Re(Visiting) Time Series Foundation Models in Finance

Eghbal Rahimikia*
Alliance Manchester Business School, University of Manchester
eghbal.rahimikia@manchester.ac.uk

Hao Ni
Department of Mathematics, University College London (UCL)
h.ni@ucl.ac.uk

Weiguan Wang
School of Economics, Shanghai University
weiguanwang@shu.edu.cn

November 2025

Abstract

Financial time series forecasting is central to trading, portfolio optimization, and risk management, yet it remains challenging due to noisy, non-stationary, and heterogeneous data. Recent advances in time series foundation models (TSFMs), inspired by large language models, offer a new paradigm for learning generalizable temporal representations from large and diverse datasets. This paper presents the first comprehensive empirical study of TSFMs in global financial markets. Using a large-scale dataset of daily excess returns across diverse markets, we evaluate zero-shot inference, fine-tuning, and pre-training from scratch against strong benchmark models. We find that off-the-shelf pre-trained TSFMs perform poorly in zero-shot and fine-tuning settings, whereas models pre-trained from scratch on financial data achieve substantial forecasting and economic improvements, underscoring the value of domain-specific adaptation. Increasing the dataset size, incorporating synthetic data augmentation, and applying hyperparameter tuning further enhance performance.

Keywords: *Time Series Foundation Models, Transformer Models, Asset Return Predictability, Cross-Sectional Stock Returns, Transfer Learning*

JEL Classification: *C53, C45, C63, G17*

---

* Corresponding author.

The authors acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR). Isambard-AI is operated by the University of Bristol and is funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023]. The authors would also like to acknowledge the assistance provided by Research IT and the use of the Computational Shared Facility at The University of Manchester. This work also made use of the facilities of the N8 Centre of Excellence in Computationally Intensive Research (N8 CIR), funded by the N8 research partnership and EPSRC (Grant No. EP/T022167/1). The Centre is coordinated by the Universities of Durham, Manchester, and York. Hao Ni is supported by the EPSRC under the program grant EP/S026347/1 and the Alan Turing Institute under the EPSRC grant EP/N510129/1. Weiguan Wang appreciates the financial support by the National Natural Science Foundation of China (No. 72201158). This work is also supported by the Shanghai Technical Service Center of Science and Engineering Computing, Shanghai University. All models are available through our portal at `FinText.ai` and the Hugging Face repository at `https://huggingface.co/FinText`.

1 Introduction

Financial time series forecasting is a central problem in quantitative finance, underpinning trading, portfolio construction, and risk management. Decades of research have produced a wide spectrum of models, from classical econometric approaches (e.g., ARIMA, GARCH) to modern machine learning (ML) models such as tree-based ensembles and deep neural networks. Yet reliable forecasting remains difficult. Financial data are noisy, non-stationary, and heterogeneous across assets and horizons, with low signal-to-noise ratios, regime shifts, and limited effective sample sizes for many instruments. These properties complicate generalization and challenge the stability of out-of-sample performance in dynamic markets.

The success of large language models (LLMs) has popularized a pre-training and fine-tuning paradigm for learning general-purpose representations that transfer across tasks and domains. This paradigm has inspired the emergence of time series foundation models (TSFMs)[1], large pre-trained architectures designed to learn universal temporal representations from vast and diverse time series corpora. TSFMs aim to deliver competitive performance on previously unseen datasets in a zero-shot or few-shot manner, while retaining strong in-domain accuracy after adaptation. Recent models illustrate two dominant design philosophies: discrete tokenization with autoregressive decoding versus continuous latent embeddings with regression-style objectives. Analogous to LLMs for text, the promise of TSFMs lies in compressing generic temporal regularities (e.g., seasonality, volatility clustering, long-memory features) into reusable representations that can be efficiently adapted to domain-specific tasks.

This paper presents the first comprehensive empirical study of TSFMs in global financial markets. Using a large-scale panel of daily excess returns spanning 34 years across 94 countries, we evaluate three TSFM regimes at the univariate level: (i) zero-shot inference with pre-trained weights, (ii) fine-tuning on financial data, and (iii) pre-training from scratch on financial time series. We test these approaches against a broad set of benchmarks, including linear models, ensemble models[2], and neural networks. Our analysis leverages one of the largest datasets ever used for financial forecasting, comprising approximately two billion observations that enable model training at the international level and testing across major markets, including the United States, Hong Kong, Taiwan, South Korea, Germany, the United Kingdom, India, and Australia. This extensive coverage supports a rigorous

[1] In Das et al. (2024), TimesFM is used as the abbreviation for time series foundation model. However, most of the recent literature adopts TSFM as the standard abbreviation for time series foundation model. Accordingly, in this paper, we use 'TSFM' as the generic term for time series foundation models, and reserve 'TimesFM' to refer to the specific model of Das et al. (2024).

[2] Throughout this study, ensemble models refer specifically to tree-based ensemble models.

evaluation of TSFMs and benchmark models, examining their robustness and generalizability across diverse market structures and institutional environments. We assess model performance from both statistical and economic perspectives, with statistical evaluation based on out-of-sample forecasting metrics under an expanding-window design, and economic evaluation translating forecasts into portfolio returns. By combining global data with more than three decades of daily observations, this study provides robust evidence on the potential and limitations of TSFMs for financial forecasting.

We begin by evaluating the performance of benchmark models in forecasting next-day excess returns using historical return information. This evaluation is crucial, as it establishes the foundation for assessing the predictive capacity of TSFMs. Nevertheless, such benchmarking is frequently overlooked in the TSFM literature, where model comparisons are typically performed on generic datasets that fail to capture the unique characteristics of financial time series. Across rolling estimation windows of 5, 21, 252, and 512 trading days and from 2001 to 2023, ensemble models (CatBoost, XGBoost, and LightGBM) consistently outperform linear benchmarks (OLS, Lasso, Ridge, Elastic Net, and principal component regression (PCR)) and neural network architectures across standard forecasting performance metrics. Specifically, when averaged across all window sizes, the linear regression model attains an out-of-sample R^2 of -0.47%, while the CatBoost model achieves -0.10% for all U.S. stocks. The direction accuracies are all just above 51% for the four windows. Also, small-capitalization firms exhibit greater predictability compared to large-capitalization firms. Portfolios constructed from ensemble-based forecasts also yield higher annualized returns and Sharpe ratios, alongside smaller drawdowns and more favorable higher-moment characteristics. Notably, CatBoost delivers the strongest risk-adjusted performance, outperforming both linear and neural network counterparts. The best-performing CatBoost model, in terms of the Sharpe ratio, achieves an annualized return of 46.50% and a Sharpe ratio of 6.79 when using a window size of 252. These results are derived from a long–short portfolio constructed using the model's daily predicted excess returns and rebalanced on a daily basis, without accounting for transaction costs.

Turning to TSFMs, our empirical analysis focuses on two widely adopted TSFMs: Chronos (Ansari et al., 2024) and TimesFM (Das et al., 2024). We begin by documenting that off-the-shelf pre-trained TSFMs perform weakly in zero-shot forecasting of daily excess returns, underperforming strong ensemble models such as CatBoost and LightGBM. For both Chronos and TimesFM families, their performance gradually improves as the larger model and wider window size are used. The Chronos (large) model with 512 past excess returns generates out-of-sample R^2 of -1.37% and directional accuracy just above 51%, while the TimesFM (500M) model attains R^2 of -2.80% and directional accuracy just below 50%. However, across all window sizes, off-the-shelf pre-trained TSFMs deliver markedly

lower R^2 than benchmarks. Also, the annualized returns of the long–short portfolios constructed using Chronos (large) and TimesFM (500M), based on a window size of 512, decline to 20.17% and −1.47%, respectively. Extending the analysis to twelve additional TSFM architectures yields generally similar results, reinforcing the robustness of these findings. Nonetheless, some evidence of generalization emerges among TSFMs pre-trained on larger-scale datasets. Next, fine-tuning of these pre-trained models on financial data yields limited improvements and fails to close the performance gap with benchmarks. Most fine-tuned TSFM performance deteriorates, except for Chronos (large). However, this improvement does not translate into economic gains. Furthermore, when focusing on goodness-of-fit, fine-tuning TSFMs on financial data once again does not fully close the performance gap relative to benchmark models.

Subsequently, we pre-train the TSFMs from scratch for each year using only the financial data available up to that point. This procedure ensures that the models are not exposed to information from future periods, thereby preventing look-ahead bias, an issue that may arise when employing off-the-shelf TSFMs. We find that TSFMs pre-trained from scratch achieve substantial gains in both predictive accuracy and portfolio performance. Considering the Chronos (small) model, its R^2 for window size 5 increases substantially to −3.18% from −77.07%, and for window size 512, it increases to −0.59% from −1.27%. However, even when pre-trained from scratch on financial data, TSFMs still remain less effective in terms of goodness-of-fit than benchmark models. In terms of annualized return and Sharpe ratio, the Chronos (small) model achieves 36.84% and 5.42, respectively, using a window size of 512, while the TimesFM (20M) model attains 30.36% and 3.66 under the same conditions. This highlights the importance of domain-specific pre-training and alignment with financial data. Moreover, when comparing different window sizes, the results again indicate that the performance of TSFM improves with longer windows, whereas the benchmark models perform relatively better over shorter windows. For instance, the TimesFM (20M) model attains −18.22% annual return predicting with only 5 past returns, while it attains 30.36% with a window size of 512. Furthermore, across all model classes, including benchmark, zero-shot, fine-tuned, and pre-trained TSFMs, the long leg of the portfolio consistently outperforms the short leg.

Expanding the training universe from U.S. to global markets yields mixed results for the benchmark models. The linear model gains an additional 0.43–0.60% R^2 when expanding the dataset, turning all R^2 positive. Other benchmark models (Lasso, Ridge, and NN) also improve quite remarkably, while the ensemble models (CatBoost, XGBoost, and LightGBM) deteriorate marginally. However, the direction accuracies and portfolio performance mostly weaken for all models. When combined with financial factors and synthetic data augmentation, TSFMs exhibit consistent improvements in

both statistical and economic outcomes, achieving higher accuracy and stronger risk-adjusted portfolio returns. Specifically, the Chronos (small) model achieves the highest directional accuracy of 51.74%, compared with 51.16% obtained by CatBoost using a window size of 512. The annualized return and Sharpe ratio change to 41.89% and 6.78, respectively, compared with 47.25% and 6.46 achieved by CatBoost under the same window size. Moreover, our analysis underscores the critical importance of hyperparameter tuning: with appropriate optimization, TSFMs are capable of outperforming benchmark models even without scaling the dataset. Finally, examining performance over time shows that both TSFMs and benchmark models experience gradual degradation in portfolio performance, reflecting evolving market dynamics and rising efficiency. However, the decline is markedly slower and less severe for TSFMs. We also extend the main empirical tests to seven major non-U.S. markets, where we observe broadly consistent results with those obtained in the U.S.

Taken together, the evidence shows that generic time series pre-training does not directly transfer to financial domains, and that finance-native pre-training and data scaling are essential for realizing the full potential of TSFMs in financial forecasting. In this sense, TSFMs represent a new class of models for financial forecasting, combining the scalability and adaptability of foundation architectures with domain-specific learning objectives. To advance future research and foster transparency and reproducibility, we publicly release our models as open-source resources.

1.1 Related Work

Our paper contributes most directly to research on return prediction for stocks, particularly on how historical returns can predict future performance. The momentum (Jegadeesh and Titman, 1993) and reversal effects (Jegadeesh, 1990) are among the most prominent, constructed by sorting stocks on past cumulative returns. Building signals from moving averages is another simple yet effective approach. For instance, Brock et al. (1992) find strong support for moving-average and trading-range break strategies. Neely et al. (2014) also show that technical indicators such as moving averages, momentum, and volatility measures possess predictive power for the equity risk premium. Moskowitz et al. (2012) document a time series momentum effect, showing that an asset's own past excess returns predict its future performance. Extending this evidence, Menkhoff et al. (2012) show that currency momentum delivers high abnormal returns unexplained by standard risk factors, while Asness et al. (2013) find that value and momentum premia are pervasive across global asset classes and linked to common risk sources. Barroso and Santa-Clara (2015) highlight that momentum strategies exhibit time-varying risk and occasional large crashes, but volatility scaling can improve risk-adjusted performance. Gupta and Kelly (2018) demonstrate that momentum also exists among

factors themselves, referred to as factor momentum, which generates strong abnormal returns and complements traditional momentum. Liu and Tsyvinski (2021) demonstrate that cryptocurrencies also exhibit strong momentum driven by market-specific factors such as investor attention. More recently, Ehsani and Linnainmaa (2022) show that momentum largely reflects factor timing: factor returns are positively autocorrelated, this autocorrelation concentrates in the leading principal components, and a factor-momentum strategy explains a substantial share of stock-level momentum.

Building on this literature, subsequent research has explored alternative approaches for capturing nonlinear and dynamic relationships in return predictability. Instead of relying on moving averages, non-parametric regression techniques have been applied to examine the predictive power of technical indicators constructed from price–volume information. For example, Lo et al. (2000) use kernel regressions to model nonlinear dependencies between past prices and future returns. With the availability of larger datasets and advances in computation, a new generation of ML and deep learning (DL) models now learn nonlinear interactions and temporal dependencies directly from raw financial time series, delivering strong out-of-sample forecasts when properly tuned. These studies typically employ a multivariate framework that incorporates a broader set of exogenous variables to enhance predictive accuracy. For instance, Gu et al. (2020) examine the predictive performance of various ML models including regularized linear regressions, tree-based models, and neural networks, while Leippold et al. (2022) conduct a similar analysis in the Chinese market. Extending this line of research, Chen et al. (2024) show that DL architectures can capture complex nonlinear relations and interactions in asset pricing. Li et al. (2025) demonstrate that ML models trained on a comprehensive universe of financial signals achieve economically meaningful out-of-sample performance only when features are carefully designed, and Kelly et al. (2025) propose new asset pricing models that embed Transformer architectures into the stochastic discount factor, enabling context-aware cross-asset information sharing and yielding substantial reductions in pricing errors and improvements in out-of-sample Sharpe ratios. Complementing these empirical advances, Kelly et al. (2024) provide a theoretical justification for the superior performance of complex models, proving that out-of-sample forecast accuracy and portfolio performance increase with model complexity when appropriate regularization is applied.[3] Overall,

[3]This claim has sparked debate. For example, Berk (2023) argues that the theoretical framework in Kelly et al. (2024) is too narrow to be of practical use in financial economics and that its assumptions are inconsistent with equilibrium asset pricing. Similarly, Buncic (2025) re-examines the empirical results and shows that the findings are largely driven by modeling choices, notably the zero-intercept restriction and the aggregation method. Nagel (2025) demonstrates that, in high-complexity ridgeless Random Fourier Features (RFF) settings with short training windows, the approach effectively reproduces a volatility-timed momentum strategy rather than uncovering genuine predictive signals. In addition, Cartea et al. (2025) develop a theoretical and empirical framework showing that when predictive features are measured with noise, increasing model complexity can degrade out-of-sample R^2 and portfolio Sharpe ratios, thereby highlighting a 'limited virtue of complexity.' In response to these critiques, Kelly and Malamud (2025) provide theoretical and empirical clarifications, showing that the main findings of Kelly et al. (2024) remain robust and extending the analysis to include limits to learning and the concept of ensemble complexity.

these studies show that ML models can successfully predict the cross-section of stock returns based on historical data, particularly for small-cap stocks, thereby challenging the Efficient Market Hypothesis (see also Martin and Nagel, 2022). We extend the existing literature by leveraging TSFMs to scale models beyond previously tested limits and to analyze how higher levels of parameterization influence asset pricing performance.

Another emerging trend in financial ML is the development of domain-specific language models tailored to financial text. One of the earliest and most influential examples is FinBERT (Huang et al., 2023), a pre-trained model specifically adapted for finance that enhances information extraction and sentiment analysis within accounting and financial documents. Building on this direction, Rahimikia and Drinkall (2024) develop domain-specific LLMs aimed at mitigating look-ahead bias and demonstrate that targeted pre-training at smaller model scales can match or even outperform large general-purpose models such as the LLaMA series in trading tasks. This line of research is further advanced by He et al. (2025), who introduce chronologically consistent language models trained on time-stamped data to ensure that only information available at each point in time is used. Their findings indicate that look-ahead bias can be reduced while maintaining strong performance in both language and financial prediction tasks. Together, these studies underscore the growing focus on temporally aware and domain-adapted LLMs, a consideration that is equally important for TSFMs, whose architectures are largely derived from LLM principles. Accordingly, ensuring temporal consistency and domain alignment in TSFMs represents another key focus of this study.

More recently, foundation and generative models have emerged as powerful approaches for time series forecasting. Foundation models are large Transformer-based architectures pre-trained on extensive and heterogeneous collections of time series data to learn generalizable temporal representations that can be transferred across different series and sampling frequencies with minimal fine-tuning. Inspired by LLMs such as InstructGPT (Ouyang et al., 2022), LLaMA (Touvron et al., 2023), and DeepSeek (Liu et al., 2024), these TSFMs adapt tokenization and embedding strategies originally developed for text to the continuous, multivariate, and irregularly sampled nature of temporal data. Representative examples include Chronos (Ansari et al., 2024), which applies discrete tokenization with a Transformer backbone, and TimesFM (Das et al., 2024), which employs continuous embeddings and a decoder-only design. Collectively, these architectures demonstrate strong cross-domain generalization and enable zero-shot and few-shot forecasting capabilities across diverse temporal domains.

In parallel, generative models such as generative adversarial networks (GANs) and diffusion models extend financial forecasting beyond point predictions to model the full conditional distribution of

returns. Notable examples include Quant GAN (Wiese et al., 2020), which generates realistic return paths while preserving key statistical properties of financial data; Fin-GAN (Vuletić et al., 2024), which incorporates economics-informed objectives for regime-aware scenario generation; Signature-Wasserstein GANs (Liao et al., 2024), which leverage path signatures to reduce the min-max game to supervised learning, enhancing the training stability; and FTS-Diffusion (Huang et al., 2024), which models scale-invariant temporal patterns and enhances predictive robustness through data augmentation. Moreover, TSFMs are inherently capable of simulating future trajectories and can thus be viewed as conditional generative models. Together, these developments mark a paradigm shift from traditional point estimation toward probabilistic forecasting. Despite these advances, empirical evaluation of TSFMs and generative approaches in financial contexts remains limited. This study addresses this gap by providing a comprehensive assessment of their applicability and performance in financial forecasting, systematically benchmarking zero-shot, fine-tuned, and from-scratch pre-training regimes of TSFMs across a large number of assets.

The remainder of this paper is organized as follows. Section 2 introduces the core concepts underlying TSFMs, detailing their architectures and training paradigms. Section 3 outlines the proposed methodology, including the modeling framework, evaluation metrics, and experimental setup. Section 4 describes the data sources, cleaning procedures, and computational resources used to conduct our experiments. Section 5 presents the main numerical findings, comparing benchmark models with TSFMs across various configurations and data volumes. Finally, Section 6 summarizes the key insights, discusses implications for future research, and concludes the study.

