 63d | Naive | 7.1680 | **7.1680** | 7.1680 | **7.1680** | **7.1680** | 7.1680 | 7.1680 | 7.1680 | 7.1680 | 7.1680 |
| | ARIMA | 7.1680 | **7.1680** | 7.1680 | 7.2710 | 7.2695 | 6.6920 | **6.7471** | **6.7038** | **6.7419** | **6.7700** |
| | TimesFM | 7.3662 | 7.3086 | 7.2656 | 7.3976 | 7.3101 | **6.6426** | 7.1514 | 7.2008 | 7.1748 | 7.1748 |
| | Chronos | **5.6701** | 7.7146 | **6.3448** | 7.1901 | 7.1807 | 7.0379 | 7.0446 | 7.0403 | 7.0199 | 7.0199 |
| | Sundial | 7.4388 | 7.4502 | 9.4200 | 9.5580 | 9.7080 | 9.3957 | 9.6126 | 8.9511 | 8.9295 | 8.9716 |

Table 5.2: RMSE for raw price forecasting, year-increment configuration (1–10 years of daily training data). Bold indicates best performance; underline indicates second best.

| Horizon | Method | 1y | 2y | 3y | 4y | 5y | 6y | 7y | 8y | 9y | 10y |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1d | Naive | **4.9629** | **4.9629** | **4.9629** | 4.9629 | 4.9629 | 4.9629 | 4.9629 | 4.9629 | 4.9629 | 4.9629 |
| | ARIMA | **4.9629** | **4.9629** | **4.9629** | **4.9252** | **4.9234** | **4.9051** | **4.9071** | **4.8699** | **4.8699** | **4.8718** |
| | TimesFM | 5.0598 | 5.0698 | 5.2665 | 5.4612 | 5.5861 | 5.4355 | 5.3139 | 5.3523 | 5.3072 | 5.3072 |
| | Chronos | 5.2107 | 5.3154 | 5.5534 | 5.5431 | 5.4747 | 5.6779 | 5.6765 | 5.6362 | 5.6345 | 5.6345 |
| | Sundial | 5.2293 | 5.1561 | 5.1774 | 5.2259 | 5.2521 | 5.3217 | 5.2986 | 5.2892 | 5.3126 | 5.3577 |
| 5d | Naive | **7.8314** | **7.8314** | **7.8314** | 7.8314 | 7.8314 | 7.8314 | 7.8314 | 7.8314 | 7.8314 | 7.8314 |
| | ARIMA | **7.8314** | **7.8314** | **7.8314** | **7.7940** | **7.7904** | **7.7037** | **7.7151** | **7.6687** | **7.6729** | **7.6766** |
| | TimesFM | 7.9604 | 8.0075 | 7.8372 | 7.9063 | 7.9703 | 8.3956 | 7.9425 | 8.5147 | 8.5268 | 8.5268 |
| | Chronos | 8.2606 | 8.3881 | 8.6075 | 8.6944 | 8.7687 | 8.9544 | 8.7101 | 8.6524 | 8.6538 | 8.6538 |
| | Sundial | 8.6145 | 8.1678 | 8.4778 | 8.5602 | 8.6748 | 8.5092 | 8.3990 | 8.0930 | 8.2977 | 8.2360 |
| 21d | Naive | **13.1372** | **13.1372** | **13.1372** | **13.1372** | **13.1372** | 13.1372 | 13.1372 | 13.1372 | 13.1372 | 13.1372 |
| | ARIMA | **13.1372** | **13.1372** | **13.1372** | 13.5722 | 13.5825 | 12.7957 | **12.8738** | **12.8738** | **12.9141** | 12.9429 |
| | TimesFM | 13.9758 | 13.4649 | 13.5688 | 13.4130 | 13.3409 | **12.2858** | 13.3711 | 12.8818 | 12.9173 | **12.9173** |
| | Chronos | 14.7447 | 14.1142 | 13.4960 | 13.5007 | 13.4756 | 13.4319 | 13.2713 | 13.3000 | 13.3092 | 13.3092 |
| | Sundial | 16.9961 | 16.7858 | 15.8698 | 17.2295 | 15.8566 | 15.3006 | 14.1566 | 14.9574 | 15.6375 | 15.6807 |
| 63d | Naive | 29.5861 | 29.5861 | 29.5861 | 29.5861 | 29.5861 | 29.5861 | 29.5861 | 29.5861 | 29.5861 | 29.5861 |
| | ARIMA | 29.5861 | 29.5861 | 29.5861 | 30.0197 | 30.0103 | 26.8069 | 27.0894 | **27.0442** | 27.2345 | 27.3778 |
| | TimesFM | 31.7005 | 32.0929 | 29.7440 | 30.1676 | 29.2256 | 28.4908 | 28.3830 | 28.2538 | 28.1817 | 28.1817 |
| | Chronos | **24.0536** | **28.8474** | **24.2515** | **26.6942** | **26.7110** | **26.4808** | **26.9253** | 27.0640 | **27.0189** | **27.0189** |
| | Sundial | 34.3845 | 34.9457 | 38.0803 | 37.7506 | 36.0612 | 35.2678 | 38.0358 | 36.1072 | 35.8970 | 36.6949 |

Table 5.3: MAE for raw price forecasting, year-increment configuration (1–10 years of daily training data). Bold indicates best performance; underline indicates second best.

| Horizon | Method | 1y | 2y | 3y | 4y | 5y | 6y | 7y | 8y | 9y | 10y |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1d | Naive | **3.7485** | **3.7485** | **3.7485** | 3.7485 | 3.7485 | 3.7485 | 3.7485 | 3.7485 | 3.7485 | 3.7485 |
| | ARIMA | **3.7485** | **3.7485** | **3.7485** | **3.7318** | **3.7316** | **3.7178** | **3.7193** | **3.6933** | **3.6934** | **3.6940** |
| | TimesFM | 3.8884 | 3.8588 | 4.1497 | 4.2709 | 4.2687 | 4.2221 | 4.0621 | 4.0725 | 4.0323 | 4.0323 |
| | Chronos | 3.9963 | 4.1247 | 4.3183 | 4.2923 | 4.2530 | 4.4444 | 4.4569 | 4.4384 | 4.4430 | 4.4430 |
| | Sundial | 3.9833 | 3.9415 | 3.9444 | 4.0409 | 4.0772 | 4.1733 | 4.1627 | 4.1612 | 4.1749 | 4.2206 |
| 5d | Naive | **6.0328** | **6.0328** | 6.0328 | **6.0328** | **6.0328** | 6.0328 | 6.0328 | 6.0328 | 6.0328 | 6.0328 |
| | ARIMA | **6.0328** | **6.0328** | 6.0328 | 6.0590 | 6.0585 | **5.9564** | **5.9691** | **5.9173** | **5.9219** | **5.9258** |
| | TimesFM | 6.1040 | 6.1303 | **6.0291** | 6.2050 | 6.2454 | 6.4963 | 6.1728 | 6.5395 | 6.5644 | 6.5644 |
| | Chronos | 6.4279 | 6.6178 | 6.8515 | 6.9471 | 6.9880 | 7.1729 | 6.9576 | 6.9674 | 6.9671 | 6.9671 |
| | Sundial | 6.7512 | 6.2714 | 6.6484 | 6.5860 | 6.7273 | 6.6611 | 6.5580 | 6.3729 | 6.5683 | 6.5006 |
| 21d | Naive | **9.5962** | **9.5962** | **9.5962** | **9.5962** | **9.5962** | 9.5962 | **9.5962** | 9.5962 | 9.5962 | 9.5962 |
| | ARIMA | **9.5962** | **9.5962** | **9.5962** | 9.9622 | 9.9669 | 9.5705 | 9.6042 | **9.5404** | **9.5567** | **9.5650** |
| | TimesFM | 10.1817 | 9.7962 | 9.9627 | 9.6428 | 9.7391 | **8.9690** | 9.8911 | 9.7028 | 9.7435 | 9.7435 |
| | Chronos | 10.6639 | 10.6342 | 10.2619 | 10.3757 | 10.2844 | 10.2561 | 10.1114 | 10.1109 | 10.1054 | 10.1054 |
| | Sundial | 12.9228 | 11.8199 | 11.7792 | 12.8623 | 12.0544 | 11.6833 | 10.6049 | 10.9944 | 11.4374 | 11.7607 |
| 63d | Naive | 23.2427 | **23.2427** | 23.2427 | 23.2427 | 23.2427 | 23.2427 | 23.2427 | 23.2427 | 23.2427 | 23.2427 |
| | ARIMA | 23.2427 | **23.2427** | 23.2427 | 23.5918 | 23.5859 | 21.7662 | **21.9389** | **21.7914** | **21.9113** | **21.9996** |
| | TimesFM | 24.1446 | 24.0747 | 23.5227 | 23.9451 | 23.5574 | **21.6186** | 22.9539 | 23.0956 | 23.0239 | 23.0239 |
| | Chronos | **18.7126** | 24.2715 | **20.1920** | **22.4571** | **22.3831** | 21.9427 | 22.0661 | 22.1359 | 22.0845 | 22.0845 |
| | Sundial | 24.9231 | 25.2164 | 29.4990 | 28.9408 | 30.2930 | 29.3023 | 30.3891 | 28.2717 | 28.2765 | 28.5703 |

**Limited-Data Regime**

Tables 5.4–5.6 present results under sub-year training lengths with logarithmic spacing (25–249 days).

ARIMA is unstable in this regime, failing to converge at several training lengths (25d, 41d, 53d, and 116d), producing errors several times larger than any other model. For instance, 5.91% MAPE at 1d with 25 days of training versus Naïve's fixed 1.23%. At converging lengths, ARIMA matches Naïve. Foundation models, by contrast, maintain stable performance regardless of training length.

Model rankings shift with horizon length. At short horizons (1d, 5d), Naïve leads and TimesFM is usually the strongest foundation model. At 21d, Naïve remains the strongest model, while Chronos is the strongest foundation model. At 63d, Chronos becomes the best-performing model at most training lengths, for example achieving a MAPE of 6.25% with 25 days of training versus Naïve's 7.17%. TimesFM and Sundial generally trail Chronos at the longer horizons, with Sundial especially weak at 21d and in many 63d settings. RMSE and MAE broadly support the same trend.

Table 5.4: MAPE (%) for raw price forecasting, logarithmically-spaced training lengths (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Horizon | Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1d | Naive | **1.2265** | **1.2265** | **1.2265** | **1.2265** | **1.2265** | **1.2265** | **1.2265** | **1.2265** | **1.2265** | **1.2265** |
| | ARIMA | 5.9137 | **1.2265** | 5.4893 | 4.1973 | **1.2265** | **1.2265** | 2.0016 | **1.2265** | **1.2265** | **1.2265** |
| | TimesFM | 1.2846 | 1.2986 | 1.3065 | 1.2985 | 1.2572 | 1.2708 | 1.2560 | 1.2813 | 1.3088 | 1.3043 |
| | Chronos | 1.3288 | 1.3256 | 1.3355 | 1.3311 | 1.3241 | 1.3279 | 1.3405 | 1.3315 | 1.3180 | 1.3476 |
| | Sundial | 1.3753 | 1.3548 | 1.3614 | 1.4096 | 1.3738 | 1.3977 | 1.3609 | 1.2989 | 1.2964 | 1.3317 |
| 5d | Naive | **1.9748** | **1.9748** | **1.9748** | **1.9748** | **1.9748** | **1.9748** | **1.9748** | **1.9748** | **1.9748** | **1.9748** |
| | ARIMA | 9.8097 | **1.9748** | 9.6981 | 8.5527 | **1.9748** | **1.9748** | 5.2590 | **1.9748** | **1.9748** | **1.9748** |
| | TimesFM | 2.0405 | 2.0582 | 2.0637 | 2.1066 | 2.0936 | 2.0918 | 2.0236 | 1.9794 | 2.0821 | 2.0210 |
| | Chronos | 2.1096 | 2.1068 | 2.1130 | 2.1093 | 2.1062 | 2.1157 | 2.1191 | 2.0879 | 2.0762 | 2.0958 |
| | Sundial | 2.4000 | 2.3791 | 2.3771 | 2.4433 | 2.4148 | 2.4863 | 2.5074 | 2.2822 | 2.2609 | 2.3478 |
| 21d | Naive | **3.0611** | **3.0611** | **3.0611** | **3.0611** | **3.0611** | **3.0611** | **3.0611** | **3.0611** | **3.0611** | **3.0611** |
| | ARIMA | 12.0167 | **3.0611** | 12.3759 | 12.2954 | **3.0611** | **3.0611** | 11.2846 | **3.0611** | **3.0611** | **3.0611** |
| | TimesFM | 4.1248 | 4.2649 | 4.1495 | 4.0554 | 3.8511 | 4.0466 | 4.0907 | 3.9303 | 3.8285 | 3.6615 |
| | Chronos | 3.5284 | 3.5969 | 3.6090 | 3.5085 | 3.4065 | 3.3970 | 3.3835 | 3.2579 | 3.2522 | 3.2970 |
| | Sundial | 4.9178 | 4.8607 | 4.7058 | 4.6481 | 4.9943 | 5.0701 | 5.0322 | 4.4844 | 4.4967 | 4.9969 |
| 63d | Naive | 7.1680 | 7.1680 | 7.1680 | 7.1680 | 7.1680 | 7.1680 | **7.1680** | **7.1680** | 7.1680 | 7.1680 |
| | ARIMA | 13.0838 | 7.1680 | 13.6759 | 13.9048 | 7.1680 | 7.1680 | 15.4559 | **7.1680** | 7.1680 | 7.1680 |
| | TimesFM | 10.1453 | 10.3788 | 9.2793 | 7.6543 | 7.9844 | 8.1839 | 8.5927 | 8.1904 | 8.4185 | 8.1016 |
| | Chronos | **6.2533** | **6.6497** | **6.3163** | **6.2782** | **6.6851** | **7.0735** | 7.6533 | 7.5118 | **6.8682** | **6.5650** |
| | Sundial | 9.7278 | 10.3260 | 9.2008 | 9.6219 | 10.4327 | 10.4883 | 11.1375 | 10.6235 | 9.8865 | 9.8787 |

Table 5.5: RMSE for raw price forecasting, logarithmically-spaced training lengths (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Horizon | Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1d | Naive | **4.9629** | **4.9629** | **4.9629** | **4.9629** | **4.9629** | **4.9629** | **4.9629** | **4.9629** | **4.9629** | **4.9629** |
| | ARIMA | 21.6781 | **4.9629** | 20.1184 | 15.4514 | **4.9629** | **4.9629** | 7.6404 | **4.9629** | **4.9629** | **4.9629** |
| | TimesFM | 5.1445 | 5.1840 | 5.1689 | 5.1171 | 5.0374 | 5.0900 | 5.0632 | 5.0821 | 5.1287 | 5.1931 |
| | Chronos | 5.3182 | 5.3034 | 5.3101 | 5.3057 | 5.2304 | 5.2944 | 5.3147 | 5.2567 | 5.2018 | 5.2782 |
| | Sundial | 5.4835 | 5.4196 | 5.4431 | 5.5464 | 5.4366 | 5.5233 | 5.3953 | 5.2277 | 5.2221 | 5.3286 |
| 5d | Naive | **7.8314** | **7.8314** | **7.8314** | **7.8314** | **7.8314** | **7.8314** | **7.8314** | 7.8314 | **7.8314** | **7.8314** |
| | ARIMA | 36.8652 | **7.8314** | 36.5931 | 32.6752 | **7.8314** | **7.8314** | 20.6205 | 7.8314 | **7.8314** | **7.8314** |
| | TimesFM | 7.9504 | 8.0503 | 8.0342 | 8.1094 | 8.1364 | 8.1393 | 8.0297 | **7.7397** | 8.1027 | 7.9412 |
| | Chronos | 8.4397 | 8.3995 | 8.3999 | 8.3529 | 8.3244 | 8.4848 | 8.3531 | 8.2128 | 8.2101 | 8.2484 |
| | Sundial | 9.2626 | 9.3435 | 9.1389 | 9.4936 | 9.4121 | 9.6534 | 9.6878 | 8.7780 | 8.6806 | 9.1882 |
| 21d | Naive | **13.1372** | **13.1372** | **13.1372** | **13.1372** | **13.1372** | **13.1372** | **13.1372** | **13.1372** | **13.1372** | **13.1372** |
| | ARIMA | 44.5202 | **13.1372** | 45.9843 | 46.0459 | **13.1372** | **13.1372** | 43.4631 | **13.1372** | **13.1372** | **13.1372** |
| | TimesFM | 16.5921 | 16.9779 | 17.0470 | 17.0455 | 16.3890 | 16.9211 | 16.9890 | 16.3068 | 15.8822 | 15.3046 |
| | Chronos | 14.5133 | 14.9663 | 15.2333 | 14.9751 | 14.8636 | 14.3785 | 14.2566 | 13.3234 | 13.4386 | 13.7102 |
| | Sundial | 19.9147 | 19.8517 | 19.4645 | 19.8621 | 20.8062 | 21.1713 | 21.1407 | 19.2204 | 19.0548 | 21.0016 |
| 63d | Naive | 29.5861 | 29.5861 | 29.5861 | 29.5861 | 29.5861 | 29.5861 | **29.5861** | **29.5861** | 29.5861 | 29.5861 |
| | ARIMA | 49.8257 | 29.5861 | 51.5196 | 52.9576 | 29.5861 | 29.5861 | 58.3815 | **29.5861** | 29.5861 | 29.5861 |
| | TimesFM | 42.9109 | 44.4887 | 39.5056 | 33.7435 | 34.2882 | 35.4591 | 37.0424 | 35.5108 | 36.5691 | 35.7251 |
| | Chronos | **24.5222** | **25.9159** | **24.9002** | **24.8279** | **26.1340** | **28.4166** | 32.1907 | 30.9610 | **29.4635** | **28.5189** |
| | Sundial | 42.6513 | 44.7302 | 40.5925 | 42.2295 | 46.5947 | 47.1968 | 50.3190 | 48.6723 | 45.0873 | 42.7062 |

Table 5.6: MAE for raw price forecasting, logarithmically-spaced training lengths (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Horizon | Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1d | Naive | **3.7485** | **3.7485** | **3.7485** | **3.7485** | **3.7485** | **3.7485** | **3.7485** | **3.7485** | **3.7485** | **3.7485** |
| | ARIMA | 19.2725 | **3.7485** | 17.8878 | 13.6387 | **3.7485** | **3.7485** | 6.3302 | **3.7485** | **3.7485** | **3.7485** |
| | TimesFM | 3.9270 | 3.9644 | 3.9856 | 3.9521 | 3.8420 | 3.8812 | 3.8547 | 3.9145 | 3.9907 | 3.9852 |
| | Chronos | 4.0621 | 4.0563 | 4.0821 | 4.0655 | 4.0377 | 4.0693 | 4.1127 | 4.0681 | 4.0339 | 4.1281 |
| | Sundial | 4.1952 | 4.1275 | 4.1544 | 4.2884 | 4.1992 | 4.2776 | 4.1748 | 3.9997 | 3.9938 | 4.1023 |
| 5d | Naive | **6.0328** | **6.0328** | **6.0328** | **6.0328** | **6.0328** | **6.0328** | **6.0328** | 6.0328 | **6.0328** | **6.0328** |
| | ARIMA | 32.1573 | **6.0328** | 31.7942 | 28.0079 | **6.0328** | **6.0328** | 17.0695 | 6.0328 | **6.0328** | **6.0328** |
| | TimesFM | 6.2218 | 6.2673 | 6.2794 | 6.3797 | 6.3518 | 6.3622 | 6.1755 | **6.0309** | 6.3202 | 6.1657 |
| | Chronos | 6.4522 | 6.4427 | 6.4601 | 6.4586 | 6.4284 | 6.4720 | 6.4834 | 6.3806 | 6.3468 | 6.4114 |
| | Sundial | 7.3542 | 7.2596 | 7.2858 | 7.4922 | 7.4297 | 7.6696 | 7.7169 | 7.0477 | 6.9692 | 7.2562 |
| 21d | Naive | **9.5962** | **9.5962** | **9.5962** | **9.5962** | **9.5962** | **9.5962** | **9.5962** | **9.5962** | **9.5962** | **9.5962** |
| | ARIMA | 39.4096 | **9.5962** | 40.6351 | 40.3912 | **9.5962** | **9.5962** | 37.0610 | **9.5962** | **9.5962** | **9.5962** |
| | TimesFM | 12.7218 | 13.1865 | 12.8878 | 12.6860 | 11.9375 | 12.5745 | 12.7128 | 12.2660 | 11.8870 | 11.3884 |
| | Chronos | 10.8410 | 11.1380 | 11.1287 | 10.8807 | 10.5267 | 10.5695 | 10.6131 | 10.2619 | 10.2681 | 10.4007 |
| | Sundial | 15.2122 | 14.9911 | 14.6547 | 14.6855 | 15.6743 | 15.9665 | 15.8535 | 14.1743 | 14.2227 | 15.7278 |
| 63d | Naive | 23.2427 | 23.2427 | 23.2427 | 23.2427 | 23.2427 | 23.2427 | **23.2427** | **23.2427** | 23.2427 | 23.2427 |
| | ARIMA | 43.0837 | 23.2427 | 45.0593 | 45.8942 | 23.2427 | 23.2427 | 51.0715 | **23.2427** | 23.2427 | 23.2427 |
| | TimesFM | 31.0596 | 31.7390 | 28.8117 | 25.1528 | 26.1020 | 26.7791 | 28.0211 | 26.7634 | 27.6040 | 26.6430 |
| | Chronos | **20.1430** | **21.4644** | **20.3851** | **20.3175** | **21.4915** | **23.1092** | 25.2293 | 24.6453 | **22.8305** | **21.7765** |
| | Sundial | 31.7948 | 33.9065 | 30.1360 | 31.8114 | 34.5515 | 34.7829 | 37.0482 | 35.2868 | 32.8635 | 32.5995 |

### 5.1.2 Fractional Change Forecasting

Because fractional change values have small magnitude, MAPE is disproportionately inflated for all models under this transformation and is therefore excluded from the analysis. RMSE, MAE, and MDA are the reported metrics throughout this subsection.