2 Time Series Foundation Models

We begin by introducing the fundamentals of LLMs in Section 2.1, which serve as the basis for TSFMs. We then present two representative TSFMs, namely Chronos (Ansari et al., 2024) and TimesFM (Das et al., 2024) in Section 2.2 and Section 2.3, respectively. We also describe ten additional TSFMs used in our numerical experiments in Appendix D. For a comprehensive overview, readers are referred to Liang et al. (2024), which categorizes TSFMs by their embedding design, architecture choice, pre-training objective, and adaptation strategy.

2.1 Large Language Models

LLMs are powerful DL models that generate text by predicting the next word in a sequence based on the context of preceding words. They operate on tokenized text inputs and produce text outputs, with

the Transformer architecture serving as their core building block. LLMs are typically trained using an unsupervised learning objective on massive corpora, allowing them to capture rich linguistic patterns. This enables them to perform a wide range of language tasks such as translation, summarization, question answering, and content generation. Due to their versatility and generalization capabilities, LLMs are often considered foundation models for natural language processing (NLP).

2.1.1 Tokenization

In language modeling, raw text data consists of sequences of words, which are typically represented as tokens. Tokenization refers to the process of segmenting and mapping an input string into a sequence of integer tokens $(z_1, \cdots, z_T)$, each drawn from a finite vocabulary $\mathcal{V}$. This process is often conducted by using subword (Kudo, 2018) or byte-pair encoding schemes (Sennrich et al., 2016). Tokens are then mapped to continuous vector embeddings via a learned embedding matrix. Positional encodings are added to preserve the order information that would otherwise be lost due to the permutation invariance of the Transformer architecture.

2.1.2 Network Architecture: Transformer

The Transformer (Vaswani et al., 2017) is a sequence-to-sequence model based on the self-attention mechanism. It maps an input sequence of embeddings to an output sequence of contextual representation. Sequence modeling is achieved through stacked layers of self-attention and position-wise feed-forward networks, each wrapped with residual connections and layer normalization. The core building block of the Transformer is the multi-head self-attention (MHSA) module, in which each attention head computes a weighted aggregation of input features based on their pairwise similarities. Formally,

$$\begin{aligned} \text{MHSA} : \mathbb{R}^{T\times d} &\rightarrow \mathbb{R}^{T\times \tilde{d}}; \\ X &\mapsto O := \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\, W^O, \end{aligned} \tag{1}$$

where d and $\tilde{d}$ are the feature dimension of the input and output sequence, respectively, and T is the time dimension. For head $i \in \{1, \ldots, h\}$:

$$Q_i = XW_i^Q, \quad K_i = XW_i^K, \quad V_i = XW_i^V, \tag{2}$$

where $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d \times d_k}$ are learnable projection matrices. The attention output for head i is given by:

$$\text{head}_i = \text{Softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_k}}\right) V_i, \tag{3}$$

where $V_i \in \mathbb{R}^{T \times d_k}$ and the Softmax operation is applied row-wise across the time dimension to yield normalized attention weights. The output O is obtained by concatenating all heads across the feature dimension and applying the learnable matrix $W^O \in \mathbb{R}^{(hd_k) \times \tilde{d}}$. The complete set of learnable parameters in the MHSA module is $\{W_i^Q, W_i^K, W_i^V\}_{i \in \{1, \cdots, h\}} \cup W^O$. Yu et al. (2025) analyze Transformers for time series data through the lens of rank structure, showing that time series embeddings tend to exhibit low-rank properties. This inherent structure makes attention layers more compact and computationally efficient, providing insight into why Transformers can be effectively compressed for time series forecasting.

2.1.3 Training and Inference

Given a sequence of tokens $z_{1:T}$ drawn from a vocabulary $\mathcal{V}$, an LLM is typically optimized to maximize the autoregressive log-likelihood:

$$\max_\theta \sum_{t=1}^{T} \log p_\theta \left(z_t | z_{<t}\right), \tag{4}$$

where $z_{<t} = (z_1, \cdots, z_{t-1})$ and p_θ denotes the conditional distribution of tokens, parametrized by θ. In practice, this objective is equivalent to minimizing the cross-entropy loss, which is optimized using stochastic gradient–based methods such as Adam (Kingma and Ba, 2015) or its variants.

At inference time, the model rolls out forecasts autoregressively. As a generative probabilistic model, the LLM simulates the next token based on the learned conditional distribution, appends it to the sequence, and repeats until the desired horizon is reached. Consequently, one can estimate various statistical properties of the predictive distribution, such as means, variances, or quantiles, by drawing multiple samples via Monte Carlo simulation.

2.2 Chronos

Chronos (Ansari et al., 2024), proposed by Amazon, is one of the most popular foundation models for time series. It closely follows the LLM architecture, differing primarily in the tokenization scheme, which is adapted from textual data to handle continuous-valued time series inputs. Therefore, it is possible to integrate a variety of LLM backbones within this framework, since the overall Transformer-based architecture remains largely unchanged.

Tokenization. Consider a time series $\mathbf{x}_{1:C+H} = [x_1, ..., x_{C+H}]$, where the first C steps are the input context used to predict the next H steps. To leverage the forecasting capacity of LLMs, Chronos transforms the continuous real-valued observations $\mathbf{x}$ into discrete tokens through a two-step process: (1) scaling and (2) quantization. Heterogeneous scaling across time series data makes model optimization challenging. Hence, one needs to map the time series values into a suitable range for quantization by mean scaling. A series is normalized by the mean of the absolute values from the historical context, as defined by the following transformation: $\tilde{x}$ is given by $\tilde{x} = x_i/s$ and $s = \frac{1}{C}\sum_{i=1}^{C} |x_i|$.

The scaled time series then needs to be mapped to a finite set of tokens, by means of quantization, in order to be processed by LLMs. The quantization function $q : \mathbb{R} \mapsto \{1, 2, ..., B\}$ assigns each value $\tilde{x}$ to a bin, given by:

$$q(\tilde{x}) = \begin{cases} 1 & \text{if } -\infty \leq \tilde{x} < b_1, \\ 2 & \text{if } b_1 \leq \tilde{x} < b_2, \\ \vdots & \\ B & \text{if } b_{B-1} \leq \tilde{x} < \infty, \end{cases} \tag{5}$$

where $(b_i)_i$ denote uniformly spaced bin edges. At inference time, each predicted token j is mapped to its corresponding bin center $d(j) = c_j$ through dequantization, and then rescaled by s to obtain the final forecast. The quantization–dequantization process constrains predictions within $[c_1, c_B]$, which corresponds to the range $[-15s, 15s]$ in Chronos's default configuration. The bin width is $30s/(B-1)$, implying that a large s reduces precision by grouping nearby values into the same token, whereas a small s risks producing values outside the representable range.

Framework. Similar to LLMs, Chronos adopts the cross-entropy loss function and a flexible Transformer backbone, which can be either encoder–decoder or decoder-only; in our experiments, we use the T5 architecture (Raffel et al., 2020). Analogous to language models that generate text autoregressively, Chronos predicts future time series values by recursively sampling from the conditional categorical distribution over tokenized observations. Each predicted token is then converted into a real-valued number through a dequantization and rescaling process tailored for time series data. Because the model outputs a full probability distribution at each step, it enables probabilistic forecasting and uncertainty quantification, which are particularly relevant in financial applications.

2.3 TimesFM

TimesFM (Das et al., 2024), proposed by Google, is another representative TSFM, differing from Chronos primarily in its continuous representation and training objective. Whereas Chronos discretizes time series data via scaling and quantization to produce tokenized sequences suitable for LLM-style autoregressive modeling, TimesFM represents time series directly in a continuous latent space and is optimized under a supervised regression loss. The model naturally extends to multivariate time series and retains a strong connection to LLMs through its Transformer-based architecture.

Embedding. Recall that $x_{1:T}$ is a d-dimensional time series of length T. Instead of tokenizing each scalar observation, TimesFM partitions the input into a sequence of patches, where each patch $\mathbf{p}_i \in \mathbb{R}^{L \times d}$ contains L consecutive time steps:

$$\mathbf{p}_i = [\mathbf{x}_{(i-1)L+1}, \mathbf{x}_{(i-1)L+2}, \ldots, \mathbf{x}_{iL}], \quad i = 1, \ldots, M := T/L. \tag{6}$$

For simplicity, in Equation (6), we assume that T is an integer multiple of L; otherwise, the sequence can be padded to satisfy this condition. This patching mechanism is analogous to the sliding-window approach commonly used in time series analysis, allowing each patch to serve as a 'token' that captures local temporal dependencies. Each patch defined on the non-overlapping time interval is encoded into a continuous embedding using a residual network f_ϕ, typically implemented as a multilayer perceptron (MLP):

$$\mathbf{e}_i = f_\phi(\mathbf{p}_i), \quad \mathbf{e}_i \in \mathbb{R}^h, \tag{7}$$

where h is the embedding dimension. To improve generalization across different context lengths, TimesFM applies a random masking strategy that probabilistically omits certain patches, ensuring the model encounters all possible prefix lengths during training. Note that the patch length of the input and output sequences need not be identical.

Framework. The resulting embeddings $[\mathbf{e}_1, \ldots, \mathbf{e}_M]$ are then passed through a stacked Transformer decoder with causal self-attention, whereby the model attends only to preceding embeddings in the sequence to predict an output in the future; this preserves the information flow in time series. The model outputs a continuous forecast embedding $\hat{\mathbf{p}}_{i+1}$, which is mapped back to the time domain through another residual network g_ψ:

$$\hat{\mathbf{p}}_{i+1} = g_\psi(\mathbf{h}_i), \tag{8}$$

where $\mathbf{h}_i$ denotes the hidden state of the Transformer at step i.

Unlike Chronos, which is trained with a cross-entropy loss over discrete tokens, TimesFM is optimized directly on the continuous prediction error using the mean squared error (MSE) loss between the observed output $\mathbf{p}_i$ and the model-estimated $\hat{\mathbf{p}}_i$ at the forecasting horizon. During inference, TimesFM supports multi-step forecasting through a rolling-window strategy, recursively feeding predicted patches back into the model to generate longer-horizon forecasts. This continuous formulation eliminates the need for quantization or dequantization, allowing TimesFM to learn smooth representations and capture fine-grained temporal variations without discretization artifacts.

3 Methodology

3.1 Problem Setup

Under the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, consider a financial market consisting of M assets. Let $S^{(i)} := \left(S_t^{(i)}\right)_{t\in\mathcal{T}}$ denote the price series of the i^{th} asset, where $\mathcal{T}$ is the range of the time index on a daily basis and $i \in \{1, \cdots, M\}$ is the asset index. Let $D_t^{(i)}$ denote the cash dividend (or other cash distribution) paid by asset i between time $t-1$ and t. The corresponding one-day return is defined by $r_t^{(i)} := \frac{S_t^{(i)}+D_t^{(i)}-S_{t-1}^{(i)}}{S_{t-1}^{(i)}}$. Next, we compute each firm's daily excess return by subtracting the daily risk-free rate from its total daily return. The daily excess return for firm i on day t is therefore defined as $r_t^{(i),ex} = r_t^{(i)} - r_{f,t}^{(d)}$, where $r_t^{(i)}$ is the one-day return, including dividends, and $r_{f,t}^{(d)}$ is the daily risk-free rate. We are interested in simulating the distribution of the next excess return of the i^{th} asset, given the information of the asset up to time t, i.e., $S_{1:t}^{(i)}$, or equivalently, the past excess return series, denoted by $r_{1:t}^{(i),ex}$.

We further assume that the conditional law of one-step excess returns is identical across assets, that is, $\mathbb{P}(r_{t+1}^{(i),ex}|r_{1:t}^{(i),ex} = r)$ does not depend on i. This assumption implies a homogeneous market structure where individual asset dynamics share a common recursive pattern. Under this setting, a single conditional generator can be trained using pooled data from all assets and then applied universally to simulate future returns for any asset. In practice, certain normalization of the return series might be applied to ensure that this assumption is valid.

Stochastic TSFM. Some TSFMs, such as Chronos, are conditional generative models capable of sampling future return paths given past observations. Let $G_\theta : \mathcal{Z} \times \mathbb{R}^C \to \mathbb{R}$ denote a conditional generative model parameterized by θ, where $\mathcal{Z}$ is the noise space and C is the length of the lagged time series. The aim of G_θ is to approximate the conditional distribution of the next excess return of

any arbitrary asset given the lagged values $r_{t-C+1:t}$, i.e.,

$$G_\theta(Z, r^{ex}_{t-C+1}, \ldots, r^{ex}_t) \approx \mathbb{P}(r^{ex}_{t+1}|\mathcal{F}_t), \tag{9}$$

where $\mathcal{F}_t$ denotes the sigma-algebra (information set) generated by the historical returns $r_{1:t}$. Once the generative model is trained, it can be used for various downstream tasks. For instance, one can estimate the conditional expectation of the next excess return via Monte Carlo sampling, i.e.,

$$\mathbb{E}[r^{ex}_{t+1}|\mathcal{F}_t] \approx \frac{1}{N}\sum_{n=1}^{N} G_\theta(Z_n, r^{ex}_{t-C+1}, \ldots, r^{ex}_t), \quad Z_n \overset{iid}{\sim} Z, \tag{10}$$

where N is the number of Monte Carlo samples and Z_n is independently drawn from the noise vector Z. Similarly, one can estimate the probability of an upward or downward movement of the next excess return from the empirical distribution of the Monte Carlo samples.

Deterministic TSFM. Deterministic TSFMs such as TimesFM directly output a point prediction for the next return. When no distributional information is provided, the sign of this point estimate is often used as a proxy for predicting the direction of the next movement. However, if the true conditional distribution is asymmetric, the sign of the conditional mean may not correctly indicate the most likely direction of return movement.

3.2 Time Series Foundation Models: Training and Inference

The TSFMs, described in Section 2, can be applied to the return-forecasting problem described above. Based on the dataset used for training and model initialization, the TSFMs can be further categorized into three types: (1) pre-trained models, (2) fine-tuned models, and (3) models pre-trained from scratch. For our above return prediction task, we have the financial time series data, denoted by $\mathcal{D}_{\text{fin}} = \{S^{(i)}\}_{i=1}^{N_{\text{fin}}}$.

3.2.1 Pre-Trained Model (Zero-Shot Inference)

Let $\mathcal{D}_{\text{pre}} = \{X^{(i)}\}_{i=1}^{N_{\text{pre}}}$ denote a large collection of heterogeneous time series datasets used for pre-training. This pre-training dataset is massive, encompassing time series with diverse statistical characteristics such as seasonality, autocorrelation, and volatility dynamics. Typically, $\mathcal{D}_{\text{pre}}$ has time series from various domains, and also includes a large amount of synthetic time series. The model parameters θ_{pre} are obtained by minimizing the loss function which is generic to time series of various kinds. The pre-trained model $G_{\theta_{\text{pre}}}$ thus learns domain-agnostic temporal structures such as seasonality, volatility clustering, and cross-asset dependencies. In zero-shot inference, $G_{\theta_{\text{pre}}}$ is directly applied to

a new dataset without parameter updates, leveraging its generalization ability to generate predictive distributions across unseen domains. In our case, we apply $G_{\theta_{\text{pre}}}$ to $\mathcal{D}_{\text{fin}}$ for generating the future return.

3.2.2 Fine-Tuned Model

Given a pre-trained initialization θ_{pre}, fine-tuning adapts the model to a specific downstream dataset (e.g., $\mathcal{D}_{\text{fin}}$ in our case) by solving:

$$\theta_{\text{fin}} = \arg\min_{\theta} \mathbb{E}_{(X,Y)\sim\mathcal{D}_{\text{fin}}} \mathcal{L}(G_\theta(X), Y)\,, \tag{11}$$

where $\mathcal{L}$ is a supervised loss tailored to the target objective (e.g., conditional return generation). The pre-trained parameters may serve as a strong prior, accelerating convergence and enhancing generalization when $\mathcal{D}_{\text{fin}}$ is relatively small. This transfer-learning setup balances domain-specific adaptation with the preservation of foundational temporal knowledge. In our case, we adopt the original loss function used in the TSFM with training pairs defined as $(X, Y) = (r^{ex}_{t-C+1:t}, r^{ex}_{t+1})$ for fine-tuning, and set all the model parameters trainable.

3.2.3 Models Pre-Trained from Scratch

Instead of using a pre-trained model for parameter initialization, model parameters are randomly initialized, $\theta_0 \sim \mathcal{P}_0$, and trained solely on $\mathcal{D}_{\text{fin}}$:

$$\theta_{\text{scratch}} = \arg\min_{\theta} \mathbb{E}_{(r^{ex}_{t-C+1:t+1})\sim\mathcal{D}_{\text{fin}}} \mathcal{L}\big(G_\theta(r^{ex}_{t-C+1:t}), r^{ex}_{t+1}\big)\,, \tag{12}$$

where $\mathcal{L}$ is the loss function of the TSFM. Pre-training from scratch typically requires larger datasets and greater computational resources, as the model must learn temporal dependencies without any prior knowledge. In our study, as the same loss function $\mathcal{L}$ is used for fine-tuning in Section 3.2.2, the main difference between Section 3.2.3 and Section 3.2.2 lies in the parameter initialization. Pre-training learns all parameters from scratch, whereas fine-tuning starts from a pre-trained model and, in some applications, may update only the final layers while freezing earlier ones.

3.3 Evaluation Metrics

To evaluate different aspects of generative models for return prediction, we consider a range of test metrics, grouped into two categories: (i) forecasting performance and (ii) portfolio performance.

3.3.1 Forecasting Performance Metrics

In our work, we focus on two forecasting performance metrics: (1) predicting the conditional mean of future returns, and (2) predicting the probability of return direction (up or down). For the first metric, we employ the coefficient of determination (R^2_{OOS}). In particular, our definition of R^2_{OOS} follows the out-of-sample metric used by Gu et al. (2020), defined as:

$$R^2_{\text{OOS}} = 1 - \frac{\sum_{i,t}(r^{(i),ex}_{t+1} - \hat{r}^{(i),ex}_{t+1})^2}{\sum_{i,t}(r^{(i),ex}_{t+1})^2}, \tag{13}$$

where $r^{(i),ex}_{t+1}$ denotes the realized excess return of asset i at time $t+1$, and $\hat{r}^{(i),ex}_{t+1}$ represents the corresponding predicted value. This version of R^2_{OOS} benchmarks forecasts against a naive prediction of zero.[4] Throughout this study, we report R^2_{OOS} in percentage terms. For the second group of metrics, we use classification accuracy (overall accuracy) and the macro-averaged F_1 score[5] as evaluation metrics for the binary classification of return direction. In addition to overall classification accuracy, we report upward accuracy and downward accuracy, which correspond respectively to cases where the realized future return is positive or negative. All directional accuracy measures are likewise reported in percentage terms.

3.3.2 Portfolio Performance Metrics

Following standard empirical asset pricing procedures for portfolio sorting, we translate the model's forecasts into an implementable long–short trading strategy. At the end of each trading day t, the model generates a one–day–ahead predicted excess return $\widehat{r}^{(i),ex}_{t+1}$ for each stock i in the investable universe. We then rank all stocks by $\widehat{r}^{(i),ex}_{t+1}$ and divide them into ten deciles. A zero-cost, equal-weighted long–short portfolio is constructed by going long the top-decile stocks and short the bottom-decile stocks. The portfolio is rebalanced daily.

The realized long–short portfolio return time series $\{r^{LS}_{t+1}\}$ provides a direct, economically interpretable measure of the model's cross-sectional ranking performance. Significant positive returns indicate that the model's predicted-return ranking successfully identifies near-term winners and losers.

[4] Gu et al. (2020) argue that using the historical mean as a baseline is inappropriate for individual stock returns, because historical mean excess returns are extremely noisy and can make even weak models appear to perform well. By contrast, benchmarking against zero avoids overstating predictive performance and provides a more stringent and economically meaningful test of out-of-sample forecasting accuracy.

[5] The macro-averaged F_1 score is defined as the unweighted mean of the class-specific F_1 scores, computed independently for each class. It jointly accounts for precision, the proportion of correctly predicted positive (upward or downward) excess return trends among all predicted positives, and recall, the proportion of correctly predicted positives among all actual positives. This macro-averaging approach ensures balanced evaluation of directional forecasting performance across both upward and downward excess return trends. Formally, macro-$F_1 = \frac{1}{C}\sum_{c=1}^{C} \frac{2\,\text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}$, where C denotes the number of classes (here, $C = 2$ for up and down excess return trends).

We report a comprehensive set of portfolio performance metrics that capture profitability, risk, and distributional characteristics. Specifically, the annualized return, standard deviation, and Sharpe ratio summarize mean performance and volatility-adjusted efficiency. The daily return (bps) provides a direct measure of the average daily economic magnitude of the long–short strategy. To assess downside risk, we include both the overall maximum drawdown (Max DD) and the largest single-day loss (Max DD (1-day)). Finally, the skewness and kurtosis statistics describe the asymmetry and tail risk of the portfolio return distribution. Annualized return, standard deviation, Max DD, and Max DD (1-day) are reported in percentage terms.

4 Data and Computational Resources

4.1 Data Source

We construct a comprehensive panel of daily firm-level excess returns spanning 1990–2023 across 94 countries.[6] The raw equity returns are adjusted for delisting effects, corporate actions, and market conventions to produce a clean, consistent measure of daily stock performance net of the local risk-free rate. Daily price and return information are obtained from two complementary sources. For the U.S., all equity data originate from the Center for Research in Security Prices (CRSP) database. For non-U.S. markets, we rely on Compustat global daily security files, which provide comparable firm identifiers, prices, and share counts. When a security is available in both CRSP and Compustat, the CRSP record is given priority because of its broader coverage of corporate actions, higher data quality, and inclusion of delisting returns.

Before computing excess returns, the daily data undergo a sequence of cleaning and harmonization procedures. Only listings on the primary trading venue of each country are retained to avoid duplicate records for cross-listed securities. Securities with nonpositive or missing prices are removed, and returns outside a range of $\pm 1000\%$ are treated as data errors. Observations belonging to the bottom five percent of the market capitalization distribution within each country-day are excluded to limit the influence of microcaps, which often display irregular trading behavior and excessive volatility.

When delisting information is available, the reported delisting return is incorporated on the final trading day of the firm. In cases where a delisting is recorded but the associated return is missing, a

[6] One methodological concern that may arise pertains to our choice of data frequency. While most prior empirical asset pricing studies employ monthly excess returns and emphasize multivariate return predictability using macroeconomic or firm-level variables, our analysis focuses on daily excess return forecasting. The use of daily data inherently limits the analysis to a univariate setting, as most predictive factors are unavailable or unreliable at this frequency. This study evaluates zero-shot and fine-tuning approaches and further examines the feasibility of pre-training TSFMs from scratch, which requires a large number of observations. Monthly data do not provide a sufficiently large sample for this purpose; therefore, we employ daily data to ensure the feasibility of the analysis.

Table 1: Cumulative Observations and Security Coverage

U.S.					
Year	**Observations**	**Securities**	**Year**	**Observations**	**Securities**
2000	86.82	35.76	2012	133.62	48.40
2001	90.77	36.51	2013	137.73	49.57
2002	94.62	37.28	2014	141.91	50.84
2003	98.40	38.05	2015	146.15	52.01
2004	102.21	39.13	2016	150.40	53.02
2005	106.06	40.15	2017	154.61	54.19
2006	109.91	41.35	2018	158.85	55.38
2007	113.84	42.89	2019	163.12	56.46
2008	117.82	43.78	2020	167.47	57.87
2009	121.71	44.53	2021	172.10	60.71
2010	125.61	45.54	2022	176.96	62.25
2011	129.55	47.01			
All Markets					
Year	**Observations**	**Securities**	**Year**	**Observations**	**Securities**
2000	132.96	67.40	2012	299.14	129.36
2001	143.05	73.79	2013	317.02	134.83
2002	153.45	77.35	2014	334.59	141.18
2003	163.99	80.78	2015	352.77	147.31
2004	174.79	84.55	2016	371.48	154.04
2005	186.16	89.13	2017	390.82	160.93
2006	200.20	94.61	2018	410.77	167.67
2007	215.11	101.68	2019	430.88	174.17
2008	231.03	106.50	2020	451.73	182.25
2009	247.17	110.70	2021	473.75	194.19
2010	263.82	116.18	2022	496.89	204.09
2011	280.96	122.32			

Note: This table presents the cumulative number of observations and unique securities for the U.S. in the top panel, and for all markets combined in the bottom panel. The years 2000 to 2022 represent individual years for which separate models are trained. Observation counts are reported in millions, and security counts in thousands.

return of -30% is assigned following the standard convention in the literature to mitigate survivorship bias. To maintain consistency between CRSP and Compustat records, daily returns from Compustat are winsorized using the CRSP return distribution as a benchmark. Specifically, Compustat returns above the 99.9^{th} percentile or below the 0.1^{st} percentile of the corresponding CRSP distribution for the same day are capped at those values. This ensures that extreme observations in international markets do not distort the overall return distribution. Following data cleaning, excess returns are computed as described in Section 3.1.

Table 1 presents the cumulative number of observations (excess returns) and unique securities for the U.S. in the top panel, and for all markets combined in the bottom panel.[7] The years 2000 to 2022 represent individual years for which separate models are trained. The training data begin in 1990. Observation counts are reported in millions, and security counts in thousands. The 2022 models employed the most extensive dataset, comprising a maximum of 176.96 million and 496.89 million

[7]The complete list of country and region codes is provided in Table A.1. In total, we compile global data from 94 countries and regions to construct a single dataset for model training for each year from 2000 to 2022.

observations for the U.S. and global data, respectively. The corresponding securities counts were 62,250 for the U.S. and 204,090 for all markets combined.

Figure A.1 reports the cumulative number of observations and unique securities for the U.S. market, while Figure A.2 presents the corresponding values for the combined global sample. The light-shaded region represents the sample period starting in 1990 (expanding window) used to train the predictive models, while the dark-shaded region denotes the period from 2001 to 2022, during which the predictive models are trained on a yearly basis. The data start in 1990 because of the lower quantity and quality of data prior to that year, especially for global markets.[8]

As another source of data, we use all factors from Jensen et al. (2023) (JKP factors), organized into thirteen conceptual clusters that reflect distinct economic mechanisms. The main clusters include 'Investment', 'Value', 'Low risk', and 'Quality', which capture firms' growth, valuation, risk, and profitability characteristics. Other clusters such as 'Seasonality', 'Profit growth', 'Leverage', 'Profitability', 'Momentum', 'Debt issuance', 'Accruals', 'Short-term reversal', and 'Size' represent additional aspects of firm behavior, including cyclical patterns, financial structure, and return dynamics. Together, these 153 monthly factors proxy for underlying economic risks and behavioral patterns that shape cross-sectional returns.

Table 2 presents the annual observations and security coverage of the JKP factors. The top panel reports data for the U.S., while the bottom panel aggregates data from all markets. As before, the observation counts are reported in millions, and security counts in thousands. The 2022 models employed the most extensive dataset, comprising a maximum of 456.62 million and 1434.05 million observations for the U.S. and global data, respectively. The corresponding securities counts were 33,180 for the U.S. and 102,820 for all markets combined. The larger number of observations results from generating up to 153 factors per security. Figure A.3 compares the cumulative number of observations and unique securities between the excess return data, the JKP data, and the combined dataset. The maximum number of observations for all data together (Total) recorded in 2022 corresponds to 1930.95 million observations and 135,990 securities.[9]

[8]The annual breakdown of cross-market excess returns is presented in Table A.2 and Table A.3, accompanied by the corresponding annual breakdown of cross-market securities shown in Table A.4 and Table A.5. The first two tables present the cumulative number of valid excess return observations by year and market, with values scaled in millions. The second two tables also present the number of securities by year and market, with values scaled in thousands. The numbers are presented cumulatively, showing the number of unique securities present up to and including each year. It is evident from all tables that in the early years, the number of available observations, particularly for global markets, is limited and in many cases close to zero. The sample size increases progressively over time as data coverage improves.

[9]The annual breakdown of JKP observations is presented in Table A.6 and Table A.7, accompanied by the corresponding annual breakdown of JKP securities shown in Table A.8 and Table A.9. The first two tables present the cumulative number of valid JKP observations by year and market, with all values scaled in millions. The second two tables present the number of securities used to construct the JKP factors, with values scaled in thousands. The higher number of observations and securities for JKP factors over time is clearly observable in all figures.

Table 2: Cumulative Observations and Security Coverage (JKP Factors)

U.S.					
Year	**Observations**	**Securities**	**Year**	**Observations**	**Securities**
2000	282.76	25.60	2012	384.81	28.99
2001	293.08	25.85	2013	391.92	29.30
2002	302.70	26.08	2014	399.11	29.70
2003	311.72	26.32	2015	406.39	30.00
2004	320.51	26.69	2016	413.51	30.22
2005	329.28	27.08	2017	420.51	30.51
2006	337.93	27.45	2018	427.48	30.83
2007	346.43	27.87	2019	434.45	31.11
2008	354.69	28.09	2020	441.40	31.63
2009	362.55	28.27	2021	448.76	32.84
2010	370.15	28.51	2022	456.62	33.18
2011	377.56	28.76			
All Markets					
Year	**Observations**	**Securities**	**Year**	**Observations**	**Securities**
2000	397.02	50.50	2012	879.34	81.53
2001	427.66	54.74	2013	929.82	83.53
2002	459.73	56.75	2014	981.88	85.52
2003	492.88	58.65	2015	1034.59	87.48
2004	527.88	60.64	2016	1087.90	89.17
2005	565.03	63.44	2017	1142.58	91.55
2006	604.63	66.49	2018	1198.29	93.55
2007	647.10	70.38	2019	1254.78	95.22
2008	691.35	72.50	2020	1311.98	97.30
2009	736.13	74.26	2021	1371.86	101.05
2010	782.30	76.57	2022	1434.05	102.82
2011	830.02	79.23			

Note: This table presents the cumulative number of observations and unique securities for the JKP factors, as defined in Jensen et al. (2023). The top panel reports data for the U.S., while the bottom panel aggregates data from all markets. The years 2000 to 2022 represent individual years for which separate models are trained. Observation counts are reported in millions, and security counts in thousands.

Based on these three datasets, we define three corresponding groups of data for model training. The first group contains only U.S. excess return data. The second extends this to include global excess return data, encompassing both U.S. and global firms. The third further augments the global data by incorporating the JKP factors. Throughout this study, we refer to these groups as U.S., global, and JKP-augmented, respectively.[10] Model estimation follows an expanding-window approach, with the first model (for year 2000) trained on data starting in 1990, and one additional model estimated for each subsequent year through 2022. Consequently, for each test conducted in this study, a total of 23 distinct training datasets are constructed. The dataset begins in 1990, as this year marks the point from which both U.S. and international markets exhibit sufficiently reliable and comprehensive data coverage.

4.2 Data Cleaning and Preprocessing

To ensure the integrity and comparability of the multi-asset return panel, and following the cleaning procedures described in Section 4.1, we implement a series of additional data cleaning and preprocessing steps to enhance data quality. These procedures serve to standardize the dataset prior to conducting the empirical analysis. To mitigate the influence of extreme values, all daily excess returns are winsorized symmetrically to the interval $[-1, 1]$:

$$\bar{r}_t^{(i),ex} = \min\big(1, \max(-1,\, r_t^{(i),ex})\big). \tag{14}$$

The threshold is time-invariant and keeps more data than the usual 99% quantile cut used in empirical research (Bali et al., 2016). This step curtails the leverage effect of outliers, which are common in multi-asset environments, and stabilizes the subsequent cross-sectional imputations. After cleaning, all firm-level series are merged into a unified annual panel,

$$\bar{r}_t^{ex} = \big(\bar{r}_t^{(1),ex}, \bar{r}_t^{(2),ex}, \ldots, \bar{r}_t^{(M_t),ex}\big), \tag{15}$$

with dates ordered in ascending sequence, where M_t denotes the number of assets available at time t and $i \in \{1, \ldots, M_t\}$ indexes assets. This ensures that all assets share a common calendar index, allowing for synchronized cross-sectional analysis at the daily level.

Missing excess returns are imputed within each country–day cross-section to preserve the institutional structure of national markets. For each trading date t, if all firms' returns from a country are

[10] The data volumes of the U.S., global, and JKP-augmented datasets are also reported in Figure A.4, measured in gigabytes (GB), reaching over 120 GB for the JKP-augmented data in 2022, which represents the largest dataset used.

Table 3: Summary of Annual Out-of-Sample Observations and Securities

Year	Observations	Securities	Year	Observations	Securities
2001	913,194	4,585	2013	735,737	3,308
2002	879,869	4,269	2014	749,785	3,346
2003	898,407	4,326	2015	728,878	3,285
2004	965,728	4,444	2016	722,053	3,349
2005	931,131	4,234	2017	731,544	3,360
2006	926,102	4,240	2018	721,802	3,271
2007	907,453	4,182	2019	697,442	3,233
2008	788,216	3,795	2020	700,507	3,363
2009	693,570	3,393	2021	783,312	3,583
2010	761,590	3,515	2022	716,029	3,398
2011	743,847	3,430	2023	738,169	3,705
2012	710,378	3,264	**Overall**	**18,144,743**	**10,171**

Note: This table reports the annual number of out-of-sample excess return observations and the count of unique securities for the U.S. market from 2001 to 2023. Excess returns are reported as raw observation counts, while the number of securities represents distinct securities included each year. The overall number of securities in the last row represents the count of unique securities across all years combined. The summary statistics are derived from CRSP data, with the sample filtered to include only ordinary common shares, excluding preferred stocks, funds, and other non-equity securities. The analysis is further restricted to U.S. common stocks traded on major exchanges (NYSE, AMEX, and NASDAQ) with a minimum share price of $5.

missing, the entire date is treated as missing (interpreted as a market-wide closure or holiday). Otherwise, following Gu et al. (2020), missing values within that country are replaced by the cross-sectional median of non-missing observations from that country on date t. Using the median instead of the mean enhances robustness to heavy-tailed or skewed return distributions, especially during volatile market conditions. The imputation is conducted independently across countries, respecting differences in trading calendars and market microstructures.

To avoid extrapolating excess returns beyond an asset's listing period, we identify the first and last valid trading dates for each asset, and any observations outside this interval are set to missing. This boundary masking ensures that the imputation procedure only affects internal gaps (occasional missing trades within the active period) and not structural absences such as pre-listing or post-delisting intervals. Also, dates for which all asset excess returns are missing are removed from the panel. These typically correspond to global holidays or systemic closures and thus contain no information relevant for the cross-sectional analysis. This operation increases computational efficiency without any information loss.

4.3 Sample and Model Sizes

To generate forecasts, each model is estimated using an expanding window that begins in 1990. Consequently, the out-of-sample evaluation period spans from 2001 to 2023. The summary of annual out-of-sample observations and securities is presented in Table 3. Two conditions are applied when

filtering the data. First, the sample includes only ordinary common shares, excluding preferred stocks, funds, and other non-equity securities. Second, low-priced stocks trading below $5 are removed to avoid distortions arising from illiquid or highly volatile penny stocks. The analysis is further restricted to firms listed on the major U.S. exchanges, including the NYSE, AMEX, and NASDAQ, to ensure data quality and consistency across actively traded and well-regulated markets. In total, all models, including benchmark models, are tested on over 18 million daily excess returns and approximately 10,000 U.S. securities. This long time span and large sample of excess returns and securities ensure reliable comparisons across all models.

Having defined the sample used in the evaluation, we now summarize the relative complexity of the forecasting models. To provide a clear picture of the differences in model size, Figure 1 presents a plot of model parameter counts (log-scaled on the x-axis) against model categories (y-axis). Bubble area is proportional to the true parameter count. Black bubbles denote benchmark models, while gray bubbles denote TSFMs. TSFMs include Chronos (tiny, mini, small, base, and large)[11], and TimesFM (with 8, 20, 200, and 500 million parameters). For models where the number of parameters depends on the chosen input window size, we report the average number of parameters. For benchmark models whose complexity depends on hyperparameter choices, we report the maximum possible number of parameters for illustration. The pronounced difference in scale between TSFMs and the benchmark models, which include conventional ML models, is evident. Reported TSFM sizes correspond to configurations employed in zero-shot inference, fine-tuning, and pre-training. TSFMs used for pre-training are highlighted with black borders. We employ all available model sizes of Chronos and TimesFM in both our zero-shot and fine-tuned experiments. However, due to computational constraints, we restrict our pre-training experiments with Chronos to the tiny, mini, and small configurations, and scale down TimesFM to versions containing approximately 8 million and 20 million parameters.[12]

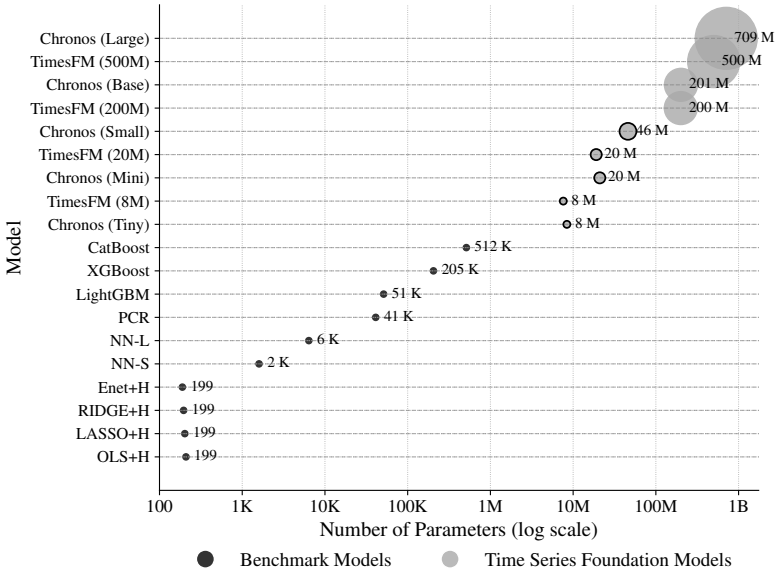
4.4 Computational Resources

Inference, fine-tuning, and particularly the pre-training of TSFMs are computationally intensive processes that necessitate access to large-scale, high-performance computing infrastructure. All model development and experimental procedures in this study were executed on servers equipped with graph-

[11] Chronos variants contain approximately 8M, 20M, 46M, 200M, and 710M parameters for the tiny, mini, small, base, and large models, respectively.