#### Year-Increment Configuration

Tables 5.7–5.9 present RMSE, MAE, and MDA for fractional change forecasting across 1–10 years of training.

For RMSE, Naïve records the highest error at every horizon (0.0190–0.0248), while ARIMA, Chronos, TimesFM, and Sundial converge into a narrow band (0.0163–0.0171). ARIMA and Chronos are effectively tied from 5d onward, both achieving 0.0165–0.0166. MAE rankings are consistent with RMSE.

MDA reveals a different ordering, where foundation models show their strongest relative performance. At 1d, Sundial often leads and reaches 56.4%, while at 5d both Sundial and TimesFM reach 58.4% and 57.6% respectively. ARIMA records 0.0% MDA at 1–3 years of training across all horizons, before recovering to 51–55% at longer training lengths. At 63d, Naïve achieves the highest MDA at a fixed 58.8%.

Table 5.7: RMSE for fractional change forecasting, year-increment configuration (1–10 years of daily training data). Bold indicates best performance; underline indicates second best.

| Horizon | Method | 1y | 2y | 3y | 4y | 5y | 6y | 7y | 8y | 9y | 10y |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1d | Naive | 0.0248 | 0.0248 | 0.0248 | 0.0248 | 0.0248 | 0.0248 | 0.0248 | 0.0248 | 0.0248 | 0.0248 |
| | ARIMA | **0.0166** | **0.0166** | 0.0166 | **0.0164** | **0.0163** | **0.0163** | **0.0163** | **0.0164** | **0.0164** | **0.0164** |
| | TimesFM | 0.0168 | 0.0167 | 0.0169 | 0.0168 | 0.0167 | 0.0167 | 0.0167 | 0.0168 | 0.0168 | 0.0168 |
| | Chronos | 0.0167 | **0.0166** | 0.0166 | 0.0166 | 0.0166 | 0.0166 | 0.0166 | 0.0166 | 0.0166 | 0.0166 |
| | Sundial | 0.0168 | 0.0167 | **0.0164** | 0.0166 | 0.0165 | 0.0166 | 0.0165 | 0.0165 | 0.0166 | 0.0165 |
| 5d | Naive | 0.0221 | 0.0221 | 0.0221 | 0.0221 | 0.0221 | 0.0221 | 0.0221 | 0.0221 | 0.0221 | 0.0221 |
| | ARIMA | **0.0166** | **0.0166** | 0.0166 | 0.0165 | **0.0165** | **0.0165** | **0.0165** | **0.0165** | **0.0165** | **0.0165** |
| | TimesFM | 0.0170 | 0.0170 | 0.0168 | 0.0167 | 0.0167 | 0.0167 | 0.0167 | 0.0166 | 0.0166 | 0.0166 |
| | Chronos | **0.0166** | **0.0166** | 0.0166 | 0.0166 | **0.0165** | 0.0166 | 0.0166 | 0.0166 | 0.0166 | 0.0166 |
| | Sundial | 0.0169 | **0.0166** | **0.0164** | **0.0163** | **0.0165** | **0.0165** | 0.0166 | 0.0166 | 0.0167 | **0.0165** |
| 21d | Naive | 0.0219 | 0.0219 | 0.0219 | 0.0219 | 0.0219 | 0.0219 | 0.0219 | 0.0219 | 0.0219 | 0.0219 |
| | ARIMA | **0.0166** | **0.0166** | 0.0166 | **0.0165** | **0.0165** | **0.0165** | **0.0165** | **0.0165** | **0.0165** | **0.0165** |
| | TimesFM | 0.0167 | 0.0167 | 0.0166 | 0.0166 | 0.0166 | 0.0166 | 0.0167 | 0.0167 | 0.0167 | 0.0167 |
| | Chronos | **0.0166** | **0.0166** | 0.0165 | **0.0165** | **0.0165** | **0.0165** | **0.0165** | **0.0165** | **0.0165** | **0.0165** |
| | Sundial | 0.0171 | **0.0166** | **0.0164** | 0.0168 | 0.0167 | 0.0166 | 0.0167 | 0.0167 | 0.0167 | 0.0169 |
| 63d | Naive | 0.0190 | 0.0190 | 0.0190 | 0.0190 | 0.0190 | 0.0190 | 0.0190 | 0.0190 | 0.0190 | 0.0190 |
| | ARIMA | **0.0166** | **0.0166** | 0.0166 | **0.0165** | 0.0165 | **0.0165** | **0.0165** | **0.0165** | **0.0165** | **0.0165** |
| | TimesFM | 0.0167 | **0.0166** | **0.0165** | 0.0166 | 0.0166 | 0.0166 | 0.0166 | 0.0166 | 0.0166 | 0.0166 |
| | Chronos | **0.0166** | **0.0166** | **0.0165** | **0.0165** | 0.0165 | **0.0165** | **0.0165** | **0.0165** | **0.0165** | **0.0165** |
| | Sundial | 0.0168 | 0.0168 | 0.0166 | **0.0165** | **0.0164** | 0.0168 | 0.0168 | 0.0166 | 0.0168 | 0.0169 |

Table 5.8: MAE for fractional change forecasting, year-increment configuration (1–10 years of daily training data). Bold indicates best performance; underline indicates second best.

| Horizon | Method | 1y | 2y | 3y | 4y | 5y | 6y | 7y | 8y | 9y | 10y |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1d | Naive | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 |
| | ARIMA | **0.0123** | **0.0123** | 0.0123 | **0.0122** | **0.0122** | **0.0122** | **0.0122** | **0.0122** | **0.0122** | **0.0122** |
| | TimesFM | 0.0125 | **0.0123** | 0.0125 | 0.0124 | 0.0124 | 0.0124 | 0.0124 | 0.0124 | 0.0124 | 0.0124 |
| | Chronos | 0.0124 | 0.0124 | 0.0124 | 0.0123 | 0.0123 | 0.0123 | 0.0123 | 0.0123 | 0.0123 | 0.0123 |
| | Sundial | 0.0125 | **0.0123** | **0.0121** | 0.0123 | **0.0122** | **0.0122** | 0.0123 | **0.0122** | 0.0123 | **0.0122** |
| 5d | Naive | 0.0168 | 0.0168 | 0.0168 | 0.0168 | 0.0168 | 0.0168 | 0.0168 | 0.0168 | 0.0168 | 0.0168 |
| | ARIMA | **0.0123** | **0.0123** | 0.0123 | **0.0122** | **0.0122** | **0.0122** | **0.0122** | **0.0123** | **0.0123** | **0.0123** |
| | TimesFM | 0.0127 | 0.0126 | 0.0124 | 0.0124 | 0.0124 | 0.0124 | 0.0123 | **0.0123** | **0.0123** | **0.0123** |
| | Chronos | 0.0124 | **0.0123** | 0.0123 | 0.0123 | 0.0123 | 0.0123 | 0.0123 | **0.0123** | **0.0123** | **0.0123** |
| | Sundial | 0.0126 | **0.0123** | **0.0122** | **0.0122** | **0.0122** | 0.0123 | **0.0122** | 0.0124 | 0.0125 | **0.0123** |
| 21d | Naive | 0.0177 | 0.0177 | 0.0177 | 0.0177 | 0.0177 | 0.0177 | 0.0177 | 0.0177 | 0.0177 | 0.0177 |
| | ARIMA | **0.0123** | **0.0123** | 0.0123 | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** |
| | TimesFM | 0.0124 | 0.0124 | 0.0124 | 0.0124 | 0.0124 | 0.0124 | 0.0124 | 0.0124 | 0.0125 | 0.0125 |
| | Chronos | 0.0124 | **0.0123** | 0.0123 | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** |
| | Sundial | 0.0127 | **0.0123** | **0.0121** | 0.0124 | 0.0124 | 0.0124 | 0.0124 | **0.0123** | 0.0124 | 0.0126 |
| 63d | Naive | 0.0148 | 0.0148 | 0.0148 | 0.0148 | 0.0148 | 0.0148 | 0.0148 | 0.0148 | 0.0148 | 0.0148 |
| | ARIMA | **0.0123** | **0.0123** | 0.0123 | **0.0123** | 0.0123 | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** |
| | TimesFM | 0.0125 | 0.0124 | 0.0123 | 0.0124 | 0.0124 | 0.0124 | **0.0123** | **0.0123** | **0.0123** | **0.0123** |
| | Chronos | 0.0124 | **0.0123** | 0.0123 | **0.0123** | 0.0123 | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** |
| | Sundial | 0.0125 | 0.0125 | **0.0122** | **0.0123** | **0.0122** | 0.0126 | 0.0126 | 0.0124 | 0.0125 | 0.0126 |

Table 5.9: Mean Directional Accuracy (%) for fractional change forecasting, year-increment configuration (1–10 years of daily training data). Bold indicates best performance; underline indicates second best.

| Horizon | Method | 1y | 2y | 3y | 4y | 5y | 6y | 7y | 8y | 9y | 10y |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1d | Naive | 50.4000 | 50.4000 | 50.4000 | 50.4000 | 50.4000 | 50.4000 | 50.4000 | 50.4000 | 50.4000 | 50.4000 |
| | ARIMA | 0.0000 | 0.0000 | 0.0000 | **54.4000** | 51.6000 | 52.4000 | **52.4000** | 54.0000 | **54.4000** | 52.8000 |
| | TimesFM | 46.0000 | 52.8000 | 45.2000 | 52.8000 | 52.4000 | 52.0000 | 50.4000 | 48.0000 | 48.4000 | 48.4000 |
| | Chronos | 49.2000 | 50.0000 | 47.2000 | 50.0000 | 50.8000 | 50.4000 | 49.2000 | 50.4000 | 50.4000 | 50.4000 |
| | Sundial | **51.0000** | **53.2000** | **56.4000** | 54.2000 | **55.0000** | **53.6000** | 52.0000 | **55.4000** | 53.0000 | **55.8000** |
| 5d | Naive | **49.6000** | 49.6000 | 49.6000 | 49.6000 | 49.6000 | 49.6000 | 49.6000 | 49.6000 | 49.6000 | 49.6000 |
| | ARIMA | 0.0000 | 0.0000 | 0.0000 | 53.2000 | 53.2000 | 53.2000 | 53.6000 | 52.8000 | 53.2000 | 51.2000 |
| | TimesFM | 46.4000 | 51.2000 | 55.2000 | 55.6000 | **54.0000** | **54.8000** | **57.6000** | **54.8000** | **53.6000** | **53.6000** |
| | Chronos | 48.4000 | 51.6000 | 50.8000 | 52.0000 | 52.0000 | 52.8000 | 51.6000 | 51.2000 | 51.2000 | 51.2000 |
| | Sundial | 48.4000 | **56.8000** | **55.6000** | **58.4000** | **54.0000** | 54.0000 | 54.8000 | 50.8000 | 50.8000 | 52.4000 |
| 21d | Naive | **52.0000** | 52.0000 | 52.0000 | 52.0000 | 52.0000 | 52.0000 | 52.0000 | 52.0000 | 52.0000 | 52.0000 |
| | ARIMA | 0.0000 | 0.0000 | 0.0000 | **54.8000** | 52.8000 | 53.2000 | 52.8000 | 52.8000 | 52.8000 | 52.4000 |
| | TimesFM | 48.8000 | 52.8000 | 50.0000 | 49.6000 | 46.8000 | 50.8000 | 50.8000 | 48.4000 | 46.8000 | 46.8000 |
| | Chronos | 47.6000 | **54.0000** | **54.0000** | 54.4000 | **53.2000** | **54.0000** | **54.4000** | **54.8000** | **54.8000** | **54.8000** |
| | Sundial | 44.8000 | 52.0000 | 53.6000 | 50.8000 | 51.2000 | 51.2000 | 52.0000 | 54.0000 | 50.4000 | 46.0000 |
| 63d | Naive | **58.8000** | **58.8000** | **58.8000** | **58.8000** | **58.8000** | **58.8000** | **58.8000** | **58.8000** | **58.8000** | **58.8000** |
| | ARIMA | 0.0000 | 0.0000 | 0.0000 | 54.4000 | 54.4000 | 54.4000 | 54.4000 | 54.0000 | 54.0000 | 54.0000 |
| | TimesFM | 48.0000 | 50.8000 | 50.0000 | 50.0000 | 47.6000 | 48.0000 | 49.6000 | 48.8000 | 48.0000 | 48.0000 |
| | Chronos | 47.6000 | 53.6000 | 54.4000 | 54.4000 | 54.4000 | 54.4000 | 54.4000 | 54.8000 | 54.8000 | 54.8000 |
| | Sundial | 52.0000 | 49.6000 | 57.2000 | 50.0000 | 55.6000 | 44.4000 | 46.4000 | 50.4000 | 46.0000 | 49.2000 |

### Limited-Data Regime

Tables 5.10–5.12 present fractional change results under sub-year training lengths with logarithmic spacing. Unlike the year-increment configuration, ARIMA converges at all training lengths.

For RMSE, all models except Naïve cluster tightly (0.0164–0.0189). ARIMA is strongest at 1d and remains among the best from 5d onward, while Chronos becomes increasingly competitive at the longer horizons. MAE shows the same broad clustering, with exact rankings less stable once the differences become very small.

For MDA, ARIMA records 0.0% at all training lengths and horizons, a consistent failure in directional prediction under this regime. Outside ARIMA, leadership is split across Naïve and the foundation models, with no single foundation model dominating across horizons. At 63d, Naïve achieves the highest MDA (58.8%) across all training lengths.

Table 5.10: RMSE for fractional change forecasting, logarithmically-spaced training lengths (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Horizon | Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1d | Naive | 0.0248 | 0.0248 | 0.0248 | 0.0248 | 0.0248 | 0.0248 | 0.0248 | 0.0248 | 0.0248 | 0.0248 |
| | ARIMA | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** |
| | TimesFM | 0.0171 | 0.0171 | 0.0170 | 0.0170 | 0.0169 | 0.0169 | 0.0170 | 0.0169 | 0.0168 | 0.0169 |
| | Chronos | 0.0171 | 0.0170 | 0.0169 | 0.0169 | 0.0169 | 0.0168 | 0.0168 | 0.0168 | 0.0167 | 0.0167 |
| | Sundial | 0.0171 | 0.0171 | 0.0173 | 0.0169 | 0.0171 | 0.0168 | 0.0167 | 0.0167 | 0.0167 | 0.0167 |
| 5d | Naive | 0.0221 | 0.0221 | 0.0221 | 0.0221 | 0.0221 | 0.0221 | 0.0221 | 0.0221 | 0.0221 | 0.0221 |
| | ARIMA | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** |
| | TimesFM | 0.0170 | 0.0170 | 0.0170 | 0.0170 | 0.0168 | 0.0169 | 0.0170 | 0.0168 | 0.0168 | 0.0168 |
| | Chronos | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** |
| | Sundial | 0.0169 | 0.0168 | 0.0167 | 0.0169 | 0.0169 | 0.0170 | 0.0167 | **0.0166** | 0.0167 | 0.0167 |
| 21d | Naive | 0.0219 | 0.0219 | 0.0219 | 0.0219 | 0.0219 | 0.0219 | 0.0219 | 0.0219 | 0.0219 | 0.0219 |
| | ARIMA | **0.0166** | **0.0166** | **0.0166** | 0.0166 | 0.0166 | 0.0166 | 0.0166 | **0.0166** | 0.0166 | **0.0166** |
| | TimesFM | 0.0167 | 0.0167 | 0.0168 | 0.0166 | 0.0168 | 0.0167 | 0.0166 | 0.0167 | 0.0166 | **0.0166** |
| | Chronos | 0.0167 | 0.0167 | **0.0166** | **0.0165** | **0.0165** | **0.0165** | **0.0165** | **0.0166** | 0.0166 | **0.0166** |
| | Sundial | 0.0168 | 0.0168 | 0.0167 | 0.0166 | 0.0168 | 0.0167 | **0.0165** | 0.0168 | **0.0165** | 0.0169 |
| 63d | Naive | 0.0190 | 0.0190 | 0.0190 | 0.0190 | 0.0190 | 0.0190 | 0.0190 | 0.0190 | 0.0190 | 0.0190 |
| | ARIMA | **0.0166** | 0.0166 | **0.0166** | **0.0166** | 0.0166 | 0.0166 | 0.0166 | **0.0166** | **0.0166** | **0.0166** |
| | TimesFM | 0.0189 | 0.0168 | **0.0166** | 0.0167 | 0.0167 | 0.0166 | 0.0166 | 0.0167 | **0.0166** | **0.0166** |
| | Chronos | 0.0168 | 0.0168 | 0.0167 | **0.0166** | **0.0165** | **0.0165** | **0.0165** | **0.0166** | **0.0166** | **0.0166** |
| | Sundial | 0.0168 | **0.0164** | 0.0168 | 0.0167 | 0.0171 | 0.0168 | 0.0166 | 0.0168 | **0.0166** | 0.0167 |

Table 5.11: MAE for fractional change forecasting, logarithmically-spaced training lengths (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Horizon | Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1d | Naive | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 |
| | ARIMA | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** |
| | TimesFM | 0.0126 | 0.0126 | 0.0126 | 0.0125 | 0.0125 | 0.0125 | 0.0126 | 0.0126 | 0.0125 | 0.0125 |
| | Chronos | 0.0126 | 0.0126 | 0.0126 | 0.0126 | 0.0125 | 0.0125 | 0.0125 | 0.0124 | 0.0124 | 0.0124 |
| | Sundial | 0.0127 | 0.0129 | 0.0130 | 0.0126 | 0.0126 | 0.0125 | 0.0124 | 0.0125 | 0.0124 | 0.0124 |
| 5d | Naive | 0.0168 | 0.0168 | 0.0168 | 0.0168 | 0.0168 | 0.0168 | 0.0168 | 0.0168 | 0.0168 | 0.0168 |
| | ARIMA | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** |
| | TimesFM | 0.0127 | 0.0128 | 0.0128 | 0.0128 | 0.0126 | 0.0127 | 0.0127 | 0.0125 | 0.0124 | 0.0126 |
| | Chronos | 0.0125 | 0.0125 | 0.0124 | 0.0124 | 0.0124 | 0.0124 | 0.0124 | 0.0124 | 0.0124 | 0.0124 |
| | Sundial | 0.0127 | 0.0125 | 0.0127 | 0.0127 | 0.0126 | 0.0128 | 0.0125 | 0.0124 | 0.0124 | **0.0123** |
| 21d | Naive | 0.0177 | 0.0177 | 0.0177 | 0.0177 | 0.0177 | 0.0177 | 0.0177 | 0.0177 | 0.0177 | 0.0177 |
| | ARIMA | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | 0.0123 | **0.0123** | **0.0123** | **0.0123** |
| | TimesFM | 0.0124 | 0.0125 | 0.0125 | 0.0125 | 0.0125 | 0.0124 | 0.0124 | 0.0125 | 0.0124 | 0.0124 |
| | Chronos | 0.0125 | 0.0126 | 0.0125 | 0.0124 | **0.0123** | **0.0123** | 0.0123 | 0.0124 | **0.0123** | **0.0123** |
| | Sundial | 0.0125 | 0.0125 | 0.0124 | 0.0124 | 0.0127 | 0.0124 | **0.0121** | 0.0126 | **0.0123** | 0.0126 |
| 63d | Naive | 0.0148 | 0.0148 | 0.0148 | 0.0148 | 0.0148 | 0.0148 | 0.0148 | 0.0148 | 0.0148 | 0.0148 |
| | ARIMA | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** |
| | TimesFM | 0.0141 | 0.0125 | **0.0123** | 0.0125 | 0.0125 | 0.0124 | 0.0125 | 0.0125 | 0.0125 | 0.0125 |
| | Chronos | 0.0127 | 0.0128 | 0.0127 | 0.0126 | 0.0124 | 0.0125 | 0.0125 | 0.0125 | 0.0124 | 0.0124 |
| | Sundial | 0.0125 | 0.0124 | 0.0125 | 0.0125 | 0.0129 | 0.0126 | 0.0124 | 0.0124 | 0.0124 | 0.0125 |