[12] TimesFM (200M) refers to the model released under the TimesFM 1 version, while TimesFM (500M) corresponds to the model released under TimesFM 2. Our scaled-down TimesFM variants are derived from the TimesFM 2. Relative to the original TimesFM 2 model configuration, which specifies 50 Transformer layers, 16 attention heads, 16 key–value heads, a head dimension of 80, a hidden size of 1280, and an intermediate size of 5120, we proportionally reduced these architectural hyperparameters to construct smaller variants. The 20M model employs 9 Transformer layers, 6 attention heads, 6 key–value heads, a head dimension of 72, a hidden size of 432, and an intermediate size of 1248, whereas the 8M model uses 7 layers, 4 attention heads, 4 key–value heads, a head dimension of 66, a hidden size of 264, and an intermediate size of 1024. All other settings remain consistent with the original TimesFM 2 specification.

Figure 1: Comparative Model Complexity



Note: This figure presents a plot of model parameter counts (log-scaled on the x-axis) against model categories (y-axis). Bubble area is proportional to the true parameter count. Black bubbles denote benchmark models, while gray bubbles denote time series foundation models (TSFMs). Benchmark models include linear (OLS+H, LASSO+H, RIDGE+H, Elastic Net+H, and PCR), ensemble (XGBoost, CatBoost, and LightGBM), and neural network (NN-S and NN-L) models. 'H' indicates that the model is estimated using the Huber loss. TSFMs include Chronos (tiny, mini, small, base, and large), and TimesFM (with 8, 20, 200, and 500 million parameters). For models where the number of parameters depends on the chosen input window size, we report the average number of parameters. For benchmark models whose complexity depends on hyperparameter choices, we report the maximum possible number of parameters for illustration. Reported TSFM sizes correspond to configurations employed in zero-shot inference, fine-tuning, and pre-training. TSFMs used for pre-training are highlighted with black borders.

ics processing units (GPUs). In total, approximately 50,000 GPU hours were utilized across all stages. This includes the pre-training of 396 Chronos models and 264 TimesFM models as part of this study. A substantial portion of these computational resources was also allocated to large-scale inference experiments, including extensive zero-shot evaluations across a diverse range of TSFMs. Benchmark models, which rely primarily on central processing unit (CPU)-based computation, also required substantial resources. Although precise usage was not recorded due to variation across servers, we estimate that roughly 25,000 CPU hours were consumed for training and testing these benchmark models.

The TSFM experiments were executed on multi-node servers equipped with NVIDIA GH200 Grace Hopper GPUs, each integrating a Grace CPU with an H100 Tensor Core GPU. Each GPU featured 96 GB of high-bandwidth memory, supported by large shared CPU memory and high-speed interconnects that enabled distributed and parallel computation. This configuration facilitated the efficient scaling of large Transformer architectures and ensured stable throughput during extended training cycles.[13] To ensure experimental consistency, random number generators (RNGs) were initialized at the beginning of all runs.

[13] A detailed account of the computational infrastructure and its underlying technical aspects is provided in McIntosh-Smith et al. (2024). Part of this research was also conducted using N8 Bede, which provides a broadly similar computational infrastructure.

5 Numerical Results

As a preliminary step, we evaluate the performance of several benchmark models to identify the best-performing one, which serves as the primary baseline for comparing TSFMs. This step ensures that subsequent comparisons are both rigorous and realistic, moving beyond the general benchmarks commonly used in the TSFM literature toward a more specialized set of models that better reflect practical financial forecasting performance. Benchmark models include linear (OLS, Lasso, Ridge, Elastic Net, and PCR), ensemble (XGBoost, CatBoost, and LightGBM), and neural network (NN-S and NN-L)[14] models. The selection of these models is motivated by several considerations. First, they are widely employed in the existing literature and represent the main categories of models commonly applied in financial forecasting tasks. Second, these models are computationally feasible to train given the large sample size utilized in this study. Finally, the set includes several models, particularly ensemble models, that are frequently recognized as among the best-performing techniques in forecasting applications.[15]

To generate forecasts, a distinct benchmark model is estimated for each year from 2000 to 2022.[16] Each model is estimated using an expanding window that begins in 1990, and the model estimated for year t is employed to produce forecasts for year $t+1$. Consequently, the out-of-sample evaluation period spans from 2001 to 2023. For the TSFMs, when pre-training or fine-tuning is performed, the same expanding-window procedure is applied to maintain consistency and avoid look-ahead bias. This approach yields 23 distinct models, each corresponding to one out-of-sample forecasting year. In contrast, for zero-shot experiments using publicly available pre-trained TSFMs, a single model is applied across all out-of-sample periods.[17] All models are also evaluated with window sizes of 5, 21, 252, and 512 trading days to examine how the amount of historical information available to each model influences its performance. The window sizes of 5, 21, and 252 trading days correspond approximately to one week, one month, and one year of past excess returns, respectively. A window size of 512 trading days is additionally included, as it is a commonly used maximum input length in

[14] NN-S refers to a single-hidden-layer neural network with 8 hidden units, while NN-L denotes a similar architecture comprising 32 hidden units. Further details are provided in Appendix C.

[15] A detailed description of the benchmark models is provided in Appendix C. Also, Table C.1 presents the hyperparameter settings for the benchmark models. The set of models is drawn from Gu et al. (2020) and Leippold et al. (2022), subject to the additional requirement that they can be trained efficiently on the large-scale dataset employed in this study. Also, hyperparameter tuning is performed only for the first year (2000), and the selected hyperparameters are subsequently applied to the remaining years. This procedure ensures computational feasibility while enabling a fair comparison across different classes of models. The key hyperparameters that were optimized are reported under 'Tuned', while those held constant throughout the analysis are reported under 'Fixed'.

[16] Throughout this study, we use the terms 'pre-training' and 'fine-tuning' when referring to TSFMs, and the terms 'training' or 'estimation' when referring to other model classes.

[17] This evaluation framework enables us to assess the generalization ability of publicly available TSFMs. However, depending on the data used to pre-train these models, look-ahead bias may arise. To mitigate this concern, we develop proprietary pre-trained models that explicitly exclude any such overlap, while retaining the zero-shot and fine-tuned results from publicly available TSFMs for comparison.

many TSFMs. Finally, our work focuses on univariate return forecasting, benchmarking all TSFM models on a univariate time series setting, although some recently proposed TSFMs are capable of handling multivariate data.

We present our empirical findings in Section 5.1, Section 5.2, and Section 5.3. In Section 5.1, both the benchmark models and the TSFMs are trained exclusively on U.S. data. Section 5.2 broadens the scope to a global context by training all models, including both benchmarks and TSFMs, on an extensive dataset encompassing 94 countries. Finally, Section 5.3 evaluates the out-of-sample performance of the benchmark models and TSFMs across seven major international markets, thereby extending our analysis beyond the U.S. market.[18]

5.1 Results with U.S. Data

The initial set of results, presented in Section 5.1.1, reports the performance of benchmark models. Moving to the TSFMs, we present numerical results from three distinct experiments. Following Section 3.2.1, the first experiment employs pre-trained models released by their respective authors and applies them directly in a zero-shot setting to forecast excess returns, as shown in Section 5.1.2. The second experiment, described in Section 3.2.2 and reported in Section 5.1.3, fine-tunes these models and provides the corresponding results. Section 5.1.4 extends the analysis by pre-training the models from scratch and presenting the resulting performance. Finally, Section 5.1.5 provides a brief overview of the training time comparison across models. For all models, only U.S. data is used. Also, unless explicitly stated otherwise, all hyperparameters are kept consistent with those proposed by the original authors.

5.1.1 Benchmark Results

Table 4 reports the forecasting performance results for the benchmark models. This table presents each metric as a set of three values, ordered from top to bottom: full sample, top 25% of firms by market capitalization (large-cap), and bottom 25% (small-cap) for various predictive models across different window sizes (5, 21, 252, and 512 trading days). Metrics are first computed separately for each calendar year using all stock-date observations within that year. The reported values represent the average of these yearly statistics. Metrics include out-of-sample R^2 (R^2_{OOS}), overall directional accuracy, upward and downward classification accuracy, and macro-averaged F1 score. Benchmark models include linear (OLS, Lasso, Ridge, Elastic Net[19], and PCR), ensemble (XGBoost, CatBoost, and LightGBM), and

[18] All pre-trained models from scratch are available through our portal at `FinText.ai` and the Hugging Face repository at `https://huggingface.co/FinText`.

[19] Throughout this study, the abbreviation Enet is employed to represent the Elastic Net model.

neural network (NN-S and NN-L) models. 'H' indicates that the model is estimated using the Huber loss. 'Overall Acc.' denotes overall directional accuracy, 'Up Acc.' and 'Down Acc.' represent the model's accuracy in predicting upward and downward excess returns respectively, and 'F1' refers to the macro-averaged F1 score.

The results show that ensemble models, specifically XGBoost, CatBoost, and LightGBM, generally achieve superior forecasting performance across window sizes compared with other models. These ensemble models exhibit higher R^2_{OOS} values and overall accuracy. For the largest window size (512), CatBoost attains an R^2_{OOS} of -0.03%, an overall prediction accuracy of 51.16%, and an F1 score of 0.49. The only model that performs better in terms of directional accuracy is OLS; however, its performance in terms of R^2_{OOS} is lower than that of the ensemble models. For the best-performing models, there is also a general trend indicating that these models exhibit better forecasting performance for small-cap stocks compared to large-cap stocks. For the CatBoost model, averaged across all window lengths, the out-of-sample R^2 for small-cap stocks is 0.51%, compared with -0.37% for large-cap stocks. The corresponding overall accuracy values are 52.29% and 50.76%, respectively, while the F1 scores are 0.52 and 0.47. A similar pattern, indicating stronger predictive performance for small-cap stocks, is generally observed across the other models as well. This finding is consistent with the results reported in Gu et al. (2020), Leippold et al. (2022), and Kelly et al. (2025). Also, neural network models (NN-S and NN-L) produce mixed outcomes. NN-L performs moderately better than NN-S with longer windows (except for 512 days), particularly in terms of directional accuracy, but still trails behind the ensemble models. Overall, the results highlight that ensemble models provide the most reliable and robust predictive performance. Across different window sizes, model performance generally improves with longer estimation windows, indicating that incorporating more historical information enhances predictive performance.[20]

[20]We also report in Table B.1 the results of the modified Diebold–Mariano (DM) tests, following the approach of Gu et al. (2020), which assess the statistical significance of out-of-sample forecasting performance differences among benchmark models. A positive and statistically significant DM statistic indicates that the model in the column outperforms the model in the corresponding row. Consistent with the benchmark performance results, XGBoost, CatBoost, and LightGBM exhibit significantly superior forecasting performance compared to traditional linear models such as OLS, Lasso, Ridge, Elastic Net, and PCR across most window sizes. Neural network models (NN-S and NN-L) show mixed results, occasionally outperforming linear models but generally lagging behind the ensemble models. Among the ensemble models, CatBoost and XGBoost most frequently deliver statistically significant improvements over competing models. When also comparing performance across different window sizes for the same model, the results indicate that longer estimation windows (252 and 512 days) generally lead to improved forecasting performance, particularly for nonlinear and ensemble models, while shorter windows (5 and 21 days) yield more volatile and less reliable results.

Table 4: Benchmark Models - Forecasting Performance

Model	**OLS+H**				**LASSO+H**				**RIDGE+H**				**Enet+H**				**PCR**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
R^2_{OOS}	-0.40	-0.41	-0.48	-0.57	-2.89	-1.42	-1.65	-1.81	-1.31	-1.35	-1.65	-1.79	-1.36	-1.36	-1.31	-1.09	-1.34	-1.24	-1.50	-2.70
	-0.88	-0.88	-0.96	-1.07	-3.75	-2.12	-2.40	-2.58	-1.96	-2.00	-2.37	-2.51	-2.04	-2.04	-2.01	-1.75	-1.85	-1.73	-2.02	-3.56
	0.31	0.30	0.26	0.19	-1.53	-0.36	-0.53	-0.67	-0.27	-0.31	-0.54	-0.67	-0.30	-0.29	-0.26	-0.13	-0.64	-0.59	-0.79	-1.54
Overall Acc.	51.92	51.93	51.86	51.82	51.34	51.24	51.15	51.18	51.24	51.27	51.27	51.32	51.22	51.28	51.15	51.12	50.99	50.92	50.90	51.04
	50.59	50.62	50.67	50.69	51.09	51.18	51.15	51.20	51.20	51.14	51.28	51.21	51.18	51.21	51.21	51.18	51.00	50.96	50.97	50.77
	54.67	54.65	54.40	54.27	52.38	51.89	51.71	51.75	51.93	52.11	51.97	52.16	51.94	52.00	51.66	51.58	51.38	51.31	51.29	52.01
Up Acc.	45.56	45.92	47.50	47.77	63.70	67.15	68.03	67.94	67.34	66.17	67.26	65.96	66.87	67.31	69.12	69.33	66.11	65.95	66.00	62.33
	42.63	43.01	44.79	45.21	63.90	68.11	69.20	69.09	68.29	66.89	68.42	66.79	67.80	68.26	70.46	70.63	68.00	68.13	68.32	63.48
	49.24	49.68	51.23	51.38	66.91	69.53	70.33	70.27	70.04	68.96	69.36	68.19	69.47	69.89	71.38	71.43	66.83	66.34	66.22	63.29
Down Acc.	58.19	57.87	56.19	55.81	39.38	35.61	34.74	34.82	35.56	36.78	35.66	37.04	36.02	35.53	33.68	33.32	36.22	36.20	36.15	39.99
	59.01	58.70	56.92	56.49	37.80	33.40	32.41	32.49	33.36	34.78	33.29	34.92	33.86	33.31	31.19	30.85	33.19	32.96	32.82	37.45
	59.48	59.05	57.23	56.81	39.63	36.16	35.26	35.36	35.94	37.20	36.60	37.97	36.50	36.04	34.22	33.96	37.67	37.95	38.06	41.97
F1	0.52	0.52	0.52	0.52	0.51	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.49	0.50	0.50	0.50	0.50
	0.50	0.50	0.51	0.51	0.50	0.49	0.49	0.49	0.49	0.50	0.49	0.50	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.50
	0.54	0.54	0.54	0.54	0.52	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.50	0.50	0.51	0.51	0.51	0.52

Model	**XGBoost**				**CatBoost**				**LightGBM**				**NN-S**				**NN-L**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
R^2_{OOS}	-0.23	0.01	-0.09	-0.12	-0.25	-0.05	-0.03	-0.03	-0.22	-0.01	-0.08	-0.13	-3.10	-3.35	-1.91	-2.35	-3.36	-2.31	-2.29	-2.58
	-0.44	-0.28	-0.33	-0.32	-0.50	-0.38	-0.32	-0.28	-0.43	-0.27	-0.31	-0.34	-4.48	-5.12	-2.54	-2.79	-5.13	-3.05	-2.82	-3.40
	0.38	0.64	0.51	0.47	0.39	0.67	0.64	0.60	0.37	0.60	0.52	0.47	-1.95	-2.08	-0.97	-1.38	-2.04	-1.39	-1.36	-1.60
Overall Acc.	51.23	51.23	51.27	51.34	50.94	51.12	51.08	51.16	51.17	51.18	51.24	51.38	50.94	50.97	50.96	51.32	51.25	51.28	51.04	50.85
	50.90	51.01	51.15	51.22	50.70	50.74	50.80	50.95	50.95	51.01	51.13	51.20	50.52	50.54	50.81	50.98	50.48	51.04	50.35	50.30
	52.52	52.44	52.31	52.46	51.96	52.48	52.33	52.39	52.34	52.30	52.34	52.59	51.94	52.06	51.74	52.33	52.79	52.24	52.34	52.05
Up Acc.	68.63	69.44	70.53	69.42	68.96	67.82	66.87	69.14	69.61	70.05	70.05	68.14	48.28	46.43	62.42	56.89	46.58	59.45	53.18	49.67
	71.26	73.96	76.27	74.88	72.05	72.97	71.06	74.39	72.42	74.80	75.86	73.46	48.02	45.96	62.93	56.89	46.08	59.20	51.97	48.62
	67.39	66.84	67.79	66.86	68.00	65.11	64.65	66.07	68.76	67.93	67.49	66.00	49.13	47.59	63.69	58.32	47.64	60.17	55.50	51.40
Down Acc.	33.97	33.15	32.25	33.36	33.14	34.60	35.59	33.37	32.95	32.53	32.73	34.73	53.19	55.28	39.64	45.35	55.32	42.51	49.16	52.30
	29.26	26.65	24.64	26.09	28.11	27.25	29.59	26.24	28.21	25.86	25.10	27.61	52.60	54.98	37.84	44.22	54.81	41.45	49.15	52.32
	38.91	39.27	38.28	39.22	37.42	41.01	41.24	39.87	37.41	38.10	38.67	40.28	54.21	55.94	41.04	46.46	56.49	44.37	49.53	52.93
F1	0.49	0.49	0.49	0.49	0.49	0.49	0.50	0.49	0.49	0.49	0.49	0.49	0.43	0.43	0.47	0.48	0.44	0.45	0.48	0.47
	0.47	0.46	0.46	0.46	0.47	0.47	0.48	0.47	0.47	0.47	0.47	0.47	0.41	0.41	0.46	0.46	0.42	0.44	0.46	0.45
	0.51	0.51	0.51	0.51	0.51	0.52	0.52	0.51	0.51	0.51	0.51	0.51	0.44	0.45	0.49	0.50	0.46	0.47	0.50	0.49

Note: This table presents each metric as a set of three values, ordered from top to bottom: full sample, top 25% of firms by market capitalization (large-cap), and bottom 25% (small-cap) for various predictive models across different window sizes (5, 21, 252, and 512 trading days). Metrics are first computed separately for each calendar year using all stock-date observations within that year. The reported values represent the average of these yearly statistics. Metrics include out-of-sample R^2 (R^2_{OOS}), overall directional accuracy, upward and downward classification accuracy, and macro-averaged F1 score. Benchmark models include linear (OLS, Lasso, Ridge, Elastic Net, and PCR), ensemble (XGBoost, CatBoost, and LightGBM), and neural network (NN-S and NN-L) models. 'Overall Acc.' denotes overall directional accuracy, 'Up Acc.' and 'Down Acc.' represent the model's accuracy in predicting upward and downward excess returns respectively, and 'F1' refers to the macro-averaged F1 score. 'H' indicates that the model is estimated using the Huber loss.