Table 5.12: Mean Directional Accuracy (%) for fractional change forecasting, logarithmically-spaced training lengths (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Horizon | Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1d | Naive | 50.4000 | 50.4000 | 50.4000 | 50.4000 | 50.4000 | 50.4000 | 50.4000 | 50.4000 | 50.4000 | 50.4000 |
| | ARIMA | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | TimesFM | **54.0000** | 51.2000 | **52.4000** | 52.8000 | **54.4000** | **54.8000** | 48.4000 | 50.4000 | 49.2000 | 47.6000 |
| | Chronos | 51.6000 | **53.6000** | 50.0000 | 54.4000 | 52.4000 | 51.6000 | 50.4000 | 51.2000 | **52.4000** | **51.2000** |
| | Sundial | 51.2000 | 47.8000 | 51.0000 | **54.6000** | 49.8000 | 51.2000 | **52.2000** | **52.2000** | 50.0000 | 50.0000 |
| 5d | Naive | 49.6000 | 49.6000 | 49.6000 | 49.6000 | 49.6000 | 49.6000 | 49.6000 | 49.6000 | 49.6000 | 49.6000 |
| | ARIMA | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | TimesFM | 49.6000 | 48.8000 | 50.4000 | 51.6000 | **52.4000** | 51.2000 | **55.6000** | **51.2000** | 52.0000 | 48.8000 |
| | Chronos | **52.8000** | **52.8000** | **51.6000** | **54.4000** | 52.0000 | **52.0000** | 51.6000 | **51.2000** | 52.8000 | **52.8000** |
| | Sundial | 49.2000 | 50.8000 | **51.6000** | 48.4000 | 46.8000 | 46.8000 | 48.0000 | 50.0000 | **54.0000** | 51.6000 |
| 21d | Naive | 52.0000 | **52.0000** | 52.0000 | 52.0000 | **52.0000** | 52.0000 | 52.0000 | **52.0000** | **52.0000** | **52.0000** |
| | ARIMA | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | TimesFM | **54.4000** | **52.0000** | 50.0000 | **54.0000** | 49.6000 | 50.0000 | 50.0000 | 49.2000 | 50.8000 | 50.0000 |
| | Chronos | 52.4000 | 50.8000 | 49.6000 | 53.6000 | **52.0000** | **52.4000** | 52.4000 | 49.6000 | 48.0000 | 47.6000 |
| | Sundial | 48.0000 | 51.2000 | **53.6000** | 53.6000 | 49.2000 | 49.6000 | **56.4000** | 51.6000 | 51.6000 | 47.2000 |
| 63d | Naive | **58.8000** | **58.8000** | **58.8000** | **58.8000** | **58.8000** | **58.8000** | **58.8000** | **58.8000** | **58.8000** | **58.8000** |
| | ARIMA | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | TimesFM | 54.4000 | 52.8000 | 54.0000 | 52.8000 | 50.0000 | 51.2000 | 48.4000 | 48.0000 | 48.4000 | 48.0000 |
| | Chronos | 51.2000 | 50.0000 | 50.0000 | 54.4000 | 54.0000 | 48.4000 | 48.0000 | 46.4000 | 50.0000 | 47.6000 |
| | Sundial | 48.0000 | 50.0000 | 50.4000 | 52.4000 | 48.0000 | 49.6000 | 51.6000 | 49.2000 | 52.0000 | 49.2000 |

## 5.2 RQ2: Temporal Frequency Effects

RQ2 examines whether switching from daily to hourly training data changes model performance. The tables in this section replicate the same configurations used in Section 5.1 with hourly-resolution input data. Unlike the daily experiments which evaluate four forecast horizons (1d, 5d, 21d, 63d), hourly configurations are evaluated at the one-step-ahead (1-hour) forecast horizon only.

### 5.2.1 Raw Price Forecasting (Hourly)

#### Year-Increment Configuration

Tables 5.13–5.15 present MAPE, RMSE, and MAE for raw price forecasting at hourly frequency with year-increment training.

Naïve and ARIMA dominate. Naïve achieves the lowest MAPE (0.3994%) at 1–6 years, while ARIMA marginally improves from 7 years onward (0.3992%). TimesFM (0.45%) and Chronos (0.53%) remain stable across all training lengths. Sundial is competitive at short training lengths (0.42% at 1–2 years) but exhibits severe instability from 3 years onward, with MAPE rising to 1.64% at 10 years and RMSE increasing well above the baseline range. RMSE rankings largely mirror MAPE, with one exception: Naïve achieves the lowest RMSE at all ten training lengths, including 7–10 years where ARIMA leads in MAPE.

Table 5.13: MAPE (%) for raw price forecasting, year-increment configuration (1–10 years of hourly training data). Bold indicates best performance; underline indicates second best.

| Method | 1y | 2y | 3y | 4y | 5y | 6y | 7y | 8y | 9y | 10y |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** | 0.3994 | 0.3994 | 0.3994 | 0.3994 |
| ARIMA | **0.3994** | **0.3994** | 0.3995 | 0.3995 | 0.3995 | 0.3995 | **0.3992** | **0.3992** | **0.3992** | **0.3992** |
| TimesFM | 0.4492 | 0.4564 | 0.4564 | 0.4564 | 0.4564 | 0.4564 | 0.4564 | 0.4564 | 0.4564 | 0.4564 |
| Chronos | 0.5288 | 0.5297 | 0.5297 | 0.5297 | 0.5297 | 0.5297 | 0.5297 | 0.5297 | 0.5297 | 0.5297 |
| Sundial | 0.4198 | 0.4167 | 0.4894 | 1.2578 | 1.5394 | 1.2949 | 1.1967 | 1.4479 | 1.4988 | 1.6396 |

Table 5.14: RMSE for raw price forecasting, year-increment configuration (1–10 years of hourly training data). Bold indicates best performance; underline indicates second best.

| Method | 1y | 2y | 3y | 4y | 5y | 6y | 7y | 8y | 9y | 10y |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** |
| ARIMA | **1.8107** | **1.8107** | 1.8129 | 1.8129 | 1.8130 | 1.8129 | 1.8120 | 1.8120 | 1.8120 | 1.8121 |
| TimesFM | 1.9612 | 1.9788 | 1.9788 | 1.9788 | 1.9788 | 1.9788 | 1.9788 | 1.9788 | 1.9788 | 1.9788 |
| Chronos | 2.2565 | 2.2589 | 2.2589 | 2.2589 | 2.2589 | 2.2589 | 2.2589 | 2.2589 | 2.2589 | 2.2589 |
| Sundial | 1.8709 | 1.8726 | 2.1079 | 5.0993 | 6.0003 | 4.7209 | 4.3420 | 5.0167 | 5.1586 | 5.5942 |

Table 5.15: MAE for raw price forecasting, year-increment configuration (1–10 years of hourly training data). Bold indicates best performance; underline indicates second best.

| Method | 1y | 2y | 3y | 4y | 5y | 6y | 7y | 8y | 9y | 10y |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** | 1.2276 | 1.2276 | 1.2276 | 1.2276 |
| ARIMA | **1.2276** | **1.2276** | 1.2278 | 1.2277 | 1.2278 | 1.2277 | **1.2268** | **1.2268** | **1.2268** | **1.2268** |
| TimesFM | 1.3838 | 1.4010 | 1.4010 | 1.4010 | 1.4010 | 1.4010 | 1.4010 | 1.4010 | 1.4010 | 1.4010 |
| Chronos | 1.6469 | 1.6486 | 1.6486 | 1.6486 | 1.6486 | 1.6486 | 1.6486 | 1.6486 | 1.6486 | 1.6486 |
| Sundial | 1.2967 | 1.2835 | 1.5036 | 3.5530 | 4.3078 | 3.7647 | 3.5298 | 4.3133 | 4.4530 | 4.9791 |

**Limited-Data Regime**

Tables 5.16–5.18 present MAPE, RMSE, and MAE for raw price forecasting at hourly frequency with log-spaced training.

Naïve and ARIMA are tied at 0.3994% MAPE across all training lengths. Sundial is the strongest foundation model (0.42–0.43%), followed by TimesFM (0.43–0.46%) and Chronos (0.49–0.53%). Unlike the year-increment configuration, Sundial is stable across all training lengths. ARIMA also remains stable, contrasting with the daily limited-data regime where it failed to converge at several training lengths. RMSE and MAE rankings are consistent with MAPE.

Table 5.16: MAPE (%) for raw price forecasting, logarithmically-spaced training lengths (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** |
| ARIMA | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** |
| TimesFM | 0.4493 | 0.4447 | 0.4440 | 0.4482 | 0.4434 | 0.4573 | 0.4456 | 0.4306 | 0.4401 | 0.4394 |
| Chronos | 0.4968 | 0.4995 | 0.4959 | 0.4933 | 0.4992 | 0.4945 | 0.5043 | 0.5288 | 0.5287 | 0.5349 |
| Sundial | 0.4239 | 0.4276 | 0.4269 | 0.4260 | 0.4236 | 0.4236 | 0.4274 | 0.4266 | 0.4243 | 0.4292 |

Table 5.17: RMSE for raw price forecasting, logarithmically-spaced training lengths (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** |
| ARIMA | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** |
| TimesFM | 1.9672 | 1.9444 | 1.9485 | 1.9514 | 1.9665 | 1.9922 | 1.9598 | 1.9147 | 1.9419 | 1.9352 |
| Chronos | 2.1367 | 2.1446 | 2.1405 | 2.1374 | 2.1571 | 2.1435 | 2.1736 | 2.2421 | 2.2475 | 2.2708 |
| Sundial | 1.8950 | 1.8997 | 1.8972 | 1.8945 | 1.8912 | 1.8850 | 1.9009 | 1.8924 | 1.8850 | 1.9044 |

Table 5.18: MAE for raw price forecasting, logarithmically-spaced training lengths (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** |
| ARIMA | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** |
| TimesFM | 1.3859 | 1.3704 | 1.3683 | 1.3789 | 1.3668 | 1.4114 | 1.3796 | 1.3294 | 1.3606 | 1.3583 |
| Chronos | 1.5454 | 1.5520 | 1.5435 | 1.5380 | 1.5573 | 1.5429 | 1.5730 | 1.6448 | 1.6465 | 1.6679 |
| Sundial | 1.3045 | 1.3159 | 1.3139 | 1.3118 | 1.3063 | 1.3059 | 1.3205 | 1.3222 | 1.3147 | 1.3303 |

### 5.2.2 Fractional Change Forecasting (Hourly)

As in Section 5.1.2, MAPE is excluded from the fractional change analysis.

#### Year-Increment Configuration

Tables 5.19–5.21 present RMSE, MAE, and MDA for fractional change forecasting at hourly frequency with year-increment training.

For RMSE, all models except Naïve converge to 0.0060, while Naïve records 0.0084. MAE mirrors this pattern.

For MDA, ARIMA records near-zero accuracy at 1–2 years of training (0.34%) before recovering to 50.3–51.4% from 3 years onward. Sundial achieves the highest MDA at several training lengths (up to 52.1% at 3 years). Naïve records a fixed 49.0%. TimesFM (48.7%) and Chronos (48.1%) remain below the 50% baseline.

Table 5.19: RMSE for fractional change forecasting, year-increment configuration (1–10 years of hourly training data). Bold indicates best performance; underline indicates second best.

| Method | 1y | 2y | 3y | 4y | 5y | 6y | 7y | 8y | 9y | 10y |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 0.0084 | 0.0084 | 0.0084 | 0.0084 | 0.0084 | 0.0084 | 0.0084 | 0.0084 | 0.0084 | 0.0084 |
| ARIMA | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** |
| TimesFM | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** |
| Chronos | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** |
| Sundial | 0.0061 | **0.0060** | **0.0060** | **0.0060** | 0.0061 | 0.0061 | **0.0060** | **0.0060** | 0.0061 | **0.0060** |

Table 5.20: MAE for fractional change forecasting, year-increment configuration (1–10 years of hourly training data). Bold indicates best performance; underline indicates second best.

| Method | 1y | 2y | 3y | 4y | 5y | 6y | 7y | 8y | 9y | 10y |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 0.0058 | 0.0058 | 0.0058 | 0.0058 | 0.0058 | 0.0058 | 0.0058 | 0.0058 | 0.0058 | 0.0058 |
| ARIMA | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** |
| TimesFM | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** |
| Chronos | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** |
| Sundial | 0.0041 | **0.0040** | **0.0040** | **0.0040** | **0.0040** | 0.0041 | **0.0040** | **0.0040** | **0.0040** | **0.0040** |

Table 5.21: Mean Directional Accuracy (%) for fractional change forecasting, year-increment configuration (1–10 years of hourly training data). Bold indicates best performance; underline indicates second best.

| Method | 1y | 2y | 3y | 4y | 5y | 6y | 7y | 8y | 9y | 10y |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **49.0286** | 49.0286 | 49.0286 | 49.0286 | 49.0286 | 49.0286 | 49.0286 | 49.0286 | 49.0286 | 49.0286 |
| ARIMA | 0.3429 | 0.3429 | 50.2857 | **51.2000** | **51.0857** | **51.0286** | 50.6286 | **51.3714** | **50.8000** | **51.3714** |
| TimesFM | 48.1714 | 48.6857 | 48.6857 | 48.6857 | 48.6857 | 48.6857 | 48.6857 | 48.6857 | 48.6857 | 48.6857 |
| Chronos | 47.6571 | 48.1143 | 48.1143 | 48.1143 | 48.1143 | 48.1143 | 48.1143 | 48.1143 | 48.1143 | 48.1143 |
| Sundial | 48.4000 | **50.8000** | **52.1143** | 49.5429 | 50.6857 | 47.8286 | **51.7143** | 50.7429 | 49.0857 | 49.4857 |

**Limited-Data Regime**

Tables 5.22–5.24 present RMSE, MAE, and MDA for fractional change forecasting at hourly frequency with log-spaced training.

For RMSE and MAE, all models except Naïve cluster tightly at 0.0060–0.0061. ARIMA records 0.34% MDA at all training lengths. Sundial achieves the highest MDA at all ten training lengths (49.5–51.3%), making it the strongest directional model in this configuration, although the margin remains close to the 50% baseline. Naïve records a fixed 49.0%, while TimesFM and Chronos range from 47.7–49.4%.

Table 5.22: RMSE for fractional change forecasting, logarithmically-spaced training lengths (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 0.0084 | 0.0084 | 0.0084 | 0.0084 | 0.0084 | 0.0084 | 0.0084 | 0.0084 | 0.0084 | 0.0084 |
| ARIMA | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** |
| TimesFM | 0.0061 | 0.0061 | 0.0061 | **0.0060** | **0.0060** | 0.0061 | 0.0061 | **0.0060** | **0.0060** | **0.0060** |
| Chronos | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** |
| Sundial | 0.0061 | 0.0061 | 0.0061 | 0.0061 | 0.0061 | 0.0061 | **0.0060** | **0.0060** | **0.0060** | 0.0061 |

Table 5.23: MAE for fractional change forecasting, logarithmically-spaced training lengths (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 0.0058 | 0.0058 | 0.0058 | 0.0058 | 0.0058 | 0.0058 | 0.0058 | 0.0058 | 0.0058 | 0.0058 |
| ARIMA | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** |
| TimesFM | 0.0041 | 0.0041 | 0.0041 | 0.0041 | 0.0041 | 0.0041 | 0.0041 | 0.0041 | **0.0040** | **0.0040** |
| Chronos | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** |
| Sundial | 0.0041 | 0.0041 | 0.0041 | 0.0041 | 0.0041 | 0.0041 | 0.0041 | **0.0040** | **0.0040** | 0.0041 |

Table 5.24: Mean Directional Accuracy (%) for fractional change forecasting, logarithmically-spaced training lengths (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 49.0286 | 49.0286 | 49.0286 | 49.0286 | 49.0286 | 49.0286 | 49.0286 | 49.0286 | 49.0286 | 49.0286 |
| ARIMA | 0.3429 | 0.3429 | 0.3429 | 0.3429 | 0.3429 | 0.3429 | 0.3429 | 0.3429 | 0.3429 | 0.3429 |
| TimesFM | 49.2000 | 48.9714 | 49.3714 | 48.6286 | 48.5143 | 49.3714 | 47.7143 | 48.2286 | 49.0286 | 49.1429 |
| Chronos | 48.6286 | 48.9143 | 49.3714 | 47.7714 | 48.1143 | 48.6286 | 48.4000 | 48.2286 | 48.2857 | 48.5143 |
| Sundial | **50.8000** | **50.0000** | **50.6857** | **50.5143** | **50.6286** | **51.0857** | **51.3143** | **51.3143** | **49.5429** | **49.6571** |

## 5.3 RQ3: Market Disruption Resilience

RQ3 examines model performance during three market disruption periods. For each period, models are trained on data immediately preceding the crisis window using logarithmically-spaced training lengths from 25 to 249 days. All RQ3 tables report results for the 1 step ahead forecast only.

### 5.3.1 2008 Global Financial Crisis

The 2008 evaluation window covers January 22, 2008 to March 19, 2008. Tables 5.25–5.27 present MAPE, RMSE, and MAE for raw price forecasting under log-spaced training.

Naïve achieves the lowest MAPE (0.5846%) across all training lengths, followed closely by ARIMA (0.5846–0.6067%). Foundation models are consistently higher: TimesFM ranges from 0.61–0.71%, Sundial from 0.60–0.68%, and Chronos from 0.66–0.88%. RMSE and MAE rankings are consistent with MAPE.

Table 5.25: MAPE (%) for raw price forecasting, 2008 Financial Crisis with logarithmically-spaced training (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **0.5846** | **0.5846** | **0.5846** | **0.5846** | **0.5846** | **0.5846** | **0.5846** | **0.5846** | **0.5846** | **0.5846** |
| ARIMA | 0.5933 | **0.5846** | **0.5846** | **0.5846** | **0.5846** | 0.6067 | 0.5863 | 0.5860 | 0.5859 | 0.5858 |
| TimesFM | 0.7115 | 0.6764 | 0.6865 | 0.6765 | 0.6568 | 0.6266 | 0.6059 | 0.6342 | 0.6393 | 0.6387 |
| Chronos | 0.6570 | 0.6865 | 0.6771 | 0.6769 | 0.7335 | 0.8848 | 0.8485 | 0.7413 | 0.6842 | 0.7342 |
| Sundial | 0.6655 | 0.6420 | 0.6640 | 0.6552 | 0.6823 | 0.6718 | 0.6130 | 0.6246 | 0.6106 | 0.5990 |

Table 5.26: RMSE for raw price forecasting, 2008 Financial Crisis with logarithmically-spaced training (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **0.2819** | **0.2819** | **0.2819** | **0.2819** | **0.2819** | **0.2819** | 0.2819 | 0.2819 | 0.2819 | 0.2819 |
| ARIMA | 0.2820 | **0.2819** | **0.2819** | **0.2819** | **0.2819** | 0.2842 | **0.2806** | **0.2806** | **0.2806** | **0.2805** |
| TimesFM | 0.3043 | 0.3036 | 0.3036 | 0.2981 | 0.2939 | 0.2914 | 0.2865 | 0.2951 | 0.2946 | 0.2952 |
| Chronos | 0.2991 | 0.3078 | 0.3051 | 0.3043 | 0.3148 | 0.3470 | 0.3378 | 0.3124 | 0.3019 | 0.3124 |
| Sundial | 0.2981 | 0.2942 | 0.3016 | 0.2977 | 0.3054 | 0.3016 | 0.2832 | 0.2870 | 0.2815 | 0.2806 |

Table 5.27: MAE for raw price forecasting, 2008 Financial Crisis with logarithmically-spaced training (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **0.1734** | **0.1734** | **0.1734** | **0.1734** | **0.1734** | **0.1734** | **0.1734** | **0.1734** | **0.1734** | **0.1734** |
| ARIMA | 0.1759 | **0.1734** | **0.1734** | **0.1734** | **0.1734** | 0.1796 | 0.1738 | 0.1737 | 0.1737 | 0.1736 |
| TimesFM | 0.2098 | 0.2006 | 0.2032 | 0.2005 | 0.1949 | 0.1862 | 0.1799 | 0.1883 | 0.1900 | 0.1895 |
| Chronos | 0.1943 | 0.2028 | 0.2002 | 0.2002 | 0.2162 | 0.2592 | 0.2487 | 0.2182 | 0.2022 | 0.2164 |
| Sundial | 0.1976 | 0.1903 | 0.1972 | 0.1944 | 0.2024 | 0.1989 | 0.1815 | 0.1849 | 0.1805 | 0.1772 |

### 5.3.2 2020 COVID-19 Crash

The 2020 evaluation window covers February 28, 2020 to April 27, 2020. Tables 5.28–5.30 present raw price forecasting results under log-spaced training.

ARIMA records the lowest MAPE across most training lengths (1.162–1.173%). Sundial is competitive and leads at several training lengths (1.154–1.200%). Naïve achieves 1.182%. Chronos records high MAPE (1.215–1.302%) and is among the weakest models in this setting. However, RMSE rankings diverge from MAPE: Sundial achieves the lowest RMSE from 116 days of training onward, while ARIMA remains the more frequent MAPE leader.

Table 5.28: MAPE (%) for raw price forecasting, 2020 COVID-19 Crash with logarithmically-spaced training (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 1.1821 | 1.1821 | 1.1821 | 1.1821 | 1.1821 | 1.1821 | 1.1821 | 1.1821 | 1.1821 | 1.1821 |
| ARIMA | **1.1709** | **1.1685** | **1.1729** | **1.1664** | **1.1717** | 1.1709 | **1.1705** | 1.1641 | 1.1629 | **1.1616** |
| TimesFM | 1.2358 | 1.2439 | 1.2099 | 1.2155 | 1.1984 | 1.2056 | 1.2349 | 1.2181 | 1.2297 | 1.1936 |
| Chronos | 1.2282 | 1.2149 | 1.2195 | 1.2212 | 1.2425 | 1.2696 | 1.3018 | 1.2732 | 1.2512 | 1.2320 |
| Sundial | 1.1889 | 1.2001 | 1.1832 | 1.1749 | 1.1838 | **1.1681** | 1.1785 | **1.1536** | **1.1606** | 1.1730 |

Table 5.29: RMSE for raw price forecasting, 2020 COVID-19 Crash with logarithmically-spaced training (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 2.7548 | 2.7548 | 2.7548 | 2.7548 | 2.7548 | 2.7548 | 2.7548 | 2.7548 | 2.7548 | 2.7548 |
| ARIMA | **2.6856** | 2.6817 | **2.6948** | **2.6768** | 2.6921 | **2.6850** | 2.6843 | 2.6738 | 2.6744 | 2.6758 |
| TimesFM | 2.7971 | 2.7952 | 2.7424 | 2.7170 | **2.6756** | 2.7209 | 2.7057 | 2.7400 | 2.7501 | 2.7063 |
| Chronos | 2.6980 | **2.6792** | 2.6995 | 2.7376 | 2.7780 | 2.8478 | 2.9101 | 2.8395 | 2.7800 | 2.7578 |
| Sundial | 2.7324 | 2.7168 | 2.7323 | 2.6963 | 2.7132 | 2.6947 | **2.6820** | **2.6334** | **2.6373** | **2.6487** |

Table 5.30: MAE for raw price forecasting, 2020 COVID-19 Crash with logarithmically-spaced training (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 1.8353 | 1.8353 | 1.8353 | 1.8353 | 1.8353 | 1.8353 | 1.8353 | 1.8353 | 1.8353 | 1.8353 |
| ARIMA | **1.8219** | **1.8182** | **1.8226** | **1.8147** | **1.8205** | 1.8187 | **1.8180** | 1.8078 | 1.8058 | **1.8039** |
| TimesFM | 1.9194 | 1.9320 | 1.8823 | 1.8856 | 1.8613 | 1.8747 | 1.9189 | 1.8943 | 1.9110 | 1.8524 |
| Chronos | 1.9103 | 1.8889 | 1.8979 | 1.8996 | 1.9337 | 1.9749 | 2.0231 | 1.9791 | 1.9467 | 1.9175 |
| Sundial | 1.8434 | 1.8606 | 1.8335 | 1.8205 | 1.8336 | **1.8117** | 1.8274 | **1.7918** | **1.8014** | 1.8215 |

### 5.3.3 2022 Market Downturn

The 2022 evaluation window covers March 31, 2022 to May 27, 2022. Tables 5.31–5.33 present raw price forecasting results under log-spaced training.

Naïve and ARIMA achieve identical MAPE of 0.5897% across all training lengths. Among foundation models, Sundial is closest (0.620–0.655%), followed by TimesFM and Chronos (both higher across the range). RMSE and MAE rankings are consistent with MAPE.

Table 5.31: MAPE (%) for raw price forecasting, 2022 Market Downturn with logarithmically-spaced training (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** |
| ARIMA | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** |
| TimesFM | 0.6588 | 0.6657 | 0.6536 | 0.6564 | 0.6362 | 0.6480 | 0.6569 | 0.7005 | 0.6354 | 0.6446 |
| Chronos | 0.6917 | 0.7101 | 0.7019 | 0.6939 | 0.7106 | 0.7106 | 0.7011 | 0.6899 | 0.7123 | 0.7324 |
| Sundial | 0.6455 | 0.6443 | 0.6429 | 0.6452 | 0.6549 | 0.6414 | 0.6247 | 0.6197 | 0.6254 | 0.6261 |

Table 5.32: RMSE for raw price forecasting, 2022 Market Downturn with logarithmically-spaced training (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** |
| ARIMA | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** |
| TimesFM | 2.4135 | 2.4181 | 2.4051 | 2.3658 | 2.3048 | 2.4045 | 2.3939 | 2.5248 | 2.2987 | 2.3370 |
| Chronos | 2.5171 | 2.5633 | 2.5408 | 2.5066 | 2.5532 | 2.5457 | 2.5365 | 2.5159 | 2.5900 | 2.6564 |
| Sundial | 2.3617 | 2.3351 | 2.3470 | 2.3430 | 2.3607 | 2.3298 | 2.2892 | 2.2643 | 2.2831 | 2.2852 |

Table 5.33: MAE for raw price forecasting, 2022 Market Downturn with logarithmically-spaced training (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** |
| ARIMA | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** |
| TimesFM | 1.8230 | 1.8421 | 1.8108 | 1.8187 | 1.7620 | 1.7972 | 1.8182 | 1.9403 | 1.7612 | 1.7836 |
| Chronos | 1.9194 | 1.9701 | 1.9447 | 1.9235 | 1.9673 | 1.9673 | 1.9402 | 1.9079 | 1.9673 | 2.0185 |
| Sundial | 1.7854 | 1.7823 | 1.7793 | 1.7837 | 1.8101 | 1.7745 | 1.7316 | 1.7164 | 1.7320 | 1.7337 |

### 5.3.4 Fractional Change Forecasting During Crises

As in Section 5.1.2, MAPE is excluded from the fractional change analysis.

#### 2008 Global Financial Crisis

Tables 5.34–5.36 present RMSE, MAE, and MDA for fractional change forecasting during the 2008 crisis with log-spaced training. For RMSE, Chronos, TimesFM, and Sundial cluster tightly at 0.0092–0.0093, and ARIMA matches this range from 32 days onward (0.0091–0.0092) with an outlier at 25 days (0.0099). Naïve records 0.0135. For MDA, ARIMA exhibits a sharp bimodal pattern: 54.7% at 25 days, collapsing

to 1.4% from 32–89 days, then recovering to 52.3–52.6% from 116 days onward. Chronos is the most consistent performer across training lengths (48.8–52.6%), while TimesFM ranges from 46.0–54.0% and Sundial from 43.9–52.3%. Naïve records a fixed 44.9%.

Table 5.34: RMSE for fractional change forecasting, 2008 Financial Crisis with logarithmically-spaced training (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 0.0135 | 0.0135 | 0.0135 | 0.0135 | 0.0135 | 0.0135 | 0.0135 | 0.0135 | 0.0135 | 0.0135 |
| ARIMA | 0.0099 | **0.0092** | **0.0092** | **0.0092** | **0.0092** | **0.0092** | **0.0091** | **0.0091** | **0.0091** | **0.0091** |
| TimesFM | 0.0093 | 0.0093 | **0.0092** | **0.0092** | **0.0092** | **0.0092** | 0.0092 | 0.0092 | 0.0092 | 0.0092 |
| Chronos | **0.0092** | **0.0092** | **0.0092** | **0.0092** | **0.0092** | **0.0092** | 0.0092 | 0.0092 | 0.0092 | 0.0092 |
| Sundial | **0.0092** | 0.0093 | 0.0093 | **0.0092** | **0.0092** | **0.0092** | 0.0092 | 0.0093 | 0.0093 | 0.0092 |

Table 5.35: MAE for fractional change forecasting, 2008 Financial Crisis with logarithmically-spaced training (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 0.0093 | 0.0093 | 0.0093 | 0.0093 | 0.0093 | 0.0093 | 0.0093 | 0.0093 | 0.0093 | 0.0093 |
| ARIMA | 0.0066 | **0.0058** | **0.0058** | **0.0058** | **0.0058** | **0.0058** | **0.0058** | **0.0058** | **0.0058** | **0.0058** |
| TimesFM | 0.0060 | 0.0059 | 0.0059 | 0.0059 | 0.0059 | 0.0059 | 0.0059 | 0.0059 | 0.0059 | 0.0059 |
| Chronos | **0.0059** | 0.0059 | **0.0058** | **0.0058** | **0.0058** | **0.0058** | **0.0058** | **0.0058** | **0.0058** | **0.0058** |
| Sundial | **0.0059** | 0.0059 | 0.0059 | 0.0060 | 0.0059 | 0.0060 | 0.0059 | 0.0060 | 0.0059 | 0.0059 |

Table 5.36: Mean Directional Accuracy (%) for fractional change forecasting, 2008 Financial Crisis with logarithmically-spaced training (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 44.9477 | 44.9477 | 44.9477 | 44.9477 | 44.9477 | 44.9477 | 44.9477 | 44.9477 | 44.9477 | 44.9477 |
| ARIMA | **54.7038** | 1.3937 | 1.3937 | 1.3937 | 1.3937 | 1.3937 | **52.2648** | 52.2648 | 52.2648 | **52.6132** |
| TimesFM | 52.9617 | **52.9617** | 49.4774 | 50.1742 | 45.9930 | **54.0070** | **52.2648** | 50.5226 | 49.8258 | 49.4774 |
| Chronos | 50.5226 | 51.5679 | **52.2648** | **52.2648** | **51.9164** | 48.7805 | 51.9164 | **52.6132** | **52.6132** | 51.5679 |
| Sundial | 50.1742 | 51.2195 | 48.0836 | 47.0383 | **51.9164** | 43.9024 | 47.7352 | 47.0383 | 49.8258 | 52.2648 |

### 2020 COVID-19 Crash

Tables 5.37–5.39 present RMSE, MAE, and MDA for fractional change forecasting during the 2020 crash with log-spaced training. For RMSE, ARIMA achieves the lowest error at all training lengths (0.0174–0.0177), followed by Sundial (0.0179–0.0182) and Chronos (0.0181–0.0182), with TimesFM trailing (0.0183–0.0187). Naïve records 0.0287. For MDA, ARIMA leads at most training lengths (51.9–55.0%), while Sundial ranks second at several lengths and peaks at 53.3%. Chronos records the lowest MDA at most training lengths (41.5–48.8%), remaining below the 50% baseline throughout. Naïve records a fixed 48.1%.

Table 5.37: RMSE for fractional change forecasting, 2020 COVID-19 Crash with logarithmically-spaced training (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 0.0287 | 0.0287 | 0.0287 | 0.0287 | 0.0287 | 0.0287 | 0.0287 | 0.0287 | 0.0287 | 0.0287 |
| ARIMA | **0.0176** | **0.0175** | **0.0177** | **0.0176** | **0.0176** | **0.0174** | **0.0174** | **0.0174** | **0.0175** | **0.0175** |
| TimesFM | 0.0186 | 0.0187 | 0.0186 | 0.0185 | 0.0184 | 0.0184 | 0.0183 | 0.0184 | 0.0185 | 0.0184 |
| Chronos | 0.0181 | 0.0181 | 0.0181 | 0.0182 | 0.0182 | 0.0181 | 0.0181 | 0.0181 | 0.0181 | 0.0181 |
| Sundial | 0.0182 | 0.0180 | 0.0182 | 0.0180 | 0.0180 | 0.0179 | 0.0181 | 0.0180 | 0.0182 | 0.0179 |

Table 5.38: MAE for fractional change forecasting, 2020 COVID-19 Crash with logarithmically-spaced training (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 |
| ARIMA | **0.0117** | **0.0117** | **0.0118** | **0.0118** | **0.0117** | **0.0116** | **0.0116** | **0.0115** | **0.0116** | **0.0116** |
| TimesFM | 0.0123 | 0.0123 | 0.0123 | 0.0122 | 0.0122 | 0.0121 | 0.0120 | 0.0121 | 0.0122 | 0.0121 |
| Chronos | 0.0119 | 0.0119 | 0.0119 | 0.0120 | 0.0120 | 0.0119 | 0.0119 | 0.0119 | 0.0119 | 0.0119 |
| Sundial | 0.0121 | 0.0120 | 0.0120 | 0.0119 | 0.0119 | 0.0117 | 0.0119 | 0.0119 | 0.0120 | 0.0118 |

Table 5.39: Mean Directional Accuracy (%) for fractional change forecasting, 2020 COVID-19 Crash with logarithmically-spaced training (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 48.0836 | 48.0836 | 48.0836 | 48.0836 | 48.0836 | 48.0836 | 48.0836 | 48.0836 | 48.0836 | 48.0836 |
| ARIMA | **52.2648** | **53.6585** | **52.9617** | **55.0523** | **54.0070** | 51.9164 | **51.9164** | **52.9617** | **51.9164** | 51.9164 |
| TimesFM | 46.6899 | 48.0836 | 45.6446 | 48.0836 | 49.4774 | 48.4321 | 48.7805 | 47.3868 | 47.7352 | 50.1742 |
| Chronos | 48.7805 | 47.3868 | 45.9930 | 45.2962 | 45.6446 | 43.5540 | 45.6446 | 44.5993 | 45.6446 | 41.4634 |
| Sundial | 48.0836 | 47.3868 | 48.0836 | 49.4774 | 51.5679 | **52.2648** | 50.1742 | 51.9164 | 46.6899 | **53.3101** |

### 2022 Market Downturn

Tables 5.40–5.42 present RMSE, MAE, and MDA for fractional change forecasting during the 2022 downturn with log-spaced training. For RMSE, ARIMA, Chronos, TimesFM, and Sundial cluster tightly at 0.0080–0.0083, while Naïve records 0.0112. For MDA, Naïve achieves the highest directional accuracy at a fixed 54.4% across all training lengths. TimesFM is usually the strongest non-Naïve model (48.1–53.0%), Chronos ranges from 49.1–51.6%, and Sundial is highly variable (45.6–54.4%). ARIMA again records 0.0% MDA across all training lengths.

Table 5.40: RMSE for fractional change forecasting, 2022 Market Downturn with logarithmically-spaced training (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 0.0112 | 0.0112 | 0.0112 | 0.0112 | 0.0112 | 0.0112 | 0.0112 | 0.0112 | 0.0112 | 0.0112 |
| ARIMA | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** |
| TimesFM | 0.0081 | 0.0081 | 0.0081 | 0.0081 | 0.0081 | 0.0081 | **0.0080** | **0.0080** | **0.0080** | **0.0080** |
| Chronos | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** |
| Sundial | **0.0080** | 0.0081 | 0.0083 | **0.0080** | 0.0082 | 0.0081 | 0.0082 | 0.0081 | 0.0081 | 0.0082 |

Table 5.41: MAE for fractional change forecasting, 2022 Market Downturn with logarithmically-spaced training (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 0.0085 | 0.0085 | 0.0085 | 0.0085 | 0.0085 | 0.0085 | 0.0085 | 0.0085 | 0.0085 | 0.0085 |
| ARIMA | **0.0059** | **0.0059** | **0.0059** | **0.0059** | **0.0059** | **0.0059** | **0.0059** | 0.0059 | **0.0059** | **0.0059** |
| TimesFM | 0.0060 | 0.0060 | 0.0060 | **0.0059** | **0.0059** | **0.0059** | **0.0059** | **0.0058** | **0.0059** | **0.0059** |
| Chronos | **0.0059** | **0.0059** | **0.0059** | **0.0059** | **0.0059** | **0.0059** | **0.0059** | 0.0059 | **0.0059** | **0.0059** |
| Sundial | 0.0060 | 0.0061 | 0.0061 | **0.0059** | 0.0061 | 0.0060 | 0.0061 | 0.0060 | 0.0060 | 0.0061 |

Table 5.42: Mean Directional Accuracy (%) for fractional change forecasting, 2022 Market Downturn with logarithmically-spaced training (25–249 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 32d | 41d | 53d | 69d | 89d | 116d | 149d | 193d | 249d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **54.3554** | **54.3554** | **54.3554** | **54.3554** | **54.3554** | **54.3554** | **54.3554** | **54.3554** | **54.3554** | **54.3554** |
| ARIMA | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| TimesFM | 51.2195 | 48.0836 | 50.5226 | 52.6132 | 51.5679 | 49.1289 | 52.9617 | 51.2195 | 52.2648 | 52.2648 |
| Chronos | 49.1289 | 49.1289 | 49.4774 | 50.5226 | 49.8258 | 49.4774 | 51.5679 | 49.8258 | 49.8258 | 50.1742 |
| Sundial | **54.3554** | 50.5226 | 47.0383 | **54.3554** | 47.0383 | 48.0836 | 48.4321 | 48.4321 | 50.5226 | 45.6446 |

# Chapter 6

# Discussion

The results chapter establishes *what* happens across 2,400 experiments. This chapter interprets the *why*. It explains the strength of the Naïve baseline, quantifies model gaps, examines Chronos's long-horizon advantage, assesses the pre-training leakage concern, answers the research questions, and translates the findings into a practical model-selection framework. The chapter closes with the study's limitations and future work.

## 6.1 Difficulty of Outperforming Naïve

Naïve is difficult to beat because the evaluated target is close to a random-walk price process. Under the weak form of the Efficient Market Hypothesis (Fama, 1970; Malkiel, 1973), past prices should not contain exploitable information about future price changes. In martingale terms, the conditional expectation of $y_{t+h}$ given the history up to $t$ is $y_t$ (Samuelson, 1965), making the Naïve predictor $\hat{y}_{t+h|t} = y_t$ the natural benchmark for level forecasts.

The empirical results are consistent with this interpretation. Across short-horizon daily forecasts, hourly forecasts, and the crisis-window level metrics, the models rarely find enough structure to improve materially on the latest observed price. This should not be read as a generic failure of foundation models, but as evidence that univariate MSFT closing prices contain little recoverable short-horizon signal. The relevant question is therefore not simply whether Naïve wins, but whether any model reduces the gap enough, in a specific regime, to justify the added complexity.

## 6.2 Quantifying the Gap Between Models

Table 6.1 reports the absolute gap from the best model in each headline regime, computed as model value minus the row's best value. For MAPE, gaps are percentage-point differences for RMSE, gaps are in the original price units. MDA is discussed separately because it measures sign rather than level accuracy.

The table shows a horizon-dependent pattern. At one and five days, Naïve and ARIMA remain the benchmark. At 21 days, TimesFM leads, but the gap is modest: Naïve and ARIMA trail by 0.1937 and 0.1752 percentage points on MAPE, and by 0.8514 and 0.5099 on RMSE. At 63 days, the advantage shifts to Chronos and becomes larger, especially on RMSE. This is the only daily horizon where the gap is wide enough to suggest a meaningful foundation-model advantage rather than a small ranking change.

The crisis rows are therefore interpreted through the 2022 downturn first. Because this is the least leakage-prone disruption window, its traditional-baseline lead is treated as the primary evidence for crisis

Table 6.1: Absolute gap from the best model on MAPE and RMSE for each headline regime. MAPE gaps are percentage-point differences; RMSE gaps are in price units. Bold zeros mark the leader or a tie within rounding. Training-length cells in parentheses indicate the configuration used.

| Regime | Metric | Naïve | ARIMA | TimesFM | Chronos | Sundial |
|---|---|---|---|---|---|---|
| Daily, 1d (1y train) | MAPE | **0.0000** | **0.0000** | +0.0495 | +0.0801 | +0.0631 |
| | RMSE | **0.0000** | **0.0000** | +0.0969 | +0.2478 | +0.2664 |
| Daily, 5d (1y train) | MAPE | **0.0000** | **0.0000** | +0.0280 | +0.1289 | +0.1962 |
| | RMSE | **0.0000** | **0.0000** | +0.1290 | +0.4292 | +0.7831 |
| Daily, 21d (6y train) | MAPE | +0.1937 | +0.1752 | **0.0000** | +0.4659 | +0.9511 |
| | RMSE | +0.8514 | +0.5099 | **0.0000** | +1.1461 | +3.0148 |
| Daily, 63d (1y train) | MAPE | +1.4979 | +1.4979 | +1.6961 | **0.0000** | +1.7687 |
| | RMSE | +5.5325 | +5.5325 | +7.6469 | **0.0000** | +10.3309 |
| Hourly, 1h (1y train) | MAPE | **0.0000** | **0.0000** | +0.0498 | +0.1294 | +0.0204 |
| | RMSE | **0.0000** | **0.0000** | +0.1505 | +0.4458 | +0.0602 |
| 2008 crisis (250d) | MAPE | **0.0000** | +0.0012 | +0.0642 | +0.1465 | +0.0152 |
| | RMSE | +0.0014 | **0.0000** | +0.0150 | +0.0318 | +0.0007 |
| 2020 crisis (150d) | MAPE | +0.0260 | +0.0079 | +0.0454 | +0.1166 | **0.0000** |
| | RMSE | +0.1323 | +0.0514 | +0.0931 | +0.2155 | **0.0000** |
| 2022 crisis (125d) | MAPE | **0.0000** | **0.0000** | +0.0947 | +0.1017 | +0.0311 |
| | RMSE | **0.0000** | **0.0000** | +0.2885 | +0.3132 | +0.0866 |

forecasting. Naïve and ARIMA tie for the lead across both MAPE and RMSE, and the full 2022 raw-price tables strengthen this point: the Naïve–ARIMA tie holds at every logarithmically-spaced training length from 25 to 249 days, so the baseline advantage is not an artifact of a single training-window choice. The 2008 and 2020 rows are used as secondary stress checks rather than equal evidence. The 2008 row is consistent with the 2022 conclusion because Naïve and ARIMA are also at the front. The 2020 row is not enough to reverse the conclusion, even though Sundial leads, because that result is exposed to the leakage concern discussed in Section 6.5.