Table 5: Benchmark Models - Portfolio Performance

Model	OLS+H				LASSO+H				RIDGE+H				Enet+H				PCR			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
	45.32	45.71	45.91	45.34	44.83	43.67	43.84	43.78	44.93	44.48	44.79	44.30	44.29	44.31	44.30	43.83	30.17	29.50	28.59	33.39
Annualized Return	29.15	29.40	29.34	28.86	28.82	27.97	28.10	28.06	28.86	28.80	28.48	28.24	28.53	28.49	28.48	28.11	19.33	19.23	18.31	21.64
	16.17	16.31	16.57	16.48	16.01	15.70	15.73	15.72	16.07	15.68	16.30	16.06	15.76	15.82	15.82	15.72	10.83	10.27	10.28	11.75
	9.32	9.34	9.34	9.31	9.25	8.99	9.00	9.06	9.22	9.24	9.36	9.35	9.18	9.17	9.14	9.01	8.36	8.24	8.13	8.47
Standard Deviation	13.89	13.92	13.90	13.89	13.84	13.66	13.66	13.68	13.84	13.80	13.77	13.79	13.78	13.78	13.76	13.67	13.11	13.04	12.99	13.17
	11.85	11.83	11.84	11.86	11.87	11.88	11.88	11.89	11.86	11.90	12.01	11.98	11.89	11.89	11.88	11.88	12.13	12.07	12.00	12.04
	4.86	4.89	4.92	4.87	4.84	4.86	4.87	4.83	4.88	4.81	4.79	4.74	4.83	4.83	4.85	4.86	3.61	3.58	3.52	3.94
Sharpe Ratio	2.10	2.11	2.11	2.08	2.08	2.05	2.06	2.05	2.09	2.09	2.07	2.05	2.07	2.07	2.07	2.06	1.47	1.47	1.41	1.64
	1.37	1.38	1.40	1.39	1.35	1.32	1.32	1.32	1.36	1.32	1.36	1.34	1.33	1.33	1.33	1.32	0.89	0.85	0.86	0.98
	17.98	18.14	18.22	17.99	17.79	17.33	17.40	17.37	17.83	17.65	17.77	17.58	17.58	17.58	17.58	17.39	11.97	11.71	11.34	13.25
Daily Return (bps)	11.57	11.67	11.64	11.45	11.44	11.10	11.15	11.13	11.45	11.43	11.30	11.21	11.32	11.30	11.30	11.15	7.67	7.63	7.27	8.59
	6.42	6.47	6.57	6.54	6.35	6.23	6.24	6.24	6.38	6.22	6.47	6.37	6.25	6.28	6.28	6.24	4.30	4.08	4.08	4.66
	17.59	17.62	17.61	19.07	17.03	14.54	14.54	14.98	17.17	17.24	17.36	18.06	17.01	16.86	15.99	14.67	15.86	16.28	16.27	14.85
Max DD	32.60	32.98	32.88	33.14	32.36	31.40	31.40	31.57	32.44	32.31	32.15	32.16	32.21	32.17	32.02	31.43	27.12	27.59	27.46	28.61
	29.03	28.45	27.57	27.15	29.50	29.02	29.02	30.69	29.68	29.43	25.99	26.31	30.58	30.50	30.76	29.02	31.73	31.92	33.28	35.12
	5.60	5.66	5.68	5.68	5.59	5.06	5.06	5.14	5.59	5.57	5.50	5.39	5.53	5.49	5.42	5.06	3.24	2.94	3.65	4.21
Max DD (1-day)	8.74	8.79	8.77	8.76	8.75	8.40	8.40	8.44	8.77	8.75	8.64	8.50	8.68	8.69	8.62	8.40	7.45	7.16	7.59	7.85
	5.11	5.08	4.94	4.95	5.25	5.48	5.48	5.47	5.23	5.31	5.19	5.30	5.33	5.29	5.33	5.48	4.64	4.88	4.92	6.13
	1.30	1.31	1.26	1.29	1.26	1.12	1.13	1.10	1.21	1.21	1.11	1.12	1.20	1.19	1.16	1.13	0.87	0.87	0.78	0.81
Skew	-0.09	-0.08	-0.10	-0.09	-0.11	-0.08	-0.07	-0.08	-0.11	-0.08	-0.10	-0.09	-0.10	-0.10	-0.09	-0.07	-0.01	-0.08	-0.10	-0.18
	0.54	0.55	0.55	0.56	0.51	0.45	0.46	0.45	0.50	0.49	0.45	0.44	0.47	0.47	0.46	0.46	0.42	0.38	0.41	0.31
	17.33	17.74	17.62	17.67	16.73	15.15	15.15	15.15	16.68	16.65	16.26	16.21	16.53	16.40	16.22	15.09	12.51	12.43	11.19	12.39
Kurt	12.04	12.24	12.14	12.13	11.86	11.08	11.07	11.18	11.87	11.84	11.63	11.57	11.72	11.73	11.66	11.07	10.11	9.80	10.04	10.50
	6.41	6.37	6.40	6.34	6.19	6.01	6.06	6.06	6.15	6.26	6.20	6.19	6.13	6.12	6.06	6.06	6.15	6.12	5.83	6.15

Model	XGBoost				CatBoost				LightGBM				NN-S				NN-L			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
	44.47	46.86	47.69	47.40	42.54	46.52	46.50	47.25	43.61	45.93	46.49	46.52	40.59	42.62	40.15	40.11	44.04	42.97	40.84	38.82
Annualized Return	27.16	28.86	30.13	30.06	26.58	28.26	29.00	29.22	26.74	28.52	29.53	29.68	27.33	27.45	25.45	25.53	28.37	27.66	26.12	24.87
	17.31	18.00	17.56	17.34	15.95	18.26	17.50	18.03	16.87	17.41	16.96	16.83	13.26	15.17	14.70	14.58	15.67	15.31	14.73	13.95
	7.68	7.17	8.14	8.30	7.40	6.86	6.85	7.31	7.72	7.51	8.22	8.40	9.29	9.19	8.97	8.93	9.22	9.29	8.84	8.87
Standard Deviation	14.15	13.90	13.94	13.98	14.14	13.88	13.71	13.81	14.23	13.99	14.05	14.10	14.00	13.85	13.62	13.71	13.84	13.83	13.70	13.57
	12.05	12.42	12.53	12.53	12.09	12.37	12.31	12.31	12.07	12.42	12.56	12.52	11.76	11.96	11.91	11.84	11.91	11.89	11.84	11.90
	5.79	6.53	5.86	5.71	5.74	6.78	6.79	6.46	5.65	6.12	5.66	5.54	4.37	4.64	4.48	4.49	4.78	4.63	4.62	4.38
Sharpe Ratio	1.92	2.08	2.16	2.15	1.88	2.04	2.12	2.12	1.88	2.04	2.10	2.11	1.95	1.98	1.87	1.86	2.05	2.00	1.91	1.83
	1.44	1.45	1.40	1.38	1.32	1.48	1.42	1.46	1.40	1.40	1.35	1.34	1.13	1.27	1.23	1.23	1.32	1.29	1.24	1.17
	17.65	18.59	18.92	18.81	16.88	18.46	18.45	18.75	17.31	18.23	18.45	18.46	16.11	16.91	15.93	15.92	17.48	17.05	16.21	15.41
Daily Return (bps)	10.78	11.45	11.96	11.93	10.55	11.22	11.51	11.59	10.61	11.32	11.72	11.78	10.84	10.89	10.10	10.13	11.26	10.97	10.36	9.87
	6.87	7.14	6.97	6.88	6.33	7.25	6.94	7.15	6.70	6.91	6.73	6.68	5.26	6.02	5.83	5.78	6.22	6.08	5.84	5.54
	12.21	12.61	15.51	16.15	12.08	13.10	13.62	14.34	13.94	16.08	16.72	17.04	18.91	18.50	16.58	16.98	17.01	18.53	14.48	17.29
Max DD	32.08	32.09	33.46	33.65	32.51	32.85	31.86	32.62	33.33	33.87	33.95	33.72	33.36	33.38	32.27	32.70	32.39	33.23	31.99	31.66
	29.82	29.85	30.91	29.99	30.31	28.43	30.20	27.85	30.67	32.56	31.17	29.89	30.49	30.76	31.78	31.25	30.53	29.63	30.79	29.22
	4.97	3.41	4.36	4.56	4.86	3.75	4.15	4.93	5.10	4.26	4.96	5.17	5.62	5.47	4.80	5.15	5.51	5.61	5.49	5.29
Max DD (1-day)	8.56	8.06	8.23	8.19	8.28	8.21	8.19	8.19	8.59	8.33	8.28	8.32	8.81	8.54	8.21	8.37	8.67	8.80	8.67	8.32
	5.51	4.74	5.63	5.12	5.56	4.49	5.03	5.27	5.66	5.36	5.56	5.52	6.31	5.53	4.98	5.28	5.58	5.63	5.18	5.19
	1.81	2.10	1.57	1.48	2.24	2.19	2.25	1.72	1.80	1.74	1.58	1.56	1.31	1.31	1.20	1.67	1.24	1.37	1.27	1.36
Skew	-0.05	0.01	-0.01	-0.03	-0.01	0.00	0.02	-0.05	-0.08	-0.06	-0.04	-0.04	-0.12	-0.15	-0.09	-0.00	-0.11	-0.11	-0.14	-0.13
	0.35	0.36	0.33	0.36	0.35	0.38	0.38	0.36	0.33	0.35	0.32	0.32	0.48	0.51	0.51	0.52	0.52	0.53	0.45	0.49
	26.84	28.34	21.08	19.97	33.34	31.81	32.72	26.79	27.33	25.88	22.98	22.46	17.34	16.69	15.43	20.70	16.54	17.36	17.32	17.86
Kurt	11.44	11.51	12.13	11.98	11.62	12.31	12.53	12.56	11.37	12.17	12.26	12.20	11.96	11.61	11.69	12.29	11.45	11.68	11.98	11.78
	6.15	5.21	5.70	5.64	6.17	5.16	5.40	5.53	6.36	5.81	5.94	5.84	7.07	6.62	6.16	6.23	6.76	6.69	6.04	6.51

Note: This table reports average yearly portfolio performance metrics across different rolling window sizes (5, 21, 252, and 512 trading days) for each model. Each cell displays three values from top to bottom: long–short portfolio, long-only leg, and short-only leg. Metrics include annualized return, standard deviation, Sharpe ratio, daily return (in basis points), maximum drawdown (Max DD), one-day maximum drawdown (Max DD (1-day)), skewness, and kurtosis of portfolio returns. Benchmark models include linear (OLS, Lasso, Ridge, Elastic Net, and PCR), ensemble (XGBoost, CatBoost, and LightGBM), and neural network (NN-S and NN-L) models. Portfolios are formed using decile sorting based on model forecasts, with equal weighting across stocks. 'H' indicates that the model is estimated using the Huber loss.

Detailed results on portfolio performance can be found in Table 5. This table reports average yearly portfolio performance metrics across different rolling window sizes for each model. Each cell displays three values from top to bottom: long–short portfolio, long-only leg, and short-only leg. Metrics include annualized return, standard deviation, Sharpe ratio, daily return (in basis points), Max DD, Max DD (1-day), skewness, and kurtosis of portfolio returns. Consistent with the results shown in Table 4, the portfolio performance metrics indicate that ensemble models outperform both linear and neural network models. Across all rolling window sizes, ensemble models, particularly CatBoost, followed by XGBoost and LightGBM, achieve the highest annualized returns, Sharpe ratios, and daily returns, showing that stronger predictive accuracy translates into higher portfolio profitability. These models also exhibit moderate standard deviations, smaller maximum drawdowns, and more favorable higher-moment characteristics, including positive skewness and lower kurtosis, which suggest improved tail-risk management and more stable return distributions. In contrast, linear models deliver consistent but moderate performance across metrics, while neural networks display higher volatility, deeper drawdowns, and lower Sharpe ratios. The highest Sharpe ratio of 6.79 is achieved by CatBoost, exceeding both the linear benchmarks (mean Sharpe ratio ≈ 4.61) and neural network models (mean Sharpe ratio ≈ 4.55). The other ensemble models achieve maximum Sharpe ratios of 6.53 for XGBoost and 6.12 for LightGBM. Also, across all benchmark models and performance metrics, the long leg generally outperforms the short leg. This pattern holds across window sizes and risk-adjusted measures. This finding aligns with the results documented in Gu et al. (2020), Chen et al. (2024), and Leippold et al. (2022). Overall, the metrics in Table 5 confirm that ensemble models, especially CatBoost, provide the best balance between profitability, risk control, and return stability.[21] Accordingly, CatBoost is adopted as the benchmark model for the remainder of this study.

5.1.2 Zero-Shot Results

The central premise of TSFMs is that they can deliver competitive zero-shot performance across a wide range of tasks and frequencies, paralleling the success of LLMs, which have demonstrated strong zero-shot learning capabilities across diverse domains. TSFMs are designed to generalize to new tasks without task-specific fine-tuning; however, whether this promise holds in complex, real-world settings remains an open question. We evaluate this claim by employing TSFMs released by their respective authors and assessing their zero-shot performance, with particular attention to whether their claimed

[21] Table B.2 reports the annualized average returns of decile spread portfolios across models and window sizes. Decile returns generally rise from low to high, with a few small non-monotonic steps. Ensemble models show the steepest and most consistent gradients and typically the largest H–L spreads. Linear models also sort positively but with flatter gradients, although PCR is notably weaker than the others. Neural networks deliver smaller spreads than ensembles and most linear models, although they occasionally overlap in some window sizes.

generalization ability extends to the highly specialized and demanding task of daily excess return forecasting. We begin our analysis by examining two TSFMs, Chronos (Ansari et al., 2024) and TimesFM (Das et al., 2024), which are widely recognized as pioneering contributions to this domain. Their architectures have provided the basis for numerous subsequent model developments.

The zero-shot forecasting performance results in Table 6 show that TSFMs, including Chronos and TimesFM, perform substantially worse than the benchmark model (CatBoost). Across all model sizes and window lengths, TSFMs produce markedly negative R^2_{OOS} values, often extremely negative for smaller Chronos models and both TimesFM versions. The Chronos (small) variant attains the best performance among the TSFMs, with an R^2_{OOS} of -1.27%, which remains substantially below the benchmark models' best value of -0.03% for the full sample. The two TimesFM variants exhibit even greater instability, yielding highly negative R^2_{OOS} values across all windows, further highlighting the severe zero-shot underperformance of TSFMs relative to the benchmark. Between TimesFM 1 and TimesFM 2, particularly with respect to R^2_{OOS}, TimesFM 1 exhibits substantially weaker performance. Although larger Chronos variants improve with longer windows, still, for many model configurations their directional accuracy remains around or below 50%. TSFMs also exhibit asymmetric forecasting behavior, occasionally overpredicting either upward or downward movements, a tendency that is also observed in most of the benchmark models. TimesFM models display even more unstable and erratic performance, with highly negative R^2_{OOS} values and below 50% directional accuracy. The weakness of TSFMs is particularly evident at shorter window sizes, where these models consistently deliver the most severe drops in predictive performance. Although their results improve with longer windows, this recovery remains limited, suggesting that TSFMs are more capable of leveraging longer historical contexts than shorter-term information. Also, results are mixed between small and large stocks, with performance varying across models and window sizes. Nonetheless, the best-performing models broadly mirror the benchmark's small-stock advantage documented in Section 5.1.1.

Table 6: Zero-Shot TSFMs - Forecasting Performance

Model	**Benchmark**				**Chronos (Tiny)**				**Chronos (Mini)**				**Chronos (Small)**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
R^2_{OOS}	-0.25	-0.05	-0.03	-0.03	-75.87	-8.34	-8.34	-1.77	-43.14	-11.60	-11.60	-0.75	-77.07	-13.04	-13.04	-1.27
	-0.50	-0.38	-0.32	-0.28	-68.38	-8.27	-8.27	-2.50	-40.16	-9.68	-9.68	-1.13	-52.28	-8.01	-8.01	-1.34
	0.39	0.67	0.64	0.60	-85.21	-8.87	-8.87	-1.15	-47.34	-13.96	-13.96	-0.40	-104.56	-18.68	-18.68	-1.16
Overall Acc.	50.94	51.12	51.08	51.16	48.39	49.31	49.31	50.46	48.60	49.32	49.32	50.54	48.54	49.50	49.50	50.99
	50.70	50.74	50.80	50.95	49.41	49.88	49.88	50.86	49.36	49.68	49.68	50.68	49.22	49.46	49.46	50.25
	51.96	52.48	52.33	52.39	46.26	48.35	48.35	50.60	46.95	48.75	48.75	51.06	46.98	49.53	49.53	52.62
Up Acc.	68.96	67.82	66.87	69.14	54.53	59.51	59.51	78.02	51.67	54.39	54.39	68.43	47.69	44.87	44.87	28.51
	72.05	72.97	71.06	74.39	55.72	62.19	62.19	82.42	53.54	58.34	58.34	74.41	49.71	49.24	49.24	41.28
	68.00	65.11	64.65	66.07	52.81	56.75	56.75	70.99	48.62	48.76	48.76	60.84	44.32	38.35	38.35	15.37
Down Acc.	33.14	34.60	35.59	33.37	42.32	39.23	39.23	23.25	45.50	44.21	44.21	32.87	49.28	53.82	53.82	72.60
	28.11	27.25	29.59	26.24	42.71	36.80	36.80	17.55	44.86	40.39	40.39	25.56	48.56	49.35	49.35	58.95
	37.42	41.01	41.24	39.87	40.37	40.84	40.84	32.21	45.39	48.64	48.64	42.22	49.27	59.38	59.38	85.77
F1	0.49	0.49	0.50	0.49	0.48	0.49	0.49	0.46	0.49	0.49	0.49	0.49	0.48	0.49	0.49	0.48
	0.47	0.47	0.48	0.47	0.49	0.49	0.49	0.44	0.49	0.49	0.49	0.47	0.49	0.49	0.49	0.49
	0.51	0.52	0.52	0.51	0.46	0.48	0.48	0.49	0.47	0.49	0.49	0.51	0.47	0.49	0.49	0.44