The practical implication is that downstream use cases should not treat a rank difference as sufficient evidence for model replacement. In most regimes, the absolute gaps are small enough that transaction costs, latency, calibration needs, or interpretability requirements could dominate the statistical difference. The main exception is the 63-day daily horizon, where Chronos reduces both typical and tail-sensitive error by a visibly larger margin. For short-horizon price forecasts, crisis monitoring, or low-latency deployment, the table supports a simple baseline unless the downstream task specifically benefits from probabilistic forecasts or long-horizon error reduction.

## 6.3 Chronos Long-Horizon Advantage

Chronos is the only model in the study that clearly improves on both traditional baselines and the other foundation models in one regime: the 63-day horizon on daily data. It records the lowest MAPE in the limited-data setting (6.25% at 25 days vs. Naïve's 7.17%) and the lowest RMSE at nearly every year-increment configuration. The result is narrow, however. Chronos does not lead at short horizons, on hourly data, or in the cleanest crisis test.

The most plausible explanation is architectural fit to the long-horizon task. Chronos generates the forecast horizon directly, whereas TimesFM rolls out autoregressively across patches, Sundial samples trajectories through its flow-matching procedure, and the traditional baselines propagate the latest state forward. At short horizons these differences have limited effect, but at 63 days error accumulation becomes more important. Direct multi-step decoding can therefore reduce long-horizon error even without extracting a stronger financial signal. This interpretation is consistent with the TimesFM authors' own observation that direct horizon prediction can outperform autoregressive decoding on long-horizon benchmarks (Das

et al., 2024).

## 6.4 Additional Patterns in Results

### 6.4.1 Empirical ARIMA order distribution

Across the rolling-forecast folds, `auto_arima` almost always selects $(0, 1, 0)$ on price levels and $(0, 0, 0)$ on returns (Figure 6.1), with higher orders appearing only in a thin tail. This is itself evidence of a near-random-walk price process: the information criterion finds little AR or MA structure worth retaining. On prices, the modal $(0, 1, 0)$ specification is a random walk whose one-step forecast matches Naïve, explaining the repeated ARIMA–Naïve ties. On returns, the modal $(0, 0, 0)$ specification forecasts the unconditional mean.

The 0.0% MDA records require a separate interpretation. In several return configurations, ARIMA's unconditional-mean forecast is numerically zero across the rolling test period. Under the fractional-change MDA convention defined in Section 4.4.3, these zero forecasts are counted as directionally correct only when the realised return is exactly zero. The 0.0% MDA records in RQ1 (1–3y), RQ2 hourly returns, and RQ3 2022 therefore indicate a zero-return forecast under this scoring convention, not a sequence of consistently wrong positive or negative calls. ARIMA is therefore a poor candidate when MDA is the target metric.

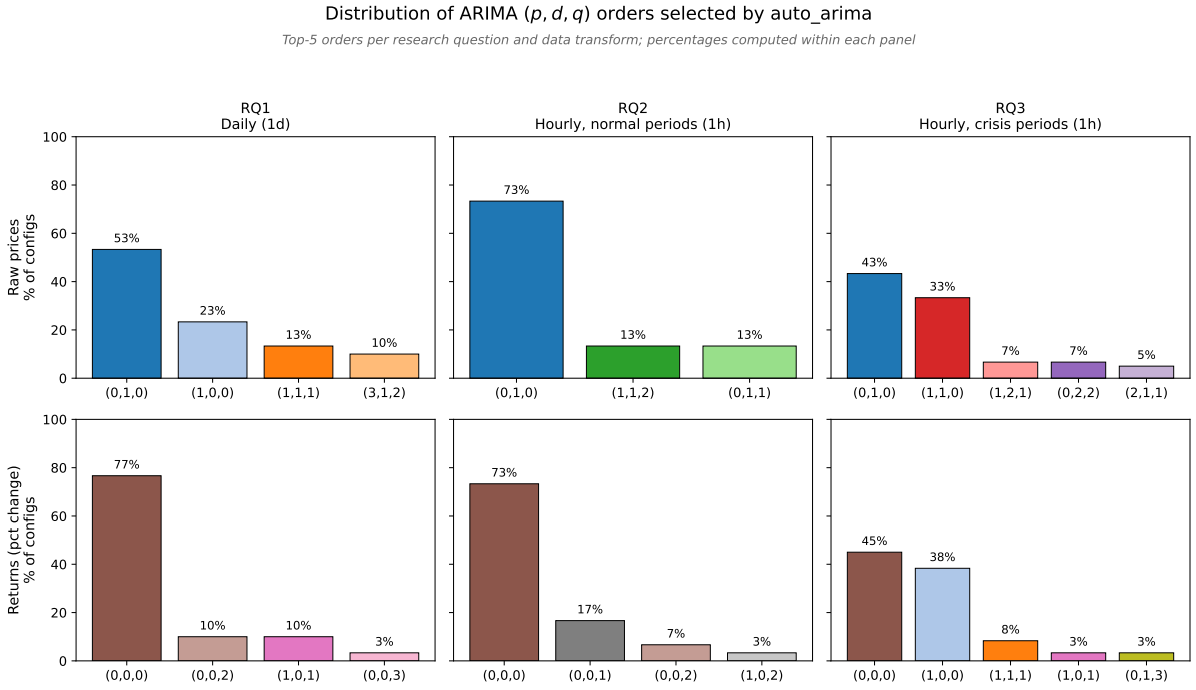


Figure 6.1: Empirical distribution of $(p, d, q)$ triples selected by `auto_arima` across every rolling-forecast fold, split by sampling frequency (daily vs. hourly) and by target representation (raw price vs. fractional return).

### 6.4.2 Near-zero forecasts on hourly returns

On hourly fractional-change returns, every non-Naïve model converges to an RMSE of approximately 0.0060, while Naïve records 0.0084 and the same separation appears in MAE. This pattern is expected if returns are close to independent zero-mean draws.

Figure 6.2 compares the daily and hourly fractional-change series. The hourly series is tightly concentrated around zero, with an average fractional change of $9.13 \times 10^{-5}$ and occasional crisis-period spikes rather than sustained directional structure. This distribution helps explain why hourly fractional-change RMSE and MAE cluster so tightly across models: once models produce forecasts close to zero, their squared and absolute errors are dominated by the same small realized return variance. In this setting, different forecast paths can therefore yield nearly identical aggregate error values.

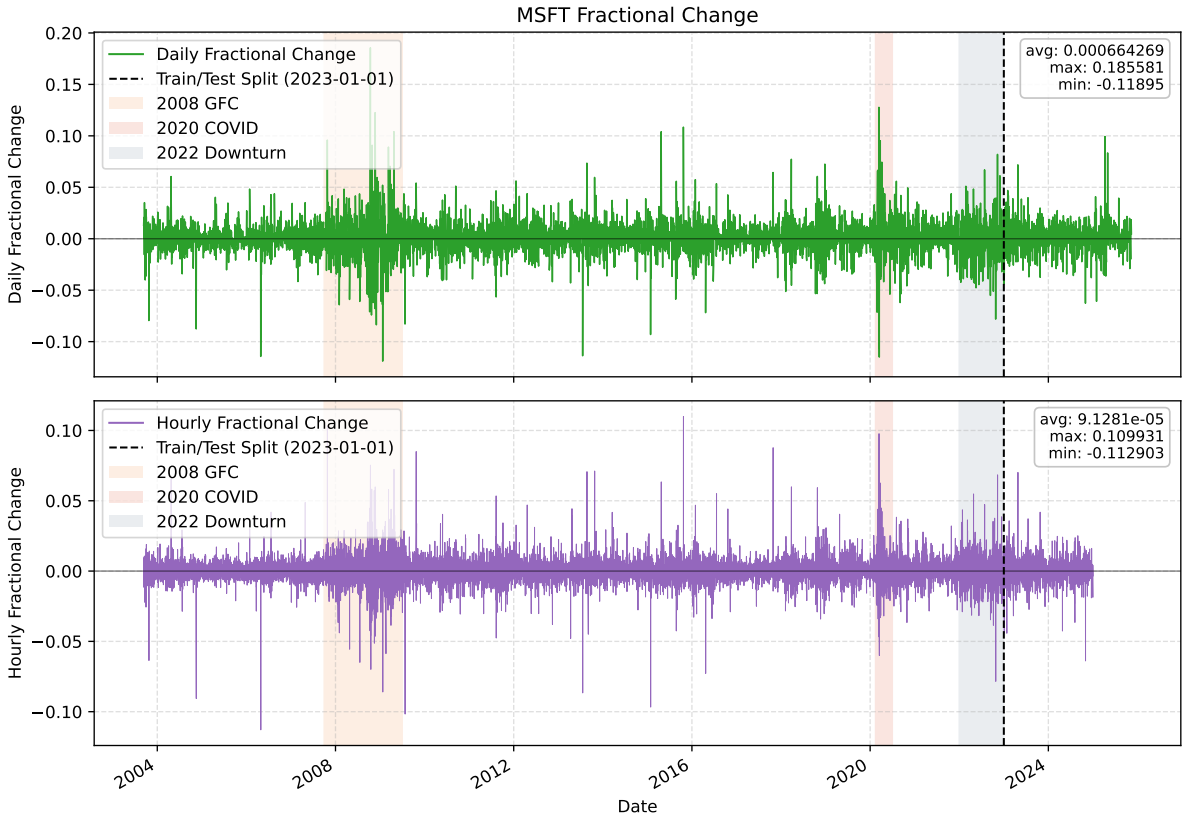


Figure 6.2: MSFT daily and hourly fractional-change series. The dashed vertical line marks the January 1, 2023 train-test split, and shaded regions indicate the 2008 Global Financial Crisis, the 2020 COVID-19 crash, and the 2022 market downturn. The hourly series is more tightly concentrated around zero than the daily series, which helps explain the narrow spread of hourly fractional-change errors.

The RMSE-optimal forecast is therefore close to the unconditional mean, yielding error $\sigma$, while Naïve carries forward last-period noise and incurs approximately $\sigma\sqrt{2}$. The empirical ratio $0.0084/0.0060 \approx 1.41$ matches that relationship. The result is therefore not evidence that the non-Naïve models extract directional structure; rather, they converge on the same near-zero return forecast under squared loss. It also explains why conclusions depend on the target representation: Naïve is structurally strong on price levels, but structurally weak on returns.

The equality of several hourly return metrics should therefore be read carefully. In the year-increment configuration, Naïve produces exactly the same RMSE, MAE, MSE, and MDA for 1–10 years because the one-step rolling rule depends only on the last observed return, not on older training history. TimesFM and Chronos produce exactly the same metrics from 2 years onward, indicating that the extra pre-split hourly history lies outside the effective context used at each rolling forecast origin. By contrast, ARIMA and Sundial mostly vary in the raw configuration summaries, but the values remain close enough that rounding to four decimals often collapses them to the same displayed number. The repeated table entries are thus a mixture of true forecast invariance and presentation-level rounding, not an independent sign of equal model skill.

### 6.4.3 Naïve Directional Accuracy at 63 Days

At the 63-day horizon, Naïve attains the highest directional accuracy in the study at 58.8% MDA. At one- and five-day horizons, directional accuracy for all models remains within a narrow band close to 50–58%, with no consistent separation between methods.

This pattern is best understood in terms of horizon-dependent structure in return signs. The Naïve predictor in this setup carries forward the most recent observed return, so its directional forecast effectively tests whether return signs tend to persist over the chosen horizon. At short horizons, sign changes are frequent and largely dominated by noise, which limits all models to near-random performance.

As the horizon increases, returns become smoother due to aggregation, and short-term reversals cancel out more frequently. This increases the persistence of the resulting sign sequence, making it easier for a simple lag-based rule to achieve above-random directional accuracy. The improvement in Naïve performance at 63 days therefore reflects increased sign persistence in longer-horizon returns rather than any form of learned conditional structure.

Consequently, elevated directional accuracy at longer horizons should be interpreted as an effect of temporal aggregation on the return process, not as evidence of improved predictive skill.

### 6.4.4 Training Length and Forecast Quality

Across all three research questions, training-set length and forecast quality are not monotonically related, and the non-monotonicity reproduces across MAPE, MAE, and RMSE. The clearest cases are the following. On hourly raw-price (Table 5.13), Sundial degrades by roughly a factor of four on MAPE while TimesFM and Chronos stay flat. On daily one-day (Table 5.1), both TimesFM and Chronos worsen as history grows. At the 63-day horizon, Chronos's best MAPE in the grid is achieved with one year of training. On the 2022 downturn (Table A.25), Chronos deteriorates as the window lengthens while Naïve and ARIMA are flat. Three mechanisms cover the pattern. Zero-shot foundation models have receptive fields that saturate, after which extra history is redundant or harmful. Traditional baselines on prices with little predictable conditional structure gain shrinking standard errors but no new signal. Across regime shifts, a longer pre-event window dilutes the recent signal. A longer record is therefore not a free improvement, and training-window length is itself a regime-specific design choice.

### 6.4.5 Foundation Model Stability in RQ2

On hourly raw-price MAPE, TimesFM and Chronos remain essentially flat across the entire year-increment ladder, at approximately 0.45% and 0.53% respectively, and the same flatness appears in their MAE and RMSE traces, while ARIMA varies as its parameters re-fit each fold and Sundial deteriorates on every level metric with longer inputs. The stability of TimesFM and Chronos is informative on two distinct levels. Operationally, it is the dual face of the cold-start robustness identified earlier: where ARIMA's quality depends materially on how much pre-crisis history happens to be available, the two stable foundation models offer near-constant performance once a moderate context has been provided, which simplifies deployment in regimes where the available history is inconsistent across assets. Theoretically, it indicates that additional historical context at the hourly frequency does not unlock additional conditional skill for these architectures: their effective receptive field saturates well within the shortest training window evaluated, and the longer windows convey no exploitable structure beyond what the shorter ones already supply.

### 6.4.6 Sundial's bimodal stability profile

Sundial is the strongest foundation model on hourly limited-data (0.42–0.43% MAPE) but markedly unstable on hourly year-increment training, with MAPE rising to 1.64% at 10 years and RMSE nearly tripling. This is an aggressive degradation in the exact setting where a long-context foundation model should become more useful.

One explanation is that the models do not use longer histories in the same way. TimesFM and Chronos both cap their effective input length at 2048 time points, so additional hourly history beyond that limit is truncated or otherwise outside the model's usable context (Das et al., 2024; Ansari, Stella, et al., 2024). Sundial, by contrast, is specified with a longer 2880-point context (Yong Liu, Qin, et al., 2025), and its public generation wrapper does not appear to enforce hard truncation to that limit. Inputs are padded to a multiple of the patch length and converted into non-overlapping patches of 16 observations, so 2880 observations correspond to 180 model tokens (THUML, 2025a).

Sundial's deterioration is therefore best read as a context-selection problem. In a univariate price series that is close to a random walk, older hourly observations can introduce stale regime information without adding usable signal. Its generation code also normalizes using the mean and standard deviation of the full supplied input, so a much longer history can affect the scale of the recent observations used for forecasting (THUML, 2025b). A model whose accuracy worsens sharply when more historical data is supplied cannot be treated as a reliable long-context forecaster without context selection, truncation, or fine-tuning.

### 6.4.7 Foundation Models as ARIMA Fallback

In the daily limited-data regime ARIMA fails to converge at 25, 41, 53, and 116 days of training, producing MAPE, MAE, and RMSE several multiples higher than any other model. The foundation models remain stable across the same range on all three level metrics. Even when a foundation model does not outperform a working ARIMA, its failure mode is gradual whereas ARIMA's is abrupt, and the cold-start use case (an asset with limited price history) constitutes the strongest argument in favour of foundation models that this study supports.

## 6.5 Pre-Training Data Leakage

A natural concern in the crisis experiments is that the foundation models may have encountered parts of the 2008 and 2020 test periods during pre-training. The traditional baselines were fitted only on the strictly pre-crisis windows supplied at inference time, so any foundation-model exposure to MSFT in those periods would make the comparison asymmetric and would weaken claims of crisis resilience.

The published documentation does not allow this possibility to be confirmed. Chronos was trained on public multi-domain time-series datasets, including some financial data, but the reported sources are benchmark collections rather than ticker-level series (Ansari, Stella, et al., 2024). TimesFM was trained primarily on Google Trends and Wikipedia pageview data, with no financial price data listed among its main sources (Das et al., 2024). Sundial was pre-trained on TimeBench, a large mixture of real and synthetic time series, but its constituent datasets are not enumerated at ticker level (Yong Liu, Qin, et al., 2025). The defensible position is therefore not that leakage occurred, but that it cannot be ruled out for all models and crisis windows.

For this reason, the 2022 downturn is the cleanest crisis test in the study: it post-dates the published pre-training cutoffs of all three foundation models. In that window, Naïve and ARIMA tie at 0.5897% MAPE on raw price, match on MAE and RMSE rankings, and outperform every foundation model on each level metric; Naïve also leads on fractional-change MDA at 54.4%. The evidence therefore does not support a foundation-model crisis-resilience advantage. Results from 2008 and 2020 should be treated as

suggestive only, because any favourable foundation-model performance in those windows is inseparable from the unresolved pre-training exposure question.

## 6.6 Answers to the Research Questions

### 6.6.1 RQ1: Forecast horizon length effects

*How does forecast horizon length affect the relative performance of time-series foundation models compared to traditional approaches for stock price forecasting?*

Horizon length is the single factor that most clearly differentiates foundation models from the traditional baselines. At one and five days ahead, no model in the study achieves a meaningful advantage over Naïve on any of the four metrics, with MAPE, MAE, and RMSE rankings agreeing and MDA scattered within a 48–56% band. At twenty-one days, TimesFM produces a marginal lead on MAPE (a relative reduction of approximately 6% at six years of training) that is partially mirrored on MAE. At sixty-three days, Chronos achieves the largest non-Naïve advantage observed in the study, with relative MAPE reductions of 13–21% over Naïve in the limited-data setting and the lowest RMSE at nine of ten year-increment training lengths, the RMSE lead being wider than the MAPE lead because Chronos's worst forecasts are systematically less extreme than the other models'. The relative ordering of model families therefore changes with horizon: traditional approaches and Naïve are essentially equivalent to foundation models at short horizons across every metric, while foundation models, and Chronos in particular, gain a regime-specific advantage at long horizons.

### 6.6.2 RQ2: Temporal frequency effects

*How does the temporal frequency of input data affect the relative performance of foundation models compared to traditional approaches for stock price forecasting?*

Moving from daily to hourly data does not produce a frequency-driven advantage for foundation models on raw-price level metrics. On the contrary, Naïve and ARIMA tie at 0.3994% MAPE on hourly one-step forecasts, with matching MAE and RMSE, and no foundation model closes the gap on any of the three. The main contribution of the foundation models at hourly frequency is therefore not higher level-metric accuracy but greater stability across training-window lengths. TimesFM and Chronos exhibit near-constant hourly raw-price MAPE, MAE, and RMSE across the full year-increment ladder (MAPE approximately 0.45% and 0.53% respectively), whereas ARIMA's quality varies with the available history and Sundial deteriorates sharply at long inputs on every level metric. The answer to RQ2 is therefore that higher temporal frequency preserves the competitiveness of traditional baselines on raw-price forecasting, while making stability across diverse training windows the clearest practical advantage of the stronger foundation models. Sundial does not share that advantage in this study, and its long-context degradation makes its use as a default long-history foundation model questionable without context selection or adaptation.

### 6.6.3 RQ3: Market disruption resilience

*How do foundation models perform relative to traditional approaches during major market disruptions?*

This study finds no evidence of a foundation-model crisis-resilience advantage on the cleanest available test. In the 2022 downturn, Naïve and ARIMA tie at 0.5897% MAPE and outperform every foundation model on raw price across MAPE, MAE, and RMSE, while Naïve also leads on fractional-change MDA at 54.4%. The earlier 2008 and 2020 windows do not overturn this conclusion, because any favourable foundation-model result in those periods must be read alongside the pre-training leakage concern discussed

in Section 6.5. The practical conclusion is therefore that the foundation models' crisis-period value lies more in cold-start stability than in superior disruption accuracy.

## 6.7 A Practical Forecasting Framework

The most directly actionable contribution of the thesis is a scenario-to-model mapping. Table 6.2 summarises a recommended choice for each combination of horizon, data availability, frequency, and task type encountered in the study, with the rationale condensed into a one-line column.

Table 6.2: Practical model-selection framework derived from the experimental record on MSFT. Recommendations are descriptive of this study. Transferability to other assets and markets is the subject of Section 6.8.

| **Scenario** | **Recommended** | **Why** |
|---|---|---|
| Short-horizon (1–5d) point forecast, liquid large-cap | Naïve | Random-walk benchmark. Foundation models add cost without edge |
| Short-horizon, need uncertainty bands | ARIMA *or* Chronos | ARIMA cheaper. Chronos when data is limited or unstable |
| Medium horizon (21d) | TimesFM | Marginally lowest MAPE at 21d in this study |
| Long horizon (63d) | Chronos | Largest long-horizon advantage, especially on RMSE |
| Limited / cold-start data (<1 year) | Foundation models | Sundial is strongest among FMs with limited data. FMs remain stable where ARIMA fails to converge |
| Hourly raw price, multi-year history | Naïve or ARIMA | Foundation models do not beat baseline |
| Hourly frequency, returns / directional task | Sundial | Only model consistently >50% MDA on hourly returns |
| Crisis / regime-shift forecasting | Naïve (FM as ensemble member) | Cleanest test (2022) shows no FM resilience edge |

One qualification frames the table. Every recommendation is validated on MSFT only, and the thresholds implied by the rows are descriptive of this study rather than prescriptive for the wider equity universe. The crisis recommendation is conservative for the leakage reasons discussed in Section 6.5.

The most notable pattern in the framework is that no single model is preferred across all configurations. The choice of forecaster depends on the horizon, the data-availability regime, the data frequency, and whether the downstream task is point-prediction, distributional, or directional. This itself constitutes the headline contribution: framing the comparison between foundation models and traditional approaches as a single, undifferentiated contest is mis-posed. The appropriate question is *when* each tool is preferable, and the table above offers one answer for the MSFT case.

## 6.8 Limitations

**Single stock.** All experiments use only MSFT price. This is a deliberate depth-over-breadth design choice (justified in Chapter 1, Section 2.6) that allows a fully crossed factorial study within a feasible compute budget, but the conclusions are properly stated as conclusions about MSFT. Smaller-cap, less liquid, or recently listed names may show stronger departures from weak-form efficiency and may have weaker pre-training exposure, both of which could shift relative model performance.

**Single feature.** The models receive only the closing-price series. Volume, order-book imbalance, options-implied volatility, sentiment, and macro indicators are excluded by design. The achievable accuracy in this thesis is therefore capped at the level supported by univariate price information alone.

**Zero-shot foundation-model inference.** No fine-tuning was performed for any foundation model. Rahimikia, Ni, and W. Wang (2025) report that fine-tuning closes some but not all of the gap on financial data. A fine-tuned foundation model evaluated against the same baselines is not addressed by this thesis and is identified as future work.

## 6.9 Future Work

**Multiple-stock evaluation.** Repeating the same experimental setup on a broader set of stocks, covering different company sizes, trading activity levels, sectors, and listing histories, would test whether the rankings reported here extend beyond MSFT. Smaller and more recently listed companies are especially informative because they are more likely to contain exploitable price structure, and foundation models are less likely to have seen much data on them during pre-training. As a result, both the Naïve baseline and the leakage concern discussed in Section 6.5 should play a smaller role.

**Multivariate inputs.** Adding variables such as volume, order-book imbalance, options-implied volatility, sentiment, macro indicators, and other relevant features would raise the ceiling observed in the univariate setting. Of the five models studied, TimesFM and Sundial already support additional inputs, and Chronos-2 (Ansari, Shchur, et al., 2025) extends the original Chronos to handle multiple variables, so the comparison can be carried out without changing the model set. The key question is whether the gap to the Naïve benchmark, which only closes at longer horizons in the univariate case, closes earlier once these additional inputs are included.

**Fine-tuning the foundation models.** A natural follow-up is to fine-tune Chronos, TimesFM, and Sundial on a held-out span of MSFT, or on the cross-asset selection described above, and then re-run the same protocol against the same baselines. Rahimikia, Ni, and W. Wang (2025) report partial gap closure under fine-tuning on financial data, the open question is whether this improvement is consistent across horizons and regimes or, as in the zero-shot results in this thesis, varies by regime.

# Chapter 7

# Conclusion

This thesis examined whether recent time-series foundation models provide a practical advantage over traditional statistical approaches for stock price forecasting. Using MSFT as a focused case study, it compared Naïve and ARIMA baselines against Chronos, TimesFM, and Sundial across 2,400 rolling-forecast experiments covering multiple horizons, temporal frequencies, data representations, training-window lengths, and market disruption periods.

The overall conclusion is that foundation models do not uniformly outperform simple statistical baselines in this financial setting. For short-horizon raw-price forecasting, Naïve remains extremely difficult to beat, and ARIMA often behaves similarly. In these regimes, the additional complexity of zero-shot foundation models is generally not justified by the observed point-forecast accuracy.

The strongest evidence in favor of foundation models appears at longer horizons and under limited-data conditions. TimesFM achieves a modest advantage at 21 days, while Chronos produces the clearest gain at 63 days, especially on RMSE. Foundation models also show value as stable fallback models when ARIMA fails to converge in very short training windows.

Moving from daily to hourly data does not create a foundation-model advantage on raw-price forecasts. Naïve and ARIMA remain strongest on hourly one-step level prediction, although TimesFM and Chronos are relatively stable across training-window lengths. The benefit of foundation models at higher frequency is therefore better described as operational stability than as superior level accuracy.

The market-disruption conclusion gives the 2022 downturn more weight than the 2008 and 2020 windows. The 2022 window is the closest test to a genuinely unseen market disruption because it post-dates the published pre-training cutoffs of the foundation models. In that window, Naïve and ARIMA outperform all foundation models on raw-price level metrics, so the primary crisis-period evidence does not support a foundation-model crisis-resilience advantage. The 2008 and 2020 findings are used as secondary context rather than equal evidence: 2008 points in the same direction as 2022, while the favourable 2020 Sundial result is treated as lower-confidence because the foundation models may have encountered that historical period during pre-training.

## Final Remarks

The findings of this thesis suggest a measured view of time-series foundation models in stock forecasting. Their generalization capabilities are real, but they do not automatically translate into superior financial forecasts. In this setting, simple baselines remain highly competitive, while foundation models are most promising for longer horizons, limited-data situations, probabilistic outputs, and cases where traditional model fitting becomes unstable.

The practical implication is therefore not that foundation models replace traditional methods, but that they expand the forecasting toolkit. For financial forecasting, the central question is not whether old or new methods are universally better, but when each method is appropriate.

# Disclaimer

During the preparation of this thesis, AI-assisted coding and writing tools were used to support selected auxiliary tasks. Specifically, these tools supported proof-reading and language polishing of drafted sections, the automated generation of results tables from experiment configuration files, and the drafting of utility scripts used throughout the experimental pipeline.

All research contributions, including the formulation of research questions, the experimental design, the implementation of the forecasting models, the analysis of results, and the interpretation of findings, are the work of the author. Any AI-assisted output was critically reviewed, verified, and edited by the author, who takes full responsibility for the final content of this thesis.

# Bibliography


Adebiyi, Ayodele A., Aderemi O. Adewumi, and Charles K. Ayo (2014). "Comparison of ARIMA and artificial neural networks models for stock price prediction". In: *Journal of Applied Mathematics* 2014, p. 614342.

Al-Alawi, Adel Ismail and Yusuf Ahmed Alaali (Jan. 2023). "Stock Market Prediction using Machine Learning Techniques: Literature Review Analysis". In: *2023 International Conference On Cyber Management And Engineering (CyMaEn)*, pp. 153–157. DOI: 10.1109/CyMaEn57228.2023.10050933. URL: https://ieeexplore.ieee.org/document/10050933/?arnumber=10050933 (visited on 10/28/2024).

Ansari, Abdul Fatir, Oleksandr Shchur, et al. (2025). *Chronos-2: From Univariate to Universal Forecasting*. arXiv: 2510.15821 [cs.LG]. URL: https://arxiv.org/abs/2510.15821.

Ansari, Abdul Fatir, Lorenzo Stella, et al. (Nov. 2024). *Chronos: Learning the Language of Time Series*. en. arXiv:2403.07815 [cs]. DOI: 10.48550/arXiv.2403.07815. URL: http://arxiv.org/abs/2403.07815 (visited on 03/07/2025).

Ariyo, Adebiyi A., Adewumi O. Adewumi, and Charles K. Ayo (Mar. 2014). "Stock Price Prediction Using the ARIMA Model". In: *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*. Cambridge, United Kingdom: IEEE, pp. 106–112. ISBN: 978-1-4799-4922-9 978-1-4799-4923-6. DOI: 10.1109/UKSim.2014.67. URL: http://ieeexplore.ieee.org/document/7046047/ (visited on 11/05/2024).

AWS Machine Learning Blog (2025). *Fast and Accurate Zero-Shot Forecasting with Chronos Bolt and AutoGluon*. URL: https://aws.amazon.com/blogs/machine-learning/fast-and-accurate-zero-shot-forecasting-with-chronos-bolt-and-autogluon/ (visited on 03/11/2026).

Bollerslev, Tim (Apr. 1986). "Generalized autoregressive conditional heteroskedasticity". In: *Journal of Econometrics* 31.3, pp. 307–327. DOI: None. URL: https://ideas.repec.org/a/eee/econom/v31y1986i3p307-327.html.

Box, George E. P. and Gwilym M. Jenkins (1976). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Buczyński, Mateusz et al. (2023). "Financial Time Series Models—Comprehensive Review of Deep Learning Approaches and Practical Recommendations". en. In: *Engineering Proceedings* 39.1. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, p. 79. ISSN: 2673-4591. DOI: 10.3390/engproc2023039079. URL: https://www.mdpi.com/2673-4591/39/1/79 (visited on 10/27/2024).

Cho, Kyunghyun et al. (Sept. 2014). *Learning Phrase Representations using RNN Encoder- Decoder for Statistical Machine Translation*. arXiv:1406.1078. DOI: 10.48550/arXiv.1406.1078. URL: http://arxiv.org/abs/1406.1078 (visited on 11/18/2024).

Das, Abhimanyu et al. (Apr. 2024). *A decoder-only foundation model for time-series forecasting*. arXiv:2310.10688. URL: http://arxiv.org/abs/2310.10688 (visited on 10/25/2024).

Ekambaram, Vijay et al. (June 2024). *Tiny Time Mixers (TTMs): Fast Pre-trained Models for Enhanced Zero/Few-Shot Forecasting of Multivariate Time Series*. arXiv:2401.03955. URL: http://arxiv.org/abs/2401.03955 (visited on 10/25/2024).

Engle, Robert F. (1982). "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation". In: *Econometrica* 50.4, pp. 987–1007.

Fama, Eugene F. (1970). "Efficient capital markets: A review of theory and empirical work". In: *The Journal of Finance* 25.2, pp. 383–417.

Fama, Eugene F. and Kenneth R. French (1996). "Multifactor Explanations of Asset Pricing Anomalies". en. In: *The Journal of Finance* 51.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.1996.tb05202.x, pp. 55–84. ISSN: 1540-6261. DOI: 10.1111/j.1540-6261.1996.tb05202.x. URL:

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1996.tb05202.x (visited on 11/17/2024).

Fattah, Jamal et al. (Jan. 2018). "Forecasting of demand using ARIMA model". en. In: *International Journal of Engineering Business Management* 10, p. 1847979018808673. ISSN: 1847-9790, 1847-9790. DOI: 10.1177/1847979018808673. URL: https://journals.sagepub.com/doi/10.1177/1847979018808673 (visited on 11/05/2024).

Fischer, Thomas and Christopher Krauss (2018). "Deep learning with long short-term memory networks for financial market predictions". In: *European Journal of Operational Research* 270.2, pp. 654–669.

Garza, Azul, Cristian Challu, and Max Mergenthaler-Canseco (May 2024). *TimeGPT-1*. arXiv:2310.03589. URL: http://arxiv.org/abs/2310.03589 (visited on 10/25/2024).

Godahewa, Rakshitha et al. (2021). "Monash Time Series Forecasting Archive". In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Goswami, Mononito et al. (Oct. 2024). *MOMENT: A Family of Open Time-series Foundation Models*. arXiv:2402.03885. URL: http://arxiv.org/abs/2402.03885 (visited on 10/25/2024).

Gu, Shihao, Bryan Kelly, and Dacheng Xiu (2020). "Empirical asset pricing via machine learning". In: *The Review of Financial Studies* 33.5, pp. 2223–2273.

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). "Long Short-Term Memory". en. In: *Neural Computation* 9.8, pp. 1735–1780. ISSN: 0899-7667, 1530-888X. DOI: 10.1162/neco.1997.9.8.1735. URL: https://direct.mit.edu/neco/article/9/8/1735-1780/6109 (visited on 11/18/2024).

Hyndman, Rob J. and Anne B. Koehler (2006). "Another look at measures of forecast accuracy". In: *International Journal of Forecasting* 22.4, pp. 679–688.

Liang, Yuxuan et al. (June 2024). *Foundation Models for Time Series Analysis: A Tutorial and Survey*. arXiv:2403.14735. URL: http://arxiv.org/abs/2403.14735 (visited on 10/25/2024).

Lim, Bryan and Stefan Zohren (2021). "Time-series forecasting with deep learning: a survey". In: *Philosophical Transactions of the Royal Society A* 379.2194, p. 20200209.

Lipman, Yaron et al. (2023). "Flow Matching for Generative Modeling". In: *International Conference on Learning Representations*.

Liu, Xiangfei et al. (2024). "TSFM-Bench: A Comprehensive and Unified Benchmark of Foundation Models for Time Series Forecasting". In: *arXiv preprint arXiv:2410.11802*.

Liu, Xu et al. (Oct. 2024). *Moirai-MoE: Empowering Time Series Foundation Models with Sparse Mixture of Experts*. arXiv:2410.10469. URL: http://arxiv.org/abs/2410.10469 (visited on 10/25/2024).

Liu, Yong, Guo Qin, et al. (2025). *Sundial: A Family of Highly Capable Time Series Foundation Models*. arXiv: 2502.00816 [cs.LG]. URL: https://arxiv.org/abs/2502.00816.

Liu, Yong, Haixu Wu, et al. (2022). *Non-stationary Transformers: Exploring the Stationarity in Time Series Forecasting*. arXiv: 2205.14415 [cs.LG]. URL: https://arxiv.org/abs/2205.14415.

Makridakis, Spyros, Evangelos Spiliotis, and Vassilios Assimakopoulos (2020). "The M4 Competition: 100,000 time series and 61 forecasting methods". In: *International Journal of Forecasting* 36.1, pp. 54–74.

— (2022). "The M5 accuracy competition: Results, findings, and conclusions". In: *International Journal of Forecasting* 38.4, pp. 1346–1364.

Malkiel, Burton. G. (1973). *A Random Walk Down Wall Street*. Norton, New York.

Nie, Yuqi et al. (2023). "A Time Series is Worth 64 Words: Long-term Forecasting with Transformers". In: *International Conference on Learning Representations (ICLR)*.

Phuoc, Tran et al. (Mar. 2024). "Applying machine learning algorithms to predict the stock price trend in the stock market – The case of Vietnam". en. In: *Humanities and Social Sciences Communications* 11.1, p. 393. ISSN: 2662-9992. DOI: 10.1057/s41599-024-02807-x. URL: https://www.nature.com/articles/s41599-024-02807-x (visited on 11/05/2024).

Radford, Alec et al. (2018). "Improving language understanding by generative pre-training". In.

Raffel, Colin et al. (2020). "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *Journal of Machine Learning Research* 21.140, pp. 1–67.

Rahimikia, Eghbal, Hao Ni, and Weiguan Wang (2025). "Re(Visiting) Time Series Foundation Models in Finance". In: *arXiv preprint arXiv:2511.18578*.

Rasul, Kashif et al. (Feb. 2024). *Lag-Llama: Towards Foundation Models for Probabilistic Time Series Forecasting*. arXiv:2310.08278. URL: http://arxiv.org/abs/2310.08278 (visited on 10/31/2024).

Samuelson, Paul A. (1965). "Proof that properly anticipated prices fluctuate randomly". In: *Industrial Management Review* 6.2, pp. 41–49.

Sezer, Omer Berat, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu (2020). "Financial time series forecasting with deep learning: A systematic literature review: 2005–2019". In: *Applied Soft Computing* 90, p. 106181.

Smith, Taylor G. (Oct. 2023). *pmdarima*. Version 2.0.4. URL: https://github.com/alkaline-ml/pmdarima.

Su, Jianlin et al. (2024). *RoFormer: Enhanced Transformer with Rotary Position Embedding*. arXiv: 2104.09864 [cs.CL]. URL: https://arxiv.org/abs/2104.09864.

Tashman, Leonard J. (2000). "Out-of-sample tests of forecasting accuracy: an analysis and review". In: *International Journal of Forecasting* 16.4. The M3- Competition, pp. 437–450. ISSN: 0169-2070. DOI: https://doi.org/10.1016/S0169-2070(00)00065-0. URL: https://www.sciencedirect.com/science/article/pii/S0169207000000650.

THUML (2025a). *Sundial Base 128M modeling code*. URL: https://huggingface.co/thuml/sundial-base-128m/blob/main/modeling_sundial.py (visited on 05/07/2026).

— (2025b). *Sundial Base 128M time-series generation code*. URL: https://huggingface.co/thuml/sundial-base-128m/blob/main/ts_generation_mixin.py (visited on 05/07/2026).

Wang, Shuzhen (2023). "A Stock Price Prediction Method Based on BiLSTM and Improved Transformer". en. In: *IEEE Access* 11, pp. 104211–104223. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2023.3296308. URL: https://ieeexplore.ieee.org/document/10185006/ (visited on 10/28/2024).

Ye, Jiexia et al. (May 2024). *A Survey of Time Series Foundation Models: Generalizing Time Series Representation with Large Language Model*. arXiv:2405.02358. URL: http://arxiv.org/abs/2405.02358 (visited on 10/25/2024).

Zhou, Ce et al. (2024). "A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT". In: *International Journal of Machine Learning and Cybernetics* 15, pp. 3225–3262.

# Appendix A

# Linear-Spaced Training Results

This appendix collects the linear-spaced training results referenced in the main text. Linear and logarithmic spacing yield consistent findings across all configurations; the log-spaced results are reported in Chapter 5 and the linear-spaced counterparts are provided here for completeness. Definitions of all metrics are given in Section 4.4.