Model	**Chronos (Base)**				**Chronos (Large)**				**TimesFM 1 (200M)**				**TimesFM 2 (500M)**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
R^2_{OOS}	-57.59	-21.47	-21.47	-1.35	-46.86	-23.84	-23.84	-1.37	-1469.95	-5004.66	-2.3e4	-6.3e4	-27.96	-11.87	-3.86	-2.80
	-38.80	-13.21	-13.21	-1.60	-35.50	-12.58	-12.58	-1.03	-610.92	-517.03	-2.1e4	-2.7e4	-27.27	-10.90	-3.79	-2.90
	-78.52	-30.27	-30.27	-1.22	-60.60	-35.63	-35.63	-1.70	-3499.35	-9809.08	-3.1e4	-1.2e5	-29.93	-13.36	-3.96	-2.64
Overall Acc.	48.59	49.19	49.19	50.80	48.70	49.13	49.13	51.01	49.55	50.00	49.87	49.60	48.34	48.70	49.59	49.82
	49.42	49.61	49.61	49.86	49.58	49.91	49.91	50.61	50.24	49.86	49.62	49.74	49.25	49.44	49.69	50.01
	46.84	48.45	48.45	52.61	46.94	47.85	47.85	52.46	48.53	50.15	50.24	49.36	46.34	47.18	49.72	49.94
Up Acc.	53.94	52.98	52.98	24.55	59.70	64.13	64.13	49.60	73.64	49.94	45.46	54.59	50.14	52.74	52.16	56.11
	55.68	56.87	56.87	30.55	61.30	66.96	66.96	59.93	73.80	49.14	44.42	53.29	52.65	56.01	56.30	61.13
	51.04	47.24	47.24	21.48	57.24	60.71	60.71	37.29	73.55	51.27	47.35	56.82	46.21	48.61	47.40	50.53
Down Acc.	43.27	45.31	45.31	76.18	37.85	34.35	34.35	52.04	26.02	50.01	54.08	44.72	46.46	44.56	46.81	43.44
	42.75	41.76	41.76	69.42	37.20	31.87	31.87	40.19	25.68	50.51	54.86	46.00	45.49	42.28	42.34	38.02
	43.02	49.33	49.33	80.34	37.70	36.35	36.35	66.00	26.26	49.15	52.79	42.75	46.38	45.81	51.66	49.31
F1	0.48	0.49	0.49	0.47	0.48	0.48	0.48	0.50	0.47	0.50	0.50	0.49	0.48	0.48	0.49	0.49
	0.49	0.49	0.49	0.47	0.49	0.48	0.48	0.49	0.47	0.50	0.49	0.49	0.49	0.49	0.49	0.49
	0.47	0.48	0.48	0.47	0.47	0.47	0.47	0.51	0.46	0.50	0.50	0.49	0.46	0.47	0.49	0.50

Note: This table presents each metric as a set of three values, ordered from top to bottom: full sample, top 25% of firms by market capitalization (large-cap), and bottom 25% (small-cap) for various predictive models across different window sizes (5, 21, 252, and 512 trading days). The benchmark model is CatBoost, the best-performing model among the benchmarks. The time series foundation models (TSFMs) include Chronos (tiny, mini, small, base, and large) and TimesFM (version 1 with 200 million and version 2 with 500 million parameters). Zero-shot inference is performed using the pre-trained models released by the respective authors. Metrics are first computed separately for each calendar year using all stock-date observations within that year. The reported values represent the average of these yearly statistics. Metrics include out-of-sample R^2 (R^2_{OOS}), overall directional accuracy, upward and downward classification accuracy, and macro-averaged F1 score. 'Overall Acc.' denotes overall directional accuracy, 'Up Acc.' and 'Down Acc.' represent the model's accuracy in predicting upward and downward excess returns respectively, and 'F1' refers to the macro-averaged F1 score. Numbers exceeding four digits are expressed in scientific notation for clarity.

Table 7: Zero-Shot TSFMs - Portfolio Performance

Model	Benchmark				Chronos (Tiny)				Chronos (Mini)				Chronos (Small)			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
	42.54	46.52	46.50	47.25	-37.07	-17.31	9.15	14.25	-32.85	-15.50	20.17	21.09	-36.73	-14.36	6.53	7.64
Annualized Return	26.58	28.26	29.00	29.22	-12.51	-3.68	9.42	11.20	-9.23	-1.13	14.44	15.40	-12.20	-1.27	7.63	8.98
	15.95	18.26	17.50	18.03	-24.56	-13.63	-0.27	3.05	-23.62	-14.37	5.73	5.69	-24.53	-13.10	-1.09	-1.34
	7.40	6.86	6.85	7.31	8.11	6.07	5.42	5.28	8.00	6.69	5.44	5.57	8.13	6.79	7.56	8.14
Standard Deviation	14.14	13.88	13.71	13.81	11.79	12.00	12.69	12.86	11.91	12.05	12.50	12.51	11.75	11.62	11.48	10.63
	12.09	12.37	12.31	12.31	13.34	13.52	13.58	13.45	13.73	13.75	13.74	13.52	13.70	14.46	15.53	15.62
	5.74	6.78	6.79	6.46	-4.57	-2.85	1.69	2.70	-4.10	-2.32	3.71	3.79	-4.52	-2.12	0.86	0.94
Sharpe Ratio	1.88	2.04	2.12	2.12	-1.06	-0.31	0.74	0.87	-0.77	-0.09	1.16	1.23	-1.04	-0.11	0.66	0.84
	1.32	1.48	1.42	1.46	-1.84	-1.01	-0.02	0.23	-1.72	-1.05	0.42	0.42	-1.79	-0.91	-0.07	-0.09
	16.88	18.46	18.45	18.75	-14.71	-6.87	3.63	5.66	-13.03	-6.15	8.00	8.37	-14.57	-5.70	2.59	3.03
Daily Return (bps)	10.55	11.22	11.51	11.59	-4.97	-1.46	3.74	4.45	-3.66	-0.45	5.73	6.11	-4.84	-0.50	3.03	3.56
	6.33	7.25	6.94	7.15	-9.74	-5.41	-0.11	1.21	-9.37	-5.70	2.27	2.26	-9.73	-5.20	-0.43	-0.53
	12.08	13.10	13.62	14.34	99.98	98.16	19.64	12.82	99.95	97.25	10.57	12.40	99.98	96.42	23.19	24.63
Max DD	32.51	32.85	31.86	32.62	95.91	69.10	35.16	34.74	91.15	55.02	27.87	29.08	95.53	54.45	35.18	29.40
	30.31	28.43	30.20	27.85	99.70	96.45	51.28	46.49	99.64	97.06	46.91	46.08	99.71	96.12	64.16	62.50
	4.86	3.75	4.15	4.93	5.34	4.91	5.00	5.45	12.18	4.67	4.89	12.24	12.11	4.98	4.29	4.31
Max DD (1-day)	8.28	8.21	8.19	8.19	6.31	6.85	6.96	6.98	7.23	6.55	6.85	6.64	6.78	6.82	7.06	6.64
	5.56	4.49	5.03	5.27	7.84	7.68	7.60	7.40	11.65	7.99	7.87	11.65	11.65	7.81	7.25	6.64
	2.24	2.19	2.25	1.72	-0.77	-1.11	-1.49	-1.36	-2.97	-0.81	-1.05	-7.78	-2.77	-1.12	-0.66	-0.50
Skew	-0.01	0.00	0.02	-0.05	-0.41	-0.42	-0.46	-0.44	-0.44	-0.41	-0.43	-0.31	-0.47	-0.52	-0.53	-0.45
	0.35	0.38	0.38	0.36	0.15	0.12	0.06	0.12	-0.29	0.17	0.12	-0.27	-0.26	0.03	0.04	0.10
	33.34	31.81	32.72	26.79	14.18	20.52	19.51	20.58	67.69	18.09	17.60	266.68	61.70	15.74	7.27	7.09
Kurt	11.62	12.31	12.53	12.56	5.66	6.45	5.47	5.74	6.34	5.08	5.24	5.64	5.99	6.18	6.17	7.73
	6.17	5.16	5.40	5.53	11.65	9.70	9.22	9.81	15.73	10.44	8.75	13.62	15.40	8.06	5.93	5.32

Model	Chronos (Base)				Chronos (Large)				TimesFM 1 (200M)				TimesFM 2 (500M)			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
	-31.94	-16.69	7.37	14.35	-30.59	-17.40	9.89	20.17	-6.83	-1.83	-1.40	-4.85	-39.62	-33.33	-7.61	-1.47
Annualized Return	-10.04	-2.99	8.21	13.50	-8.95	-3.05	8.82	13.13	-1.42	4.39	5.00	3.06	-13.04	-11.11	1.02	4.73
	-21.90	-13.70	-0.84	0.85	-21.64	-14.35	1.07	7.04	-5.41	-6.22	-6.41	-7.91	-26.58	-22.22	-8.64	-6.19
	7.65	6.74	8.31	8.52	7.66	6.90	7.62	6.91	6.40	9.48	5.06	5.33	9.52	9.23	8.55	7.95
Standard Deviation	12.07	11.94	10.58	10.10	12.20	11.90	10.96	10.72	12.98	13.18	11.06	11.26	11.79	11.57	11.36	11.57
	13.61	14.00	15.63	15.63	13.61	13.71	14.97	14.27	11.37	14.28	12.88	12.63	14.32	14.53	14.89	14.86
	-4.18	-2.48	0.89	1.68	-3.99	-2.52	1.30	2.92	-1.07	-0.19	-0.28	-0.91	-4.16	-3.61	-0.89	-0.18
Sharpe Ratio	-0.83	-0.25	0.78	1.34	-0.73	-0.26	0.80	1.22	-0.11	0.33	0.45	0.27	-1.11	-0.96	0.09	0.41
	-1.61	-0.98	-0.05	0.05	-1.59	-1.05	0.07	0.49	-0.48	-0.44	-0.50	-0.63	-1.86	-1.53	-0.58	-0.42
	-12.68	-6.62	2.92	5.69	-12.14	-6.91	3.92	8.00	-2.71	-0.73	-0.56	-1.93	-15.72	-13.23	-3.02	-0.58
Daily Return (bps)	-3.98	-1.19	3.26	5.36	-3.55	-1.21	3.50	5.21	-0.56	1.74	1.99	1.22	-5.18	-4.41	0.41	1.88
	-8.69	-5.44	-0.33	0.34	-8.59	-5.69	0.43	2.79	-2.15	-2.47	-2.54	-3.14	-10.55	-8.82	-3.43	-2.46
	99.94	97.90	23.95	19.72	99.92	98.22	19.87	15.80	81.06	59.97	32.50	70.13	99.99	99.96	84.86	46.40
Max DD	92.83	64.37	31.04	27.80	90.86	64.45	38.92	28.55	55.51	41.87	35.81	41.90	96.33	94.25	54.11	45.57
	99.46	96.52	65.64	53.41	99.42	96.99	53.10	45.85	76.29	84.46	81.81	87.24	99.82	99.50	90.14	84.32
	4.99	4.98	3.93	5.97	12.08	4.60	5.05	4.31	3.62	6.64	3.84	4.83	7.12	7.88	11.93	12.19
Max DD (1-day)	6.99	6.44	6.18	6.01	6.97	6.39	5.63	5.42	6.92	7.88	5.59	5.66	7.85	6.71	6.29	6.28
	8.34	8.17	7.42	6.25	11.65	7.72	7.53	6.21	6.02	7.79	7.15	7.86	9.23	9.22	11.65	11.65
	-0.79	-0.97	-0.57	-0.54	-3.18	-1.03	-0.84	-0.72	-0.08	-0.43	-0.34	-0.14	-1.51	-1.68	-3.27	-3.91
Skew	-0.43	-0.48	-0.52	-0.48	-0.41	-0.50	-0.55	-0.43	-0.30	-0.14	-0.43	-0.37	-0.59	-0.65	-0.66	-0.56
	0.12	0.08	0.02	0.08	-0.30	0.12	0.08	0.15	0.37	-0.10	0.21	0.14	0.02	0.01	-0.30	-0.29
	13.79	15.52	6.44	6.74	74.25	16.83	11.78	9.52	8.62	11.36	17.94	35.83	19.07	21.97	59.69	81.67
Kurt	6.12	5.80	7.22	9.55	6.04	5.61	5.34	5.51	6.50	9.73	5.81	6.21	6.41	5.95	5.33	5.03
	10.77	9.53	5.53	4.58	15.77	10.66	7.17	6.42	8.65	8.73	8.65	9.81	12.69	12.29	14.20	14.08

Note: This table reports average yearly portfolio performance metrics across different rolling window sizes (5, 21, 252, and 512 trading days) for each model. The benchmark model is CatBoost, the best-performing model among the benchmarks. The time series foundation models (TSFMs) include Chronos (tiny, mini, small, base, and large) and TimesFM (version 1 with 200 million and version 2 with 500 million parameters). Zero-shot inference is performed using the pre-trained models released by the respective authors. Each cell displays three values from top to bottom: long–short portfolio, long-only leg, and short-only leg. Metrics include annualized return, standard deviation, Sharpe ratio, daily return (in basis points), maximum drawdown (Max DD), one-day maximum drawdown (Max DD (1-day)), skewness, and kurtosis of portfolio returns. Portfolios are formed using decile sorting based on model forecasts, with equal weighting across stocks.

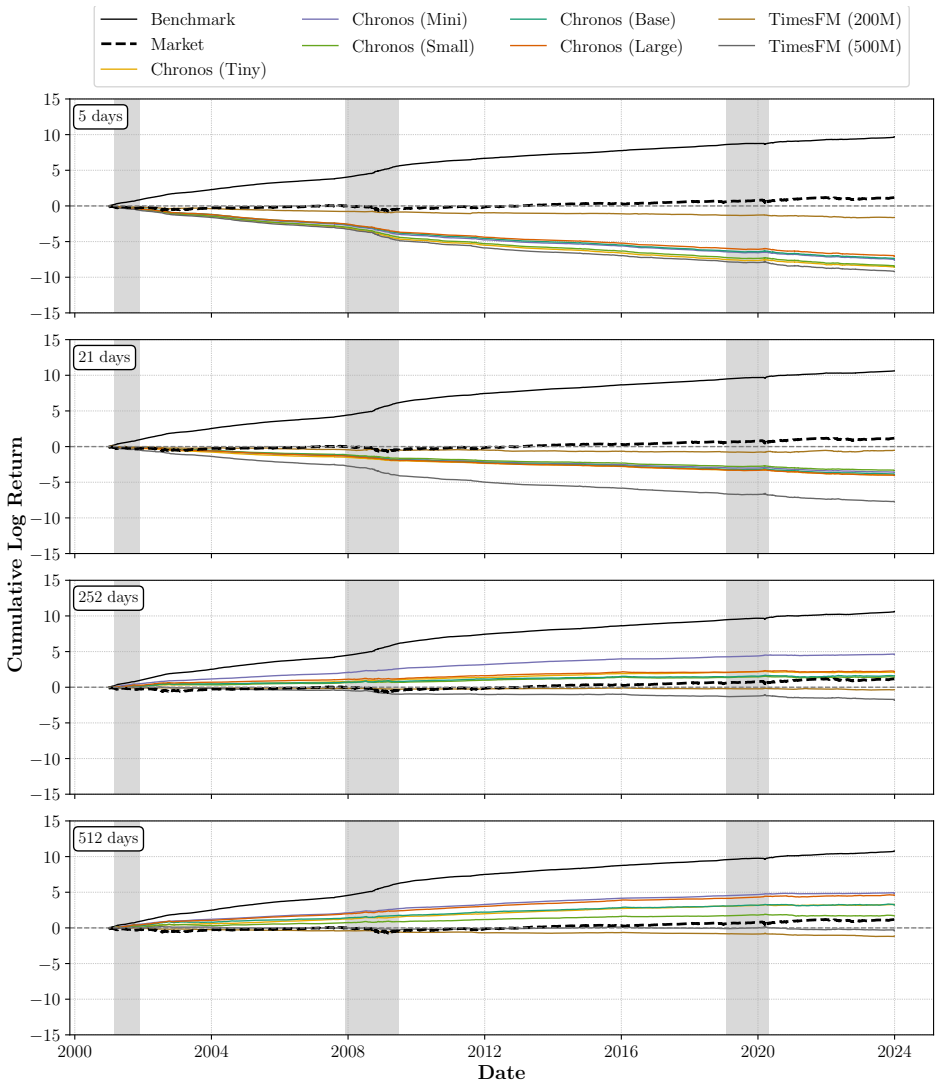
In terms of portfolio performance, the zero-shot forecasts generated by TSFMs do not translate into economically meaningful trading strategies. As reported in Table 7, portfolios constructed using TSFM signals deliver substantially lower annualized returns and Sharpe ratios relative to the benchmark model, and in many cases yield negative performance, particularly for smaller Chronos variants and both TimesFM models. These portfolios are further characterized by higher volatility, larger maximum drawdowns, and unfavorable higher-moment statistics (negative skewness and elevated kurtosis). While larger Chronos models exhibit marginal improvements when longer historical windows are employed, their profitability remains well below those of the benchmark. Among all configurations, the best-performing model is Chronos (large) with a 512-day window, yielding an annualized return of 20.17% and a Sharpe ratio of 2.92, which are substantially lower than the 47.25% annualized return and 6.46 Sharpe ratio reported for the benchmark model with the same window size. For the two TimesFM variants, the strongest configuration is TimesFM 2 with a 512-day window, yielding an annualized return of –1.47% and a Sharpe ratio of –0.18. Also, the relative performance of the long and short legs is generally consistent with the benchmark results. Across models and window sizes, the long leg tends to deliver higher performance than the short leg. Overall, unlike the benchmark model whose forecasting accuracy translates into economically significant results, these TSFMs lack the zero-shot generalization capability necessary for effective portfolio construction.[22]

Figure 2 illustrates the cumulative log returns of long–short portfolios generated from model forecasts across different window sizes. Each panel shows how an initial investment evolves over time when trading on TSFM signals, compared with the benchmark model and the market (S&P 500). Shaded regions correspond to U.S. recession periods as identified by the National Bureau of Economic Research (NBER). The benchmark produces steadily rising and stable cumulative return paths across all window sizes, consistently outperforming the market. In contrast, portfolios constructed using TSFM forecasts display flat or declining cumulative returns, with many TSFMs, particularly smaller Chronos variants and both TimesFM versions, exhibiting persistent losses and sharp drawdowns. Even the larger Chronos models show only slight improvements at longer window sizes and still fail to generate sustained positive returns. Overall, these results reinforce our previous findings and indicate that TSFM-based forecasts do not yield profitable long–short trading strategies in a zero-shot setting, in clear contrast to the benchmark portfolio.[23]

[22]The full cross-sectional return distribution for zero-shot TSFMs is reported in Table B.3. While the benchmark model produces a clear and monotonic spread in returns across deciles in Table B.2, TSFMs generally fail to generate a stable ranking structure. In many cases, decile returns appear noisy or inverted, and the H–L spreads are small or negative, particularly for smaller Chronos and TimesFM models. Even larger TSFM variants only exhibit weak and inconsistent spread patterns, indicating limited ability to order stocks by expected returns in a zero-shot setting.