## A.1 RQ1: Limited-Data Regime (Daily)

### A.1.1 Raw Price Forecasting

Table A.1: MAPE (%) for raw price forecasting, linearly-spaced training lengths (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Horizon | Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1d | Naive | **1.2265** | **1.2265** | **1.2265** | **1.2265** | **1.2265** | **1.2265** | **1.2265** | **1.2265** | **1.2265** | **1.2265** |
| | ARIMA | 5.9137 | 4.5792 | **1.2265** | **1.2265** | 1.5279 | **1.2265** | **1.2265** | **1.2265** | **1.2265** | **1.2265** |
| | TimesFM | 1.2846 | 1.2957 | 1.2607 | 1.2731 | 1.2537 | 1.2772 | 1.2974 | 1.3076 | 1.2923 | 1.3085 |
| | Chronos | 1.3288 | 1.3413 | 1.3238 | 1.3376 | 1.3298 | 1.3258 | 1.3174 | 1.3131 | 1.3303 | 1.3502 |
| | Sundial | 1.3623 | 1.3831 | 1.3908 | 1.4161 | 1.3434 | 1.2949 | 1.2942 | 1.2851 | 1.3260 | 1.3166 |
| 5d | Naive | **1.9748** | **1.9748** | **1.9748** | **1.9748** | **1.9748** | **1.9748** | **1.9748** | **1.9748** | **1.9748** | **1.9748** |
| | ARIMA | 9.8097 | 8.9506 | **1.9748** | **1.9748** | 3.8370 | **1.9748** | **1.9748** | **1.9748** | **1.9748** | **1.9748** |
| | TimesFM | 2.0405 | 2.0998 | 2.1006 | 2.0696 | 2.0353 | 1.9769 | 2.0520 | 2.0761 | 2.0865 | 2.0184 |
| | Chronos | 2.1096 | 2.1147 | 2.1203 | 2.1339 | 2.1169 | 2.0877 | 2.0595 | 2.0775 | 2.0979 | 2.0930 |
| | Sundial | 2.4237 | 2.3996 | 2.4982 | 2.5754 | 2.4673 | 2.2103 | 2.3268 | 2.2710 | 2.3634 | 2.3398 |
| 21d | Naive | **3.0611** | **3.0611** | **3.0611** | **3.0611** | **3.0611** | **3.0611** | **3.0611** | **3.0611** | **3.0611** | **3.0611** |
| | ARIMA | 12.0167 | 12.3618 | **3.0611** | **3.0611** | 9.3681 | **3.0611** | **3.0611** | **3.0611** | **3.0611** | **3.0611** |
| | TimesFM | 4.1248 | 4.0393 | 3.9660 | 4.1425 | 4.0832 | 3.9488 | 3.8283 | 3.8503 | 3.6913 | 3.6600 |
| | Chronos | 3.5284 | 3.5511 | 3.4215 | 3.3462 | 3.3567 | 3.2401 | 3.2520 | 3.2902 | 3.3407 | 3.2919 |
| | Sundial | 5.0038 | 4.7735 | 5.0363 | 5.0317 | 4.9761 | 4.6004 | 4.4098 | 4.6720 | 4.7639 | 4.7266 |
| 63d | Naive | 7.1680 | 7.1680 | 7.1680 | **7.1680** | **7.1680** | **7.1680** | 7.1680 | 7.1680 | 7.1680 | 7.1680 |
| | ARIMA | 13.0838 | 13.7079 | 7.1680 | **7.1680** | 14.3847 | **7.1680** | 7.1680 | 7.1680 | 7.1680 | 7.1680 |
| | TimesFM | 10.1453 | 6.5344 | 7.9999 | 8.6112 | 8.6503 | 8.2120 | 8.2541 | 8.3387 | 8.3517 | 8.0762 |
| | Chronos | **6.2533** | **6.1770** | **6.4723** | 7.2246 | 8.0485 | 7.4067 | **6.9116** | **6.5532** | **6.3408** | **6.5582** |
| | Sundial | 10.2129 | 9.6459 | 10.4568 | 10.7176 | 11.2414 | 11.0896 | 9.7751 | 9.8487 | 10.0686 | 9.8594 |

Table A.2: RMSE for raw price forecasting, linearly-spaced training lengths (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Horizon | Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1d | Naive | **4.9629** | **4.9629** | **4.9629** | **4.9629** | **4.9629** | **4.9629** | **4.9629** | **4.9629** | **4.9629** | **4.9629** |
| | ARIMA | 21.6781 | 16.8272 | **4.9629** | **4.9629** | 6.0522 | **4.9629** | **4.9629** | **4.9629** | **4.9629** | **4.9629** |
| | TimesFM | 5.1445 | 5.0929 | 5.0528 | 5.1105 | 5.0152 | 5.0695 | 5.1274 | 5.1666 | 5.1412 | 5.1906 |
| | Chronos | 5.3182 | 5.3328 | 5.2371 | 5.3202 | 5.2784 | 5.2422 | 5.1999 | 5.1940 | 5.2332 | 5.2825 |
| | Sundial | 5.4456 | 5.5037 | 5.4922 | 5.5782 | 5.3398 | 5.2022 | 5.2294 | 5.1455 | 5.2905 | 5.2747 |
| 5d | Naive | **7.8314** | **7.8314** | **7.8314** | **7.8314** | **7.8314** | 7.8314 | **7.8314** | **7.8314** | **7.8314** | **7.8314** |
| | ARIMA | 36.8652 | 34.0535 | **7.8314** | **7.8314** | 15.1837 | 7.8314 | **7.8314** | **7.8314** | **7.8314** | **7.8314** |
| | TimesFM | 7.9504 | 8.0894 | 8.1540 | 8.1227 | 8.0854 | **7.7368** | 8.0461 | 8.1187 | 8.1766 | 7.9246 |
| | Chronos | 8.4397 | 8.3529 | 8.4160 | 8.5190 | 8.2756 | 8.2132 | 8.1476 | 8.2051 | 8.2781 | 8.2384 |
| | Sundial | 9.4486 | 9.3364 | 9.7041 | 9.9696 | 9.6056 | 8.5144 | 8.9374 | 8.8166 | 9.1267 | 9.1502 |
| 21d | Naive | **13.1372** | **13.1372** | **13.1372** | **13.1372** | **13.1372** | **13.1372** | **13.1372** | **13.1372** | **13.1372** | **13.1372** |
| | ARIMA | 44.5202 | 46.1805 | **13.1372** | **13.1372** | 36.7632 | **13.1372** | **13.1372** | **13.1372** | **13.1372** | **13.1372** |
| | TimesFM | 16.5921 | 17.0116 | 16.6035 | 17.2095 | 16.8769 | 16.3895 | 15.9122 | 16.0897 | 15.6657 | 15.2765 |
| | Chronos | 14.5133 | 14.9403 | 15.0804 | 13.9336 | 14.0712 | 13.2762 | 13.3019 | 13.6283 | 13.7887 | 13.7451 |
| | Sundial | 20.1512 | 20.6117 | 20.9342 | 20.7808 | 20.8085 | 19.5135 | 18.6987 | 19.4085 | 20.1419 | 20.0764 |
| 63d | Naive | 29.5861 | 29.5861 | 29.5861 | **29.5861** | **29.5861** | **29.5861** | 29.5861 | 29.5861 | 29.5861 | 29.5861 |
| | ARIMA | 49.8257 | 52.3399 | 29.5861 | **29.5861** | 55.4233 | **29.5861** | 29.5861 | 29.5861 | 29.5861 | 29.5861 |
| | TimesFM | 42.9109 | 29.4771 | 34.6809 | 36.3749 | 37.2034 | 35.6952 | 36.1839 | 36.5546 | 36.7334 | 35.6071 |
| | Chronos | **24.5222** | **24.4295** | **25.9008** | 29.6210 | 33.0199 | 30.6651 | **28.6267** | **28.7971** | **27.8274** | **28.7230** |
| | Sundial | 43.6529 | 42.7027 | 46.6279 | 48.4284 | 51.1510 | 50.4700 | 43.7245 | 42.7735 | 44.7529 | 43.1289 |

Table A.3: MAE for raw price forecasting, linearly-spaced training lengths (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Horizon | Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1d | Naive | **3.7485** | **3.7485** | **3.7485** | **3.7485** | **3.7485** | **3.7485** | **3.7485** | **3.7485** | **3.7485** | **3.7485** |
| | ARIMA | 19.2725 | 14.8970 | **3.7485** | **3.7485** | 4.7455 | **3.7485** | **3.7485** | **3.7485** | **3.7485** | **3.7485** |
| | TimesFM | 3.9270 | 3.9479 | 3.8567 | 3.8910 | 3.8445 | 3.9019 | 3.9706 | 3.9963 | 3.9568 | 3.9982 |
| | Chronos | 4.0621 | 4.1001 | 4.0399 | 4.1047 | 4.0632 | 4.0491 | 4.0332 | 4.0229 | 4.0715 | 4.1365 |
| | Sundial | 4.1523 | 4.2175 | 4.2529 | 4.3367 | 4.1272 | 3.9899 | 3.9870 | 3.9558 | 4.0793 | 4.0549 |
| 5d | Naive | **6.0328** | **6.0328** | **6.0328** | **6.0328** | **6.0328** | 6.0328 | **6.0328** | **6.0328** | **6.0328** | **6.0328** |
| | ARIMA | 32.1573 | 29.3235 | **6.0328** | **6.0328** | 12.3248 | 6.0328 | **6.0328** | **6.0328** | **6.0328** | **6.0328** |
| | TimesFM | 6.2218 | 6.3779 | 6.3671 | 6.2820 | 6.2056 | **6.0250** | 6.2491 | 6.3240 | 6.3554 | 6.1552 |
| | Chronos | 6.4522 | 6.4674 | 6.4657 | 6.5207 | 6.4679 | 6.3846 | 6.3009 | 6.3582 | 6.4146 | 6.4010 |
| | Sundial | 7.4139 | 7.3704 | 7.6708 | 7.9226 | 7.6210 | 6.8434 | 7.1753 | 7.0043 | 7.2897 | 7.2509 |
| 21d | Naive | **9.5962** | **9.5962** | **9.5962** | **9.5962** | **9.5962** | **9.5962** | **9.5962** | **9.5962** | **9.5962** | **9.5962** |
| | ARIMA | 39.4096 | 40.6093 | **9.5962** | **9.5962** | 30.7089 | **9.5962** | **9.5962** | **9.5962** | **9.5962** | **9.5962** |
| | TimesFM | 12.7218 | 12.6138 | 12.2810 | 12.8690 | 12.6949 | 12.3224 | 11.8834 | 11.9508 | 11.5060 | 11.3798 |
| | Chronos | 10.8410 | 10.9762 | 10.5734 | 10.4479 | 10.5232 | 10.2100 | 10.2530 | 10.4008 | 10.5413 | 10.3870 |
| | Sundial | 15.4205 | 15.0609 | 15.8267 | 15.8091 | 15.5818 | 14.5247 | 13.9510 | 14.7045 | 15.0787 | 14.9551 |
| 63d | Naive | 23.2427 | 23.2427 | 23.2427 | **23.2427** | **23.2427** | **23.2427** | 23.2427 | 23.2427 | 23.2427 | 23.2427 |
| | ARIMA | 43.0837 | 45.2497 | 23.2427 | **23.2427** | 47.6474 | **23.2427** | 23.2427 | 23.2427 | 23.2427 | 23.2427 |
| | TimesFM | 31.0596 | 21.5431 | 26.1750 | 27.9961 | 28.1967 | 26.8554 | 27.0855 | 27.3686 | 27.4123 | 26.5692 |
| | Chronos | **20.1430** | **19.9799** | **20.9526** | 23.7027 | 26.1963 | 24.3371 | **22.7850** | **21.8287** | **21.1070** | **21.7884** |
| | Sundial | 33.4506 | 31.8908 | 34.6578 | 35.6201 | 37.3689 | 36.8622 | 32.4022 | 32.4707 | 33.3985 | 32.5286 |

### A.1.2 Fractional Change Forecasting

Table A.4: RMSE for fractional change forecasting, linearly-spaced training lengths (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Horizon | Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1d | Naive | 0.0248 | 0.0248 | 0.0248 | 0.0248 | 0.0248 | 0.0248 | 0.0248 | 0.0248 | 0.0248 | 0.0248 |
| | ARIMA | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** |
| | TimesFM | 0.0171 | 0.0170 | 0.0170 | 0.0169 | 0.0170 | 0.0169 | 0.0168 | 0.0169 | 0.0168 | 0.0169 |
| | Chronos | 0.0171 | 0.0169 | 0.0168 | 0.0168 | 0.0168 | 0.0167 | 0.0167 | 0.0167 | **0.0166** | 0.0167 |
| | Sundial | 0.0172 | 0.0171 | 0.0169 | 0.0168 | 0.0168 | 0.0168 | 0.0167 | 0.0170 | 0.0168 | 0.0167 |
| 5d | Naive | 0.0221 | 0.0221 | 0.0221 | 0.0221 | 0.0221 | 0.0221 | 0.0221 | 0.0221 | 0.0221 | 0.0221 |
| | ARIMA | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** | 0.0166 | **0.0166** | **0.0166** | **0.0166** | **0.0166** |
| | TimesFM | 0.0170 | 0.0170 | 0.0167 | 0.0170 | 0.0168 | 0.0168 | 0.0168 | 0.0167 | 0.0169 | 0.0169 |
| | Chronos | **0.0166** | **0.0166** | **0.0166** | **0.0166** | **0.0166** | 0.0166 | **0.0166** | **0.0166** | **0.0166** | **0.0166** |
| | Sundial | 0.0169 | 0.0170 | 0.0169 | 0.0168 | 0.0170 | **0.0164** | **0.0166** | 0.0169 | 0.0167 | 0.0169 |
| 21d | Naive | 0.0219 | 0.0219 | 0.0219 | 0.0219 | 0.0219 | 0.0219 | 0.0219 | 0.0219 | 0.0219 | 0.0219 |
| | ARIMA | 0.0166 | **0.0166** | 0.0166 | 0.0166 | 0.0166 | **0.0166** | **0.0166** | 0.0166 | **0.0166** | 0.0166 |
| | TimesFM | 0.0167 | 0.0170 | 0.0166 | 0.0166 | 0.0166 | 0.0167 | **0.0166** | 0.0166 | **0.0166** | 0.0166 |
| | Chronos | 0.0167 | **0.0166** | **0.0165** | 0.0165 | **0.0165** | **0.0166** | **0.0166** | 0.0166 | **0.0166** | 0.0166 |
| | Sundial | **0.0165** | **0.0166** | 0.0166 | **0.0164** | 0.0166 | 0.0169 | **0.0166** | **0.0164** | **0.0166** | **0.0165** |
| 63d | Naive | 0.0190 | 0.0190 | 0.0190 | 0.0190 | 0.0190 | 0.0190 | 0.0190 | 0.0190 | 0.0190 | 0.0190 |
| | ARIMA | **0.0166** | **0.0166** | 0.0166 | 0.0166 | 0.0166 | **0.0166** | 0.0166 | 0.0166 | **0.0166** | 0.0166 |
| | TimesFM | 0.0189 | 0.0172 | 0.0166 | 0.0166 | 0.0167 | 0.0167 | 0.0167 | 0.0166 | **0.0166** | 0.0166 |
| | Chronos | 0.0168 | **0.0166** | **0.0165** | **0.0165** | **0.0165** | **0.0166** | 0.0166 | 0.0166 | **0.0166** | 0.0166 |
| | Sundial | **0.0166** | 0.0168 | 0.0166 | 0.0167 | **0.0165** | 0.0171 | **0.0165** | **0.0165** | 0.0167 | **0.0164** |

Table A.5: MAE for fractional change forecasting, linearly-spaced training lengths (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Horizon | Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1d | Naive | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 |
| | ARIMA | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** |
| | TimesFM | 0.0126 | 0.0126 | 0.0126 | 0.0126 | 0.0127 | 0.0126 | 0.0125 | 0.0125 | 0.0125 | 0.0125 |
| | Chronos | 0.0126 | 0.0125 | 0.0125 | 0.0125 | 0.0125 | 0.0124 | 0.0124 | 0.0124 | 0.0124 | 0.0124 |
| | Sundial | 0.0127 | 0.0128 | 0.0126 | 0.0125 | 0.0125 | 0.0126 | 0.0124 | 0.0127 | **0.0123** | 0.0125 |
| 5d | Naive | 0.0168 | 0.0168 | 0.0168 | 0.0168 | 0.0168 | 0.0168 | 0.0168 | 0.0168 | 0.0168 | 0.0168 |
| | ARIMA | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** |
| | TimesFM | 0.0127 | 0.0127 | 0.0126 | 0.0128 | 0.0126 | 0.0125 | 0.0125 | 0.0124 | 0.0127 | 0.0126 |
| | Chronos | 0.0125 | 0.0124 | 0.0124 | 0.0124 | 0.0124 | 0.0124 | 0.0124 | 0.0124 | **0.0123** | 0.0124 |
| | Sundial | 0.0125 | 0.0127 | 0.0127 | 0.0125 | 0.0127 | **0.0123** | **0.0123** | 0.0128 | **0.0123** | 0.0126 |
| 21d | Naive | 0.0177 | 0.0177 | 0.0177 | 0.0177 | 0.0177 | 0.0177 | 0.0177 | 0.0177 | 0.0177 | 0.0177 |
| | ARIMA | **0.0123** | **0.0123** | **0.0123** | 0.0123 | **0.0123** | **0.0123** | 0.0123 | 0.0123 | **0.0123** | **0.0123** |
| | TimesFM | 0.0124 | 0.0128 | 0.0124 | 0.0124 | 0.0124 | 0.0125 | 0.0124 | 0.0124 | 0.0124 | 0.0124 |
| | Chronos | 0.0125 | 0.0124 | **0.0123** | 0.0123 | **0.0123** | 0.0124 | 0.0123 | 0.0123 | **0.0123** | **0.0123** |
| | Sundial | 0.0124 | 0.0124 | **0.0123** | **0.0122** | **0.0123** | 0.0126 | **0.0122** | **0.0122** | **0.0123** | **0.0123** |
| 63d | Naive | 0.0148 | 0.0148 | 0.0148 | 0.0148 | 0.0148 | 0.0148 | 0.0148 | 0.0148 | 0.0148 | 0.0148 |
| | ARIMA | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | **0.0123** | 0.0123 | **0.0123** | **0.0123** | 0.0123 |
| | TimesFM | 0.0141 | 0.0130 | 0.0124 | 0.0124 | 0.0125 | 0.0125 | 0.0125 | 0.0125 | 0.0125 | 0.0124 |
| | Chronos | 0.0127 | 0.0126 | 0.0124 | 0.0124 | 0.0125 | 0.0125 | 0.0124 | 0.0124 | 0.0124 | 0.0124 |
| | Sundial | 0.0124 | 0.0127 | 0.0125 | 0.0126 | 0.0125 | 0.0127 | **0.0122** | 0.0125 | 0.0124 | **0.0121** |

Table A.6: Mean Directional Accuracy (%) for fractional change forecasting, linearly-spaced training lengths (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Horizon | Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1d | Naive | 50.4000 | 50.4000 | 50.4000 | 50.4000 | 50.4000 | **50.4000** | 50.4000 | 50.4000 | 50.4000 | 50.4000 |
| | ARIMA | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | TimesFM | **54.0000** | **54.4000** | **52.0000** | 50.8000 | 51.2000 | 49.6000 | 47.6000 | 48.0000 | 47.2000 | 50.0000 |
| | Chronos | 51.6000 | 53.6000 | 51.6000 | **52.4000** | 51.2000 | **50.4000** | **51.6000** | **51.6000** | 52.0000 | **52.0000** |
| | Sundial | 51.2000 | 51.8000 | 48.2000 | 51.0000 | **51.6000** | 48.2000 | 51.0000 | 48.8000 | **53.4000** | 51.2000 |
| 5d | Naive | 49.6000 | 49.6000 | 49.6000 | 49.6000 | 49.6000 | 49.6000 | 49.6000 | 49.6000 | 49.6000 | 49.6000 |
| | ARIMA | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | TimesFM | 49.6000 | 52.8000 | 52.0000 | 51.6000 | 51.2000 | 52.4000 | 49.6000 | 51.6000 | 46.0000 | 46.4000 |
| | Chronos | **52.8000** | **53.6000** | **52.4000** | **52.0000** | **52.0000** | 50.0000 | 52.4000 | **52.0000** | 52.4000 | **51.6000** |
| | Sundial | **52.8000** | 50.4000 | 46.8000 | 48.8000 | 48.4000 | **53.2000** | **54.8000** | 48.4000 | **55.2000** | 50.4000 |
| 21d | Naive | 52.0000 | 52.0000 | 52.0000 | 52.0000 | **52.0000** | **52.0000** | 52.0000 | 52.0000 | 52.0000 | 52.0000 |
| | ARIMA | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | TimesFM | **54.4000** | 49.6000 | 52.0000 | 53.2000 | 49.6000 | 49.6000 | 47.2000 | 49.6000 | 48.4000 | 49.6000 |
| | Chronos | 52.4000 | 50.4000 | 53.2000 | 53.2000 | **52.0000** | 46.8000 | 49.2000 | 47.6000 | 50.0000 | 48.0000 |
| | Sundial | 48.4000 | **54.0000** | **55.6000** | **54.0000** | 49.6000 | 49.6000 | **55.2000** | **54.8000** | **57.2000** | **54.4000** |
| 63d | Naive | **58.8000** | **58.8000** | **58.8000** | **58.8000** | **58.8000** | **58.8000** | **58.8000** | **58.8000** | **58.8000** | **58.8000** |
| | ARIMA | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | TimesFM | 54.4000 | 48.8000 | 50.8000 | 52.8000 | 49.2000 | 48.4000 | 46.8000 | 48.0000 | 48.0000 | 48.0000 |
| | Chronos | 51.2000 | 52.4000 | 53.6000 | 52.0000 | 47.6000 | 46.8000 | 47.2000 | 50.0000 | 52.8000 | 48.0000 |
| | Sundial | 51.2000 | 45.2000 | 51.6000 | 47.6000 | 49.2000 | 52.4000 | 56.0000 | 52.4000 | 51.6000 | 57.2000 |

## A.2 RQ2: Limited-Data Regime (Hourly)

### A.2.1 Raw Price Forecasting

Table A.7: MAPE (%) for raw price forecasting, linearly-spaced training lengths (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** |
| ARIMA | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** | **0.3994** |
| TimesFM | 0.4493 | 0.4498 | 0.4555 | 0.4482 | 0.4416 | 0.4353 | 0.4403 | 0.4420 | 0.4420 | 0.4421 |
| Chronos | 0.4968 | 0.4978 | 0.4979 | 0.5000 | 0.5100 | 0.5271 | 0.5289 | 0.5274 | 0.5293 | 0.5345 |
| Sundial | 0.4245 | 0.4284 | 0.4212 | 0.4245 | 0.4286 | 0.4285 | 0.4267 | 0.4266 | 0.4286 | 0.4282 |

Table A.8: RMSE for raw price forecasting, linearly-spaced training lengths (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** |
| ARIMA | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** | **1.8107** |
| TimesFM | 1.9672 | 1.9796 | 1.9815 | 1.9688 | 1.9534 | 1.9372 | 1.9489 | 1.9393 | 1.9493 | 1.9462 |
| Chronos | 2.1367 | 2.1463 | 2.1527 | 2.1610 | 2.1879 | 2.2396 | 2.2465 | 2.2369 | 2.2487 | 2.2692 |
| Sundial | 1.8920 | 1.9029 | 1.8827 | 1.8901 | 1.9049 | 1.9006 | 1.8974 | 1.8913 | 1.9010 | 1.8993 |

Table A.9: MAE for raw price forecasting, linearly-spaced training lengths (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** |
| ARIMA | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** | **1.2276** |
| TimesFM | 1.3859 | 1.3889 | 1.4088 | 1.3855 | 1.3653 | 1.3430 | 1.3603 | 1.3666 | 1.3669 | 1.3654 |
| Chronos | 1.5454 | 1.5507 | 1.5536 | 1.5610 | 1.5900 | 1.6387 | 1.6467 | 1.6419 | 1.6498 | 1.6668 |
| Sundial | 1.3060 | 1.3186 | 1.3001 | 1.3091 | 1.3256 | 1.3286 | 1.3228 | 1.3220 | 1.3287 | 1.3266 |

### A.2.2 Fractional Change Forecasting

Table A.10: RMSE for fractional change forecasting, linearly-spaced training lengths (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 0.0084 | 0.0084 | 0.0084 | 0.0084 | 0.0084 | 0.0084 | 0.0084 | 0.0084 | 0.0084 | 0.0084 |
| ARIMA | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** |
| TimesFM | 0.0061 | **0.0060** | 0.0061 | 0.0061 | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** |
| Chronos | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** | **0.0060** |
| Sundial | 0.0061 | 0.0061 | 0.0061 | 0.0061 | **0.0060** | **0.0060** | **0.0060** | 0.0061 | 0.0061 | **0.0060** |

Table A.11: MAE for fractional change forecasting, linearly-spaced training lengths (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 0.0058 | 0.0058 | 0.0058 | 0.0058 | 0.0058 | 0.0058 | 0.0058 | 0.0058 | 0.0058 | 0.0058 |
| ARIMA | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** |
| TimesFM | 0.0041 | 0.0041 | 0.0041 | 0.0041 | 0.0041 | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** |
| Chronos | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** | **0.0040** |
| Sundial | 0.0041 | 0.0041 | 0.0041 | 0.0041 | 0.0041 | **0.0040** | **0.0040** | 0.0041 | 0.0041 | 0.0041 |

Table A.12: Mean Directional Accuracy (%) for fractional change forecasting, linearly-spaced training lengths (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 49.0286 | 49.0286 | 49.0286 | 49.0286 | 49.0286 | 49.0286 | **49.0286** | 49.0286 | 49.0286 | 49.0286 |
| ARIMA | 0.3429 | 0.3429 | 0.3429 | 0.3429 | 0.3429 | 0.3429 | 0.3429 | 0.3429 | 0.3429 | 0.3429 |
| TimesFM | **49.2000** | 48.5143 | **49.8286** | 48.0571 | 47.3143 | 48.2286 | 48.6857 | 48.6286 | 48.8000 | **49.3143** |
| Chronos | 48.6286 | 48.2857 | 48.6286 | 48.6857 | 47.3143 | 48.3429 | 48.2857 | 48.1143 | **49.0857** | 48.2286 |
| Sundial | 48.4571 | **49.6571** | 49.6571 | **50.1714** | **49.2000** | **49.4286** | **49.0286** | **49.1429** | 47.8286 | 48.5714 |

## A.3 RQ3: Market Disruption Resilience (Linear-Spaced)

### A.3.1 2008 Crisis: Raw Prices

Table A.13: MAPE (%) for raw price forecasting, 2008 Financial Crisis with linearly-spaced training (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **0.5846** | **0.5846** | **0.5846** | **0.5846** | **0.5846** | **0.5846** | **0.5846** | **0.5846** | **0.5846** | **0.5846** |
| ARIMA | 0.5933 | **0.5846** | **0.5846** | 0.5862 | 0.5861 | 0.5860 | 0.5863 | 0.5859 | 0.5859 | 0.5858 |
| TimesFM | 0.7115 | 0.6677 | 0.6806 | 0.6398 | 0.6399 | 0.6200 | 0.6330 | 0.6631 | 0.6345 | 0.6488 |
| Chronos | 0.6570 | 0.6899 | 0.7532 | 0.8309 | 0.8282 | 0.7382 | 0.6982 | 0.6869 | 0.7084 | 0.7311 |
| Sundial | 0.6623 | 0.6603 | 0.6682 | 0.6236 | 0.6191 | 0.6336 | 0.6148 | 0.6050 | 0.5998 | 0.5998 |

Table A.14: RMSE for raw price forecasting, 2008 Financial Crisis with linearly-spaced training (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **0.2819** | **0.2819** | **0.2819** | 0.2819 | 0.2819 | 0.2819 | 0.2819 | 0.2819 | 0.2819 | 0.2819 |
| ARIMA | 0.2820 | **0.2819** | **0.2819** | **0.2807** | **0.2806** | **0.2806** | **0.2806** | **0.2805** | **0.2805** | **0.2805** |
| TimesFM | 0.3043 | 0.2977 | 0.3033 | 0.2911 | 0.2925 | 0.2907 | 0.2965 | 0.2973 | 0.2930 | 0.2955 |
| Chronos | 0.2991 | 0.3058 | 0.3171 | 0.3311 | 0.3332 | 0.3121 | 0.3046 | 0.3023 | 0.3072 | 0.3123 |
| Sundial | 0.2988 | 0.2974 | 0.3028 | 0.2869 | 0.2857 | 0.2876 | 0.2842 | 0.2808 | 0.2806 | 0.2812 |

Table A.15: MAE for raw price forecasting, 2008 Financial Crisis with linearly-spaced training (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **0.1734** | **0.1734** | **0.1734** | **0.1734** | **0.1734** | **0.1734** | **0.1734** | **0.1734** | **0.1734** | **0.1734** |
| ARIMA | 0.1759 | **0.1734** | **0.1734** | 0.1738 | 0.1737 | 0.1737 | 0.1738 | 0.1737 | 0.1737 | 0.1736 |
| TimesFM | 0.2098 | 0.1981 | 0.2025 | 0.1896 | 0.1897 | 0.1840 | 0.1883 | 0.1968 | 0.1885 | 0.1924 |
| Chronos | 0.1943 | 0.2040 | 0.2216 | 0.2437 | 0.2431 | 0.2173 | 0.2061 | 0.2029 | 0.2090 | 0.2155 |
| Sundial | 0.1966 | 0.1959 | 0.1986 | 0.1846 | 0.1833 | 0.1876 | 0.1819 | 0.1788 | 0.1774 | 0.1774 |

### A.3.2 2008 Crisis: Fractional Change

Table A.16: RMSE for fractional change forecasting, 2008 Financial Crisis with linearly-spaced training (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 0.0135 | 0.0135 | 0.0135 | 0.0135 | 0.0135 | 0.0135 | 0.0135 | 0.0135 | 0.0135 | 0.0135 |
| ARIMA | 0.0099 | **0.0092** | **0.0092** | **0.0091** | **0.0091** | **0.0091** | **0.0091** | **0.0091** | **0.0091** | **0.0091** |
| TimesFM | 0.0093 | **0.0092** | **0.0092** | 0.0092 | 0.0092 | 0.0092 | 0.0092 | 0.0092 | 0.0092 | 0.0092 |
| Chronos | **0.0092** | **0.0092** | **0.0092** | 0.0092 | 0.0092 | 0.0092 | 0.0092 | 0.0092 | 0.0092 | 0.0092 |
| Sundial | 0.0093 | **0.0092** | **0.0092** | 0.0093 | 0.0092 | 0.0092 | 0.0092 | 0.0092 | 0.0092 | 0.0092 |

Table A.17: MAE for fractional change forecasting, 2008 Financial Crisis with linearly-spaced training (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 0.0093 | 0.0093 | 0.0093 | 0.0093 | 0.0093 | 0.0093 | 0.0093 | 0.0093 | 0.0093 | 0.0093 |
| ARIMA | 0.0066 | **0.0058** | **0.0058** | **0.0058** | **0.0058** | **0.0058** | **0.0058** | **0.0058** | **0.0058** | **0.0058** |
| TimesFM | 0.0060 | 0.0059 | 0.0059 | 0.0059 | 0.0059 | 0.0059 | 0.0059 | 0.0059 | 0.0059 | 0.0059 |
| Chronos | **0.0059** | **0.0058** | **0.0058** | **0.0058** | **0.0058** | **0.0058** | **0.0058** | **0.0058** | **0.0058** | **0.0058** |
| Sundial | 0.0060 | 0.0059 | 0.0059 | 0.0060 | 0.0059 | 0.0059 | 0.0060 | 0.0060 | 0.0060 | 0.0059 |

Table A.18: Mean Directional Accuracy (%) for fractional change forecasting, 2008 Financial Crisis with linearly-spaced training (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 44.9477 | 44.9477 | 44.9477 | 44.9477 | 44.9477 | 44.9477 | 44.9477 | 44.9477 | 44.9477 | 44.9477 |
| ARIMA | **54.7038** | 1.3937 | 1.3937 | **52.2648** | **52.2648** | 52.2648 | **52.2648** | 52.2648 | **52.6132** | **52.2648** |
| TimesFM | 52.9617 | 50.5226 | 47.7352 | 51.2195 | 50.1742 | 50.1742 | 50.8711 | 49.4774 | 49.4774 | 48.4321 |
| Chronos | 50.5226 | **52.6132** | 52.2648 | 50.8711 | **52.2648** | **52.6132** | **52.2648** | **52.6132** | 51.9164 | 51.2195 |
| Sundial | 49.4774 | 50.1742 | **53.6585** | 49.1289 | 50.1742 | 44.9477 | 43.9024 | 47.7352 | 48.0836 | 47.7352 |

### A.3.3 2020 Crash: Raw Prices

Table A.19: MAPE (%) for raw price forecasting, 2020 COVID-19 Crash with linearly-spaced training (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 1.1821 | 1.1821 | 1.1821 | 1.1821 | 1.1821 | 1.1821 | 1.1821 | 1.1821 | 1.1821 | 1.1821 |
| ARIMA | **1.1709** | **1.1719** | 1.1711 | **1.1652** | 1.1703 | 1.1640 | **1.1634** | 1.1626 | **1.1619** | **1.1616** |
| TimesFM | 1.2358 | 1.2194 | 1.2260 | 1.2170 | 1.2389 | 1.2015 | 1.2139 | 1.2259 | 1.2134 | 1.1936 |
| Chronos | 1.2282 | 1.2294 | 1.2536 | 1.2901 | 1.3015 | 1.2727 | 1.2628 | 1.2475 | 1.2328 | 1.2320 |
| Sundial | 1.1876 | 1.1843 | **1.1687** | 1.1654 | **1.1636** | **1.1561** | 1.1673 | **1.1538** | 1.1677 | 1.1679 |

Table A.20: RMSE for raw price forecasting, 2020 COVID-19 Crash with linearly-spaced training (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 2.7548 | 2.7548 | 2.7548 | 2.7548 | 2.7548 | 2.7548 | 2.7548 | 2.7548 | 2.7548 | 2.7548 |
| ARIMA | **2.6856** | **2.6928** | **2.6917** | 2.6737 | 2.6841 | 2.6739 | 2.6743 | 2.6746 | 2.6755 | 2.6758 |
| TimesFM | 2.7971 | 2.7239 | 2.6974 | 2.7214 | 2.7185 | 2.7156 | 2.7320 | 2.7479 | 2.7377 | 2.7063 |
| Chronos | 2.6980 | 2.7453 | 2.7971 | 2.8873 | 2.9021 | 2.8380 | 2.8041 | 2.7777 | 2.7557 | 2.7578 |
| Sundial | 2.7341 | 2.7280 | 2.6972 | **2.6716** | **2.6552** | **2.6225** | **2.6409** | **2.6253** | **2.6490** | **2.6455** |

Table A.21: MAE for raw price forecasting, 2020 COVID-19 Crash with linearly-spaced training (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 1.8353 | 1.8353 | 1.8353 | 1.8353 | 1.8353 | 1.8353 | 1.8353 | 1.8353 | 1.8353 | 1.8353 |
| ARIMA | **1.8219** | **1.8210** | 1.8196 | 1.8094 | 1.8177 | 1.8076 | **1.8065** | 1.8054 | **1.8043** | **1.8039** |
| TimesFM | 1.9194 | 1.8914 | 1.9046 | 1.8943 | 1.9230 | 1.8695 | 1.8878 | 1.9122 | 1.8901 | 1.8524 |
| Chronos | 1.9103 | 1.9142 | 1.9507 | 2.0058 | 2.0226 | 1.9778 | 1.9639 | 1.9409 | 1.9187 | 1.9175 |
| Sundial | 1.8418 | 1.8337 | **1.8112** | **1.8062** | **1.8055** | **1.7946** | 1.8121 | **1.7908** | 1.8121 | 1.8132 |

### A.3.4 2020 Crash: Fractional Change

Table A.22: RMSE for fractional change forecasting, 2020 COVID-19 Crash with linearly-spaced training (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 0.0287 | 0.0287 | 0.0287 | 0.0287 | 0.0287 | 0.0287 | 0.0287 | 0.0287 | 0.0287 | 0.0287 |
| ARIMA | **0.0176** | **0.0176** | 0.0176 | **0.0174** | **0.0174** | **0.0175** | **0.0174** | **0.0174** | **0.0175** | **0.0175** |
| TimesFM | 0.0186 | 0.0186 | 0.0184 | 0.0183 | 0.0183 | 0.0184 | 0.0185 | 0.0185 | 0.0184 | 0.0184 |
| Chronos | 0.0181 | 0.0182 | 0.0182 | 0.0181 | 0.0181 | 0.0181 | 0.0181 | 0.0181 | 0.0181 | 0.0181 |
| Sundial | 0.0183 | 0.0182 | **0.0175** | 0.0179 | 0.0180 | 0.0181 | 0.0179 | 0.0181 | 0.0180 | 0.0181 |

Table A.23: MAE for fractional change forecasting, 2020 COVID-19 Crash with linearly-spaced training (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 | 0.0186 |
| ARIMA | **0.0117** | **0.0117** | **0.0117** | **0.0116** | **0.0116** | **0.0116** | **0.0116** | **0.0116** | **0.0116** | **0.0116** |
| TimesFM | 0.0123 | 0.0123 | 0.0122 | 0.0121 | 0.0121 | 0.0121 | 0.0122 | 0.0123 | 0.0121 | 0.0121 |
| Chronos | 0.0119 | 0.0120 | 0.0120 | 0.0119 | 0.0119 | 0.0119 | 0.0119 | 0.0119 | 0.0119 | 0.0119 |
| Sundial | 0.0120 | 0.0119 | **0.0117** | 0.0117 | 0.0118 | 0.0120 | 0.0117 | 0.0119 | 0.0118 | 0.0120 |

Table A.24: Mean Directional Accuracy (%) for fractional change forecasting, 2020 COVID-19 Crash with linearly-spaced training (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 48.0836 | 48.0836 | 48.0836 | 48.0836 | 48.0836 | 48.0836 | 48.0836 | 48.0836 | 48.0836 | 48.0836 |
| ARIMA | **52.2648** | **52.6132** | **54.7038** | 51.9164 | **52.9617** | **51.9164** | 51.9164 | 51.9164 | **51.9164** | **51.9164** |
| TimesFM | 46.6899 | 47.7352 | 48.0836 | 46.6899 | 48.7805 | 46.6899 | 47.0383 | 48.4321 | 51.2195 | 50.1742 |
| Chronos | 48.7805 | 46.3415 | 42.8571 | 43.9024 | 46.6899 | 45.6446 | 44.9477 | 44.5993 | 43.2056 | 41.4634 |
| Sundial | **52.2648** | 50.5226 | **54.7038** | **54.3554** | **52.9617** | 47.3868 | **53.3101** | **52.9617** | 51.5679 | 47.7352 |

### A.3.5 2022 Downturn: Raw Prices

Table A.25: MAPE (%) for raw price forecasting, 2022 Market Downturn with linearly-spaced training (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** |
| ARIMA | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** | **0.5897** |
| TimesFM | 0.6588 | 0.6521 | 0.6491 | 0.6547 | 0.6844 | 0.7011 | 0.6350 | 0.6363 | 0.6309 | 0.6446 |
| Chronos | 0.6917 | 0.6972 | 0.7060 | 0.7197 | 0.6914 | 0.6907 | 0.7092 | 0.7111 | 0.7066 | 0.7324 |
| Sundial | 0.6466 | 0.6505 | 0.6464 | 0.6262 | 0.6208 | 0.6243 | 0.6314 | 0.6262 | 0.6246 | 0.6249 |

Table A.26: RMSE for raw price forecasting, 2022 Market Downturn with linearly-spaced training (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** |
| ARIMA | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** | **2.1999** |
| TimesFM | 2.4135 | 2.3532 | 2.3318 | 2.3719 | 2.4884 | 2.5412 | 2.3080 | 2.3063 | 2.3078 | 2.3370 |
| Chronos | 2.5171 | 2.5174 | 2.5295 | 2.5883 | 2.5131 | 2.5165 | 2.5627 | 2.5734 | 2.5562 | 2.6564 |
| Sundial | 2.3585 | 2.3497 | 2.3563 | 2.2934 | 2.2865 | 2.2813 | 2.3158 | 2.2959 | 2.2871 | 2.2798 |

Table A.27: MAE for raw price forecasting, 2022 Market Downturn with linearly-spaced training (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** |
| ARIMA | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** | **1.6353** |
| TimesFM | 1.8230 | 1.8043 | 1.7974 | 1.8133 | 1.8942 | 1.9381 | 1.7588 | 1.7650 | 1.7480 | 1.7836 |
| Chronos | 1.9194 | 1.9326 | 1.9552 | 1.9908 | 1.9138 | 1.9106 | 1.9605 | 1.9638 | 1.9514 | 2.0185 |
| Sundial | 1.7878 | 1.7995 | 1.7893 | 1.7351 | 1.7205 | 1.7281 | 1.7493 | 1.7339 | 1.7290 | 1.7309 |

### A.3.6 2022 Downturn: Fractional Change

Table A.28: RMSE for fractional change forecasting, 2022 Market Downturn with linearly-spaced training (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 0.0112 | 0.0112 | 0.0112 | 0.0112 | 0.0112 | 0.0112 | 0.0112 | 0.0112 | 0.0112 | 0.0112 |
| ARIMA | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** |
| TimesFM | 0.0081 | 0.0081 | 0.0081 | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** |
| Chronos | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** | **0.0080** |
| Sundial | 0.0081 | 0.0083 | 0.0083 | 0.0081 | **0.0080** | 0.0081 | 0.0081 | 0.0081 | 0.0081 | 0.0081 |

Table A.29: MAE for fractional change forecasting, 2022 Market Downturn with linearly-spaced training (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | 0.0085 | 0.0085 | 0.0085 | 0.0085 | 0.0085 | 0.0085 | 0.0085 | 0.0085 | 0.0085 | 0.0085 |
| ARIMA | **0.0059** | **0.0059** | **0.0059** | **0.0059** | **0.0059** | 0.0059 | **0.0059** | **0.0059** | **0.0059** | **0.0059** |
| TimesFM | 0.0060 | 0.0060 | 0.0060 | **0.0059** | **0.0059** | **0.0058** | **0.0059** | **0.0059** | **0.0059** | **0.0059** |
| Chronos | **0.0059** | **0.0059** | **0.0059** | **0.0059** | **0.0059** | 0.0059 | **0.0059** | **0.0059** | **0.0059** | **0.0059** |
| Sundial | 0.0061 | 0.0062 | 0.0062 | 0.0061 | **0.0059** | 0.0060 | 0.0060 | 0.0060 | 0.0060 | 0.0060 |

Table A.30: Mean Directional Accuracy (%) for fractional change forecasting, 2022 Market Downturn with linearly-spaced training (25–250 trading days). Bold indicates best performance; underline indicates second best.

| Method | 25d | 50d | 75d | 100d | 125d | 150d | 175d | 200d | 225d | 250d |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive | **54.3554** | **54.3554** | **54.3554** | **54.3554** | **54.3554** | **54.3554** | **54.3554** | **54.3554** | **54.3554** | **54.3554** |
| ARIMA | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| TimesFM | 51.2195 | 51.5679 | 50.5226 | **54.3554** | 52.6132 | 52.6132 | 52.2648 | 50.5226 | 51.5679 | 52.2648 |
| Chronos | 49.1289 | 49.4774 | 49.1289 | 48.0836 | 49.1289 | 50.1742 | 50.1742 | 50.1742 | 49.8258 | 50.1742 |
| Sundial | 46.3415 | 52.6132 | 47.0383 | 47.0383 | 51.5679 | 45.9930 | 50.8711 | 49.1289 | 50.5226 | 50.1742 |