[23]Figure B.1 presents the cumulative log returns of the long and short portfolio legs separately for each model. The results are consistent with our previous conclusions: while the benchmark model delivers stable performance across the long and short portfolio legs, TSFM-based portfolios exhibit limited economic value in the zero-shot setting.

Figure 2: Cumulative Log Returns of Zero-Shot TSFMs: Long–Short Portfolios



Note: This figure displays the cumulative log returns of long–short portfolios constructed using various forecasting models over rolling windows of 5, 21, 252, and 512 trading days. The benchmark model is CatBoost, the best-performing model among the benchmarks. The time series foundation models (TSFMs) include Chronos (tiny, mini, small, base, and large) and TimesFM (version 1 with 200 million and version 2 with 500 million parameters). Zero-shot inference is performed using the pre-trained models released by the respective authors. Each subplot corresponds to a specific horizon, as indicated by the text labels in the upper-left corners. The benchmark model (CatBoost) is highlighted in black with bold lines. The dashed black line represents the cumulative log return of the market (S&P 500). Shaded areas indicate U.S. recession periods, as defined by the National Bureau of Economic Research (NBER). All portfolios are equally weighted.

As discussed earlier, one of the central claims motivating the development of TSFMs is their purported ability to achieve strong zero-shot forecasting performance. However, the empirical results presented thus far demonstrate that, at least in the case of Chronos and TimesFM, this claim does not hold within the financial forecasting domain. While our analysis focuses on these two widely used models, we extend the zero-shot evaluation to a broader set of TSFMs, encompassing all major publicly available models and their respective size variants. This expanded analysis offers a more comprehensive and representative assessment of how current TSFMs perform relative to established benchmark models. Specifically, our evaluation includes Chronos-Bolt (tiny, mini, small, and base), TimesFM 2.5, Moirai (small, base, and large) and Moirai 2 (Woo et al., 2024), Kairos (10M, 23M, and 50M parameters) (Feng et al., 2025), Moment (small, base, and large) (Goswami et al., 2024), Lag-Llama (Rasul et al., 2023), TiRex (Auer et al., 2025), FlowState (Graf et al., 2025), TTM (Ekambaram et al., 2024), Toto (Cohen et al., 2024), and Sundial (Liu et al., 2025). The forecasting performance results are reported in Table B.4 and Table B.5, portfolio performance results in Table B.6 and Table B.7, and spread portfolio performance results in Table B.8 and Table B.9.[24]

The extended zero-shot evaluation across a diverse suite of TSFMs reveals pronounced heterogeneity in forecasting accuracy, robustness, and portfolio level performance. Forecasting performance variation appears strongly associated with model scale, with larger models generally delivering improved accuracy. The most salient finding, however, concerns the pronounced weakness in goodness-of-fit, as indicated by consistently low R^2_{OOS} values across all evaluated TSFMs. In nearly all tested TSFMs, these models underperform relative to the benchmark, underscoring persistent challenges in achieving reliable goodness-of-fit. At the portfolio level, Toto demonstrates the strongest zero-shot performance, followed by Lag-Llama, Sundial, and TiRex. The subsequent group of models, including Moirai 2, TimesFM 2.5, and Moment (large), exhibit moderate performance among the TSFMs. The remaining TSFMs can be classified as relatively weak performers.

A comparison of forecasting performance indicates that, in terms of goodness-of-fit, even Toto performs relatively poorly, yielding substantially lower R^2_{OOS} values relative to the benchmark. In its best-performing configuration, the model attains an R^2_{OOS} of -114.18% for the 512-day window, which is markedly worse than the benchmark model's value of -0.03% for the same window size. Despite

[24]Different TSFM architectures can generate outputs in fundamentally different ways, meaning that a consistent aggregation approach is not always achievable. To ensure comparability across models, we adapt our procedure accordingly: if the model provides multiple generated outputs, we use the mean; otherwise, we use the median, and if neither is available, we use the model's point forecasting. The number of generated samples for each model follows the proposed configuration, with Chronos, Kairos, Sundial, and TTM producing 20 samples each, and Lag-Llama, Toto, and Moirai producing 100 samples each, consistent with the values reported in the respective studies, except where computational constraints required a proportional reduction. For TSFMs designed to generate a single output per forecast, only one prediction is produced accordingly.

this, the model exhibits more balanced directional accuracy. For the same configuration, it attains upward and downward accuracy rates of 55.59% and 45.81%, respectively, compared with 69.14% and 33.37% for the benchmark model. Consequently, it achieves a higher F1 score of 0.50, relative to the benchmark model's 0.49. Using a 512-day window, the benchmark's long–short Sharpe is 6.46 (its peak occurs at 252 days with Sharpe 6.79), with standard deviation 7.31%, maximum drawdown 14.34%, skewness +1.72, and kurtosis 26.79. In comparison, Toto attains a slightly lower Sharpe of 6.22 but lower standard deviation 4.14% and maximum drawdown 8.96%. However, Toto's return distribution exhibits materially higher kurtosis (82.53) and negative skewness (−2.66), consistent with heavier downside tails. Thus, relative to the benchmark, Toto trades lower typical variability for greater exposure to rare but severe losses; it does not unambiguously dominate the benchmark on a risk–return basis. The details of the pre-training datasets for the TSFMs are presented in Table D.2, which highlights the substantial data volumes used by these models, particularly Toto with approximately 2.3 trillion observations. The enhanced performance observed in TSFMs such as Toto cannot be explained solely by architectural design improvements; it is more plausibly driven, at least in part, by the scale of the pre-training data employed.

Overall, the zero-shot evaluation reveals pronounced heterogeneity in performance across existing TSFMs. While several models fail to generate meaningful predictive signals and perform close to random, others achieve moderate accuracy and exhibit early signs of generalization, particularly when longer window sizes are employed. Moreover, while TSFMs tend to display inferior performance in terms of R^2_{OOS} compared to benchmark models, this pattern is not generally observed for directional accuracy. R^2_{OOS} measures the calibration of point forecasts, which is a stringent criterion given the high noise in daily excess returns. In contrast, portfolio performance depends primarily on discrimination, that is, the ability to correctly capture the sign and cross-sectional ranking of excess returns. A model may exhibit a low R^2_{OOS} because its predicted values are biased or mis-scaled, yet still generate an informative ordering of assets. When returns are sorted and a top-minus-bottom long–short portfolio is constructed, even modest predictive advantages in ranking can accumulate into economically significant Sharpe ratios. It is also noteworthy that, as presented in Table D.2, apart from a few TSFMs pre-trained on Bitcoin price series, no directly related return data are utilized for pre-training. The fact that some models can reach these levels of performance despite having limited or no exposure to financial data during pre-training is noteworthy, as it suggests that generic time series representations may already encode structural and temporal features transferable to complex financial forecasting tasks. However, these mixed outcomes indicate that zero-shot performance alone may not fully capture the underlying potential of these models. Consequently, a systematic investigation involving both

fine-tuning and domain-adaptive pre-training is essential to assess whether currently low-performing architectures such as Chronos and TimesFM can be improved to deliver more stable, consistent, and economically meaningful performance.

5.1.3 Fine-Tuned Results

Although the primary appeal of TSFMs lies in their zero-shot performance, most studies also recommend fine-tuning as a secondary step to enhance predictive accuracy when zero-shot results are suboptimal. This approach is consistent with evidence from LLMs, where fine-tuning has been shown to improve performance on domain-specific tasks. Accordingly, in this study, we extend our analysis by fine-tuning the Chronos and TimesFM models of varying sizes to investigate how this process influences both forecasting and portfolio performance.

As shown in Table 8, most fine-tuned Chronos variants and both TimesFM models continue to exhibit weak out-of-sample fit, with overall accuracy hovering around 50%, similar to their zero-shot counterparts in Table 6. The notable exception is Chronos (large), which, after fine-tuning, achieves positive R^2_{OOS} ($R^2_{OOS} \approx 0.46$–0.48) across window sizes and market-caps. Nonetheless, there are some improvements, particularly in the R^2_{OOS} values of TimesFM models and, in some cases, slight gains in overall accuracy. Also, similar to the zero-shot setting, results after fine-tuning are mixed between small-cap and large-cap stocks, though some models show a slight edge for small-cap stocks. For portfolio results, the pattern is even clearer: relative to the zero-shot portfolio results in Table 7, fine-tuning generally deteriorates performance, often reducing the Sharpe ratio and, in several cases, reversing it from positive to negative values. Annualized returns for the fine-tuned TSFM portfolios are mostly negative, with Sharpe ratios ranging from –3.34 to 0.07. Even for Chronos (large), the improved forecasting metrics do not translate into competitive trading performance.[25] The best-performing configuration is the Chronos (large) model, which attains an annualized return of 0.23% and a Sharpe ratio of 0.07, both far below the benchmark model's 46.50% and 6.79 for the same window size. Also, fine-tuned TSFMs show a similar pattern in which the long leg outperforms the short leg. Figure 3 presents the cumulative log returns of long–short portfolios for fine-tuned TSFMs. Compared with the results in Figure 2, the findings further confirm that fine-tuning generally degrades the performance of TSFM across all window sizes, relative to the zero-shot setting.[26]

[25]Table B.10 presents the results obtained after fine-tuning, while Table B.3 reports the spread portfolio performance of the zero-shot TSFMs. Across Chronos and TimesFM and all window sizes, fine-tuning compresses the decile return profile as lows rise, highs fall, and middle deciles bunch, most strongly at short horizons and still present at longer horizons. This compression reduces portfolio profitability and risk-adjusted performance.

[26]Figure B.2 displays the cumulative log returns of the long and short portfolio legs separately for each model. The results are consistent with our previous findings, suggesting that fine-tuning the TSFMs does not lead to any improvement in portfolio performance.

Table 8: Fine-Tuned TSFMs - Forecasting Performance

Model	**Benchmark**				**Chronos (Tiny)**				**Chronos (Mini)**				**Chronos (Small)**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
R^2_{OOS}	-0.25	-0.05	-0.03	-0.03	-284.11	-260.19	-204.18	-206.36	-562.12	-127.03	-75.19	-71.96	-288.93	-86.10	-58.67	-51.89
	-0.50	-0.38	-0.32	-0.28	-286.65	-293.73	-282.37	-300.82	-565.51	-149.73	-115.26	-117.61	-268.75	-97.87	-78.13	-70.01
	0.39	0.67	0.64	0.60	-290.49	-233.94	-140.04	-133.98	-578.10	-108.11	-47.08	-42.98	-322.05	-75.72	-42.58	-38.33
Overall Acc.	50.94	51.12	51.08	51.16	49.69	49.71	49.77	49.82	49.86	49.86	50.02	50.04	49.92	49.79	49.86	49.84
	50.70	50.74	50.80	50.95	50.53	50.83	50.92	50.95	50.55	50.81	50.81	50.80	50.40	50.78	50.82	50.76
	51.96	52.48	52.33	52.39	48.60	48.30	48.45	48.54	49.05	48.81	49.32	49.38	49.46	48.65	48.79	48.79
Up Acc.	68.96	67.82	66.87	69.14	76.67	87.92	90.35	90.30	72.87	84.18	79.70	78.87	67.87	85.31	84.45	82.83
	72.05	72.97	71.06	74.39	77.47	89.46	94.07	94.42	74.20	87.64	87.99	87.85	70.13	88.00	88.76	86.77
	68.00	65.11	64.65	66.07	75.25	84.75	83.13	82.36	70.44	77.63	66.92	65.45	63.72	80.14	77.32	76.23
Down Acc.	33.14	34.60	35.59	33.37	23.35	12.34	10.09	10.20	27.17	16.25	21.01	21.90	32.25	15.03	16.02	17.57
	28.11	27.25	29.59	26.24	22.48	10.52	5.93	5.60	25.68	12.35	12.10	12.25	29.72	11.97	11.30	13.27
	37.42	41.01	41.24	39.87	24.80	15.68	17.37	18.19	29.63	22.89	33.55	35.03	36.44	20.31	23.10	24.11
F1	0.49	0.49	0.50	0.49	0.46	0.42	0.40	0.40	0.47	0.43	0.45	0.46	0.48	0.43	0.43	0.44
	0.47	0.47	0.48	0.47	0.46	0.41	0.38	0.38	0.47	0.42	0.42	0.42	0.48	0.42	0.42	0.42
	0.51	0.52	0.52	0.51	0.46	0.42	0.43	0.43	0.47	0.45	0.48	0.48	0.49	0.44	0.45	0.46

Model	**Chronos (Base)**				**Chronos (Large)**				**TimesFM 1 (200M)**				**TimesFM 2 (500M)**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
R^2_{OOS}	-1507.22	-11.16	-4.81	-5.42	0.47	0.47	0.46	0.46	-30.17	-15.94	-5.30	-5.10	-532.65	-565.05	-176.49	-147.66
	-1496.14	-11.72	-5.02	-5.74	0.46	0.45	0.45	0.45	-29.65	-14.56	-5.54	-5.63	-520.63	-561.43	-141.04	-124.07
	-1548.14	-10.78	-4.67	-5.22	0.48	0.48	0.48	0.48	-32.07	-17.49	-5.18	-4.73	-568.61	-562.05	-178.58	-147.66
Overall Acc.	50.03	49.95	49.99	49.99	49.97	50.24	50.27	50.27	48.72	49.21	49.82	49.88	48.93	49.47	50.00	49.97
	50.00	50.19	49.84	49.94	49.98	49.85	49.82	49.79	49.28	49.78	50.23	50.19	49.41	50.00	50.38	50.29
	50.14	49.64	50.10	49.95	50.02	50.86	50.98	50.98	47.34	48.19	49.38	49.44	47.82	48.59	49.56	49.53
Up Acc.	49.62	58.73	43.83	48.48	49.72	36.13	35.33	35.02	50.95	56.60	59.55	59.89	47.30	55.10	57.48	57.58
	50.66	59.56	41.55	46.29	50.37	36.72	36.89	36.41	53.11	59.11	61.65	62.52	49.37	57.45	59.77	60.01
	47.75	57.32	47.21	51.18	48.71	35.53	33.22	33.17	47.67	54.09	57.84	57.49	44.31	52.64	55.43	55.66
Down Acc.	50.22	41.31	55.96	51.35	50.01	63.85	64.72	65.02	46.44	41.99	40.56	40.34	50.32	43.67	42.62	42.46
	49.12	40.38	58.39	53.61	49.30	63.26	63.06	63.53	45.10	39.93	38.55	37.56	49.10	41.74	40.41	39.97
	51.94	42.67	52.62	48.73	50.85	64.38	66.71	66.75	46.97	43.07	42.11	42.49	50.83	44.84	44.37	44.09
F1	0.49	0.50	0.50	0.50	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.48	0.47	0.48	0.48
	0.49	0.50	0.49	0.50	0.49	0.49	0.48	0.48	0.49	0.49	0.49	0.49	0.48	0.48	0.48	0.48
	0.49	0.49	0.50	0.50	0.49	0.49	0.49	0.49	0.47	0.48	0.49	0.49	0.46	0.47	0.48	0.47

Note: This table presents each metric as a set of three values, ordered from top to bottom: full sample, top 25% of firms by market capitalization (large-cap), and bottom 25% (small-cap) for various predictive models across different window sizes (5, 21, 252, and 512 trading days). The benchmark model is CatBoost, the best-performing model among the benchmarks. The time series foundation models (TSFMs) include Chronos (tiny, mini, small, base, and large) and TimesFM (version 1 with 200 million and version 2 with 500 million parameters). The models released by the respective authors are fine-tuned on an annual basis. Metrics are first computed separately for each calendar year using all stock-date observations within that year. The reported values represent the average of these yearly statistics. Metrics include out-of-sample R^2 (R^2_{OOS}), overall directional accuracy, upward and downward classification accuracy, and macro-averaged F1 score. 'Overall Acc.' denotes overall directional accuracy, 'Up Acc.' and 'Down Acc.' represent the model's accuracy in predicting upward and downward excess returns respectively, and 'F1' refers to the macro-averaged F1 score.

Table 9: Fine-Tuned TSFMs - Portfolio Performance

Model	Benchmark				Chronos (Tiny)				Chronos (Mini)				Chronos (Small)			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
	42.54	46.52	46.50	47.25	-3.43	-2.59	-2.46	-1.01	-3.34	-1.84	-0.32	-0.86	-2.87	-2.72	-1.36	-1.67
Annualized Return	26.58	28.26	29.00	29.22	3.59	3.42	3.69	4.57	3.73	3.94	4.39	4.52	3.13	3.94	4.61	4.67
	15.95	18.26	17.50	18.03	-7.01	-6.01	-6.15	-5.58	-7.07	-5.78	-4.71	-5.38	-6.00	-6.66	-5.97	-6.34
	7.40	6.86	6.85	7.31	3.71	5.84	7.17	6.82	3.58	5.77	6.21	5.63	3.51	5.71	5.84	5.07
Standard Deviation	14.14	13.88	13.71	13.81	13.74	14.73	14.79	14.41	13.64	14.65	14.48	13.99	13.47	14.42	14.12	13.63
	12.09	12.37	12.31	12.31	11.77	10.12	8.79	8.80	11.55	10.09	9.56	9.64	11.37	10.06	9.67	9.79
	5.74	6.78	6.79	6.46	-0.92	-0.44	-0.34	-0.15	-0.93	-0.32	-0.05	-0.15	-0.82	-0.48	-0.23	-0.33
Sharpe Ratio	1.88	2.04	2.12	2.12	0.26	0.23	0.25	0.32	0.27	0.27	0.30	0.32	0.23	0.27	0.33	0.34
	1.32	1.48	1.42	1.46	-0.60	-0.59	-0.70	-0.63	-0.61	-0.57	-0.49	-0.56	-0.53	-0.66	-0.62	-0.65
	16.88	18.46	18.45	18.75	-1.36	-1.03	-0.97	-0.40	-1.32	-0.73	-0.13	-0.34	-1.14	-1.08	-0.54	-0.66
Daily Return (bps)	10.55	11.22	11.51	11.59	1.42	1.36	1.46	1.81	1.48	1.56	1.74	1.79	1.24	1.56	1.83	1.85
	6.33	7.25	6.94	7.15	-2.78	-2.38	-2.44	-2.21	-2.81	-2.29	-1.87	-2.13	-2.38	-2.64	-2.37	-2.51
	12.08	13.10	13.62	14.34	56.25	52.31	52.83	35.19	55.21	44.49	28.12	27.81	50.94	50.14	31.62	38.69
Max DD	32.51	32.85	31.86	32.62	43.19	41.73	43.97	42.05	42.14	41.51	40.52	41.40	44.79	42.37	42.41	37.93
	30.31	28.43	30.20	27.85	83.96	79.55	78.45	76.23	84.08	78.12	72.25	75.02	79.97	81.08	77.73	79.76
	4.86	3.75	4.15	4.93	5.06	2.07	3.20	2.32	2.14	2.37	3.04	3.50	1.45	6.79	6.95	2.60
Max DD (1-day)	8.28	8.21	8.19	8.19	7.12	7.33	6.87	7.13	7.34	7.25	6.73	6.99	6.90	7.41	6.82	6.93
	5.56	4.49	5.03	5.27	5.80	5.07	4.49	4.08	5.34	5.00	3.87	6.20	5.66	6.44	6.52	6.20
	2.24	2.19	2.25	1.72	-1.63	0.20	0.19	0.21	-0.05	1.09	1.13	0.20	0.13	-0.96	-0.94	0.08
Skew	-0.01	0.00	0.02	-0.05	-0.21	-0.13	-0.18	-0.18	-0.26	-0.15	-0.12	-0.20	-0.22	-0.18	-0.18	-0.21
	0.35	0.38	0.38	0.36	0.24	0.45	0.63	0.68	0.30	0.44	0.62	0.36	0.36	0.19	0.26	0.28
	33.34	31.81	32.72	26.79	40.29	3.72	3.25	2.95	4.62	18.07	17.92	7.02	2.98	24.64	24.78	4.55
Kurt	11.62	12.31	12.53	12.56	6.45	5.97	4.70	4.86	6.88	6.00	5.35	5.42	6.51	6.03	5.19	5.68
	6.17	5.16	5.40	5.53	7.94	8.93	10.64	10.44	7.83	8.18	7.81	8.66	8.31	10.20	10.17	9.56

Model	Chronos (Base)				Chronos (Large)				TimesFM 1 (200M)				TimesFM 2 (500M)			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
	-2.00	-0.61	-1.44	-1.03	-2.11	-0.69	0.23	-0.94	-32.08	-20.45	-1.73	-2.47	-28.72	-17.40	-2.44	-4.02
Annualized Return	3.79	4.58	4.45	4.77	4.47	4.78	5.10	4.69	-9.94	-5.05	3.68	3.63	-7.94	-3.99	3.62	3.13
	-5.79	-5.20	-5.89	-5.80	-6.57	-5.47	-4.87	-5.63	-22.14	-15.40	-5.41	-6.10	-20.78	-13.41	-6.06	-7.15
	2.80	2.50	2.98	2.98	2.62	2.76	3.10	2.97	9.62	9.80	6.04	5.55	9.84	10.33	6.53	5.67
Standard Deviation	12.59	12.99	11.34	11.15	12.39	11.99	11.85	11.74	11.79	11.64	12.54	12.45	10.81	10.97	12.12	12.04
	12.87	12.17	12.22	11.93	12.53	12.60	11.62	11.66	14.24	14.51	12.01	11.70	14.52	14.60	11.98	11.66
	-0.72	-0.25	-0.48	-0.35	-0.80	-0.25	0.07	-0.32	-3.34	-2.09	-0.29	-0.45	-2.92	-1.68	-0.37	-0.71
Sharpe Ratio	0.30	0.35	0.39	0.43	0.36	0.40	0.43	0.40	-0.84	-0.43	0.29	0.29	-0.73	-0.36	0.30	0.26
	-0.45	-0.43	-0.48	-0.49	-0.52	-0.43	-0.42	-0.48	-1.55	-1.06	-0.45	-0.52	-1.43	-0.92	-0.51	-0.61
	-0.79	-0.24	-0.57	-0.41	-0.84	-0.28	0.09	-0.37	-12.73	-8.12	-0.69	-0.98	-11.40	-6.91	-0.97	-1.60
Daily Return (bps)	1.50	1.82	1.77	1.89	1.77	1.90	2.02	1.86	-3.94	-2.00	1.46	1.44	-3.15	-1.58	1.44	1.24
	-2.30	-2.06	-2.34	-2.30	-2.61	-2.17	-1.93	-2.23	-8.79	-6.11	-2.15	-2.42	-8.25	-5.32	-2.41	-2.84
	42.03	16.90	28.74	21.45	42.76	20.85	12.22	21.40	99.94	99.16	50.50	49.32	99.88	98.33	56.73	63.01
Max DD	44.33	41.18	37.91	41.15	42.81	39.57	37.87	40.26	92.55	77.16	55.79	54.16	87.92	69.97	41.25	40.59
	81.53	78.51	79.68	79.34	83.02	80.12	77.33	79.02	99.49	97.69	81.64	79.98	99.32	96.41	83.81	83.49
	1.12	1.09	5.12	5.59	1.13	1.54	1.00	5.97	7.96	12.01	3.02	6.53	8.00	12.01	3.73	6.53
Max DD (1-day)	7.22	6.80	6.87	6.30	6.74	6.97	6.72	6.81	5.94	5.53	7.05	6.88	5.28	5.50	6.99	6.78
	6.07	5.55	4.94	5.21	5.96	5.78	5.05	5.37	8.13	11.65	5.87	6.22	8.15	11.65	5.43	6.22
	1.96	-0.01	-3.53	-4.81	-0.07	-0.52	9.30	-5.54	-1.60	-2.85	-0.35	-1.74	-1.30	-2.02	-0.26	-1.30
Skew	-0.31	-0.23	-0.37	-0.39	-0.28	-0.39	-0.25	-0.36	-0.48	-0.56	-0.48	-0.45	-0.43	-0.40	-0.33	-0.33
	0.20	0.29	0.27	0.25	0.24	0.23	0.34	0.27	-0.07	-0.45	0.16	0.15	0.04	-0.17	0.36	0.31
	51.11	2.63	96.45	138.03	2.87	6.48	351.97	179.60	18.60	44.29	5.76	28.67	17.19	38.36	7.67	25.88
Kurt	7.04	6.57	6.97	6.45	7.17	7.97	7.93	7.76	5.54	5.27	6.80	6.19	4.30	4.96	6.46	6.11
	7.15	7.07	6.21	6.61	6.90	6.88	6.05	6.24	12.47	16.51	5.20	7.06	11.98	17.10	6.00	7.45

Note: This table reports average yearly portfolio performance metrics across different rolling window sizes (5, 21, 252, and 512 trading days) for each model. The benchmark model is CatBoost, the best-performing model among the benchmarks. The time series foundation models (TSFMs) include Chronos (tiny, mini, small, base, and large) and TimesFM (version 1 with 200 million and version 2 with 500 million parameters). The models released by the respective authors are fine-tuned on an annual basis. Each cell displays three values from top to bottom: long–short portfolio, long-only leg, and short-only leg. Metrics include annualized return, standard deviation, Sharpe ratio, daily return (in basis points), maximum drawdown (Max DD), one-day maximum drawdown (Max DD (1-day)), skewness, and kurtosis of portfolio returns. Portfolios are formed using decile sorting based on model forecasts, with equal weighting across stocks.

5.1.4 Pre-Trained Results

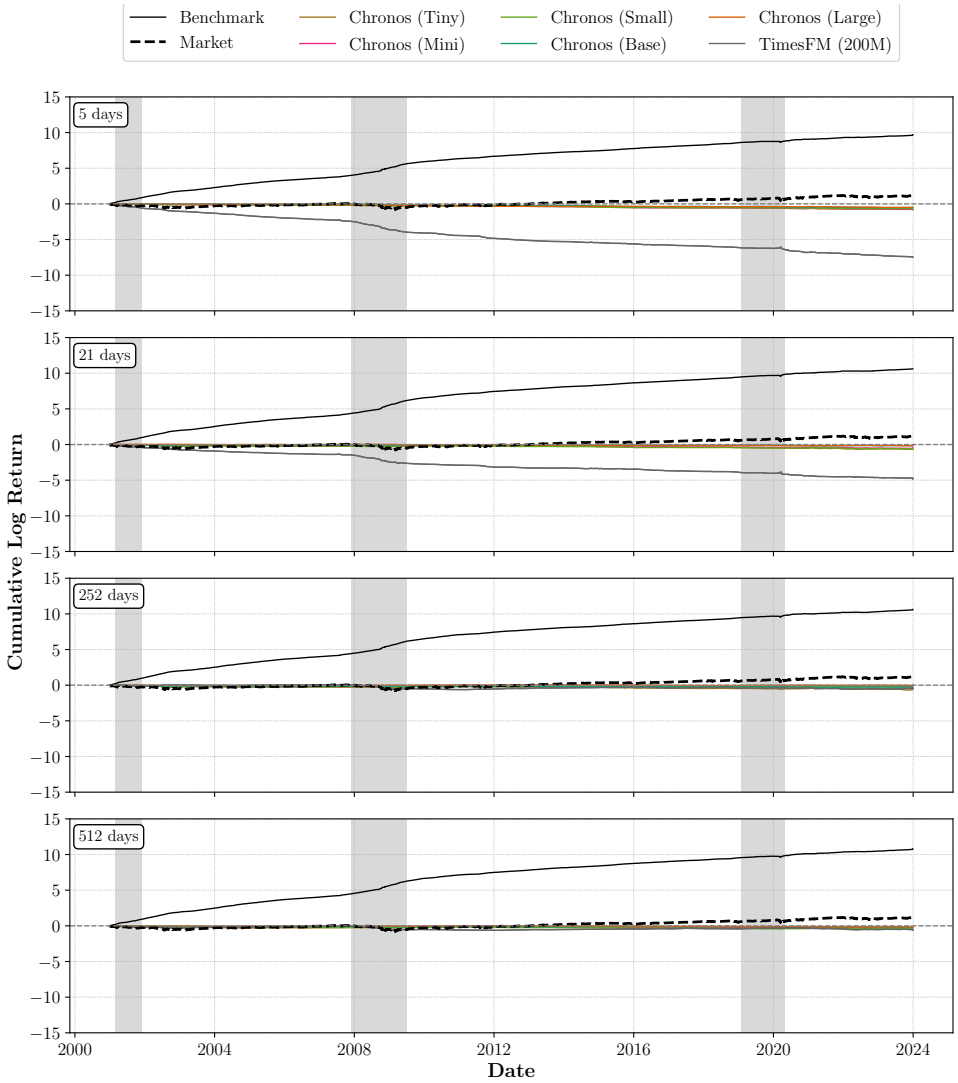
The final test evaluates the performance of TSFMs when pre-trained entirely from scratch. In this context, pre-training refers to the initialization of model parameters using large-scale U.S. excess return data, consistent with the procedure applied to all benchmark models in Section 5.1.1. We construct TSFMs following the Chronos and TimesFM architectures but pre-train them exclusively on our proprietary excess return dataset. This design enables a controlled comparison between models pre-trained on broad, cross-domain datasets and those trained solely on domain-specific financial data. Through this experiment, we aim to determine whether pre-training on financial data enhances the models' ability to capture return dynamics, a challenge that most existing TSFMs have struggled to address effectively, as evidenced by both the zero-shot results in Section 5.1.2 and the fine-tuned results in Section 5.1.3.

Table 10 reports the forecasting performance of TSFMs that have been pre-trained on U.S. excess return data. A comparison of these results with the zero-shot results in Table 6 and the fine-tuned results in Table 8 clearly shows that domain-specific pre-training leads to substantial improvements in predictive accuracy, stability, and consistency across different window sizes. While zero-shot and fine-tuned TSFMs exhibit limited forecasting ability, pre-trained models, especially Chronos, demonstrate substantially better forecasting performance. Across all model specifications, and relative to the zero-shot results, on average Chronos achieves 1.43% improvement in directional accuracy, and 20.57% in R^2_{OOS}. TimesFM demonstrates a decline in both out-of-sample R^2_{OOS} and F1 score metrics; however, it attains an improvement of up to 2.18% in directional accuracy.[27] Also, the results align with the benchmarks, showing a clear edge for small-cap stocks across models and window sizes. Although a clear and consistent improvement is observed when moving from zero-shot and fine-tuned R^2_{OOS} results to pre-trained models, especially for Chronos, the performance gap relative to the benchmark models remains substantial. The best-performing pre-trained TSFM is Chronos (small), which attains an R^2_{OOS} of -0.59% for the 512-day window. In absolute terms, this is approximately 20 times lower than the benchmark models' R^2_{OOS} of -0.03% for the same window size. Once again, this persistent discrepancy highlights the inherent limitations of TSFMs in achieving accurate goodness-of-fit, which stem from their architectural design rather than from the size or nature of the data used for pre-training.[28]

[27] The modified DM tests in Table B.11 corroborate our main findings. Across window sizes, the benchmark (CatBoost) delivers consistently superior out-of-sample accuracy relative to all pre-trained TSFMs. Within the TSFMs, larger models and longer windows generally help, with Chronos (small) tending to outperform Chronos (tiny) and Chronos (mini), and TimesFM (20M) often improving on TimesFM (8M) at medium to long windows. Comparing TSFM families, pre-trained Chronos variants typically outperform pre-trained TimesFM models at comparable window sizes.

[28] To further examine the forecasting performance of the TSFMs, we visualize the predictive distributions produced by different variants of the Chronos models. Figure B.4 presents 30-day-ahead interval forecasts (P_{10}–P_{90}) of daily

Figure 3: Cumulative Log Returns of Fine-Tuned TSFMs: Long–Short Portfolios



Note: This figure displays the cumulative log returns of long–short portfolios constructed using various forecasting models over rolling windows of 5, 21, 252, and 512 trading days. The benchmark model is CatBoost, the best-performing model among the benchmarks. The time series foundation models (TSFMs) include Chronos (tiny, mini, small, base, and large) and TimesFM (version 1 with 200 million and version 2 with 500 million parameters). The models released by the respective authors are fine-tuned on an annual basis. Each subplot corresponds to a specific horizon, as indicated by the text labels in the upper-left corners. The benchmark model (CatBoost) is highlighted in black with bold lines. The dashed black line represents the cumulative log return of the market (S&P 500). Shaded areas indicate U.S. recession periods, as defined by the National Bureau of Economic Research (NBER). All portfolios are equally weighted.

Table 10: Pre-Trained TSFMs - Forecasting Performance

Model	**Benchmark**				**Chronos (Tiny)**				**Chronos (Mini)**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512
R^2_{OOS}	-0.25	-0.05	-0.03	-0.03	-3.77	-1.09	-0.65	-0.61	-4.50	-1.26	-0.60	-0.61
	-0.50	-0.38	-0.32	-0.28	-3.85	-1.24	-0.87	-0.89	-4.43	-1.54	-0.78	-0.83
	0.39	0.67	0.64	0.60	-3.74	-0.83	-0.30	-0.24	-4.61	-0.87	-0.26	-0.23
Overall Acc.	50.94	51.12	51.08	51.16	50.32	50.65	51.27	51.32	50.24	50.57	51.31	51.34
	50.70	50.74	50.80	50.95	49.29	49.86	50.81	50.87	49.31	49.86	50.90	50.87
	51.96	52.48	52.33	52.39	51.77	52.11	52.80	52.91	51.53	51.96	52.88	53.01
Up Acc.	68.96	67.82	66.87	69.14	18.50	31.78	46.82	47.66	22.99	33.22	47.43	46.58
	72.05	72.97	71.06	74.39	17.97	32.02	55.26	55.40	22.31	33.20	55.85	53.22
	68.00	65.11	64.65	66.07	19.80	33.38	41.48	42.68	24.63	35.01	42.63	43.56
Down Acc.	33.14	34.60	35.59	33.37	81.36	68.79	55.20	54.54	76.82	67.25	54.67	55.61
	28.11	27.25	29.59	26.24	81.84	68.00	45.48	45.51	77.43	66.80	45.02	47.72
	37.42	41.01	41.24	39.87	80.29	68.62	62.76	61.90	75.47	66.88	61.85	61.26
F1	0.49	0.49	0.50	0.49	0.44	0.48	0.50	0.50	0.46	0.48	0.50	0.50
	0.47	0.47	0.48	0.47	0.43	0.47	0.48	0.48	0.45	0.47	0.49	0.49
	0.51	0.52	0.52	0.51	0.45	0.50	0.52	0.52	0.47	0.50	0.52	0.52

Model	**Chronos (Small)**				**TimesFM (8M)**				**TimesFM (20M)**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512
R^2_{OOS}	-3.18	-1.43	-0.65	-0.59	-33.01	-16.71	-22.64	-28.15	-35.12	-16.69	-13.83	-15.03
	-3.17	-1.40	-0.92	-0.84	-32.39	-17.85	-25.75	-32.83	-32.93	-16.99	-15.66	-17.60
	-3.20	-1.33	-0.17	-0.15	-35.65	-15.26	-20.82	-25.83	-39.67	-16.09	-12.37	-13.09
Overall Acc.	50.40	50.79	51.36	51.36	49.73	50.80	51.14	51.15	49.88	50.88	51.14	51.12
	49.58	50.08	50.86	50.84	49.20	49.94	50.20	50.26	49.84	50.66	50.95	50.98
	51.65	52.27	53.12	53.20	50.19	52.28	52.82	52.78	49.77	51.53	51.82	51.82
Up Acc.	26.10	36.59	45.39	45.96	39.75	34.48	34.98	35.52	49.57	54.82	57.46	57.60
	25.60	37.04	52.00	51.52	41.50	36.12	36.26	36.80	51.06	55.89	57.98	58.45
	27.57	37.76	42.72	43.93	37.11	32.77	34.21	34.48	47.46	54.05	57.70	57.38
Down Acc.	73.85	64.36	56.85	56.35	59.15	66.32	66.39	65.84	49.34	46.20	44.34	44.14
	74.16	63.16	48.97	49.52	56.93	63.98	64.48	64.00	47.46	44.04	42.74	42.26
	72.89	65.01	62.33	61.38	61.48	69.11	68.59	68.22	51.17	48.71	46.17	46.44
F1	0.47	0.49	0.50	0.50	0.47	0.45	0.43	0.43	0.47	0.47	0.45	0.45
	0.46	0.48	0.49	0.49	0.47	0.45	0.42	0.43	0.47	0.46	0.45	0.45
	0.48	0.51	0.52	0.52	0.47	0.46	0.44	0.44	0.47	0.47	0.46	0.46

Note: This table presents each metric as a set of three values, ordered from top to bottom: full sample, top 25% of firms by market capitalization (large-cap), and bottom 25% (small-cap) for various predictive models across different window sizes (5, 21, 252, and 512 trading days). The benchmark model is CatBoost, the best-performing model among the benchmarks. The time series foundation models (TSFMs) include Chronos (tiny, mini, and small) and TimesFM (with 8 million and 20 million parameters). The models are pre-trained from scratch using U.S. excess return data on an annual basis. Metrics are first computed separately for each calendar year using all stock-date observations within that year. The reported values represent the average of these yearly statistics. Metrics include out-of-sample R^2 (R^2_{OOS}), overall directional accuracy, upward and downward classification accuracy, and macro-averaged F1 score. 'Overall Acc.' denotes overall directional accuracy, 'Up Acc.' and 'Down Acc.' represent the model's accuracy in predicting upward and downward excess returns respectively, and 'F1' refers to the macro-averaged F1 score.

Moving to portfolio performance results in Table 11, and in comparison with the zero-shot results in Table 7 and the fine-tuned results in Table 9, a clear improvement in portfolio performance is observed following domain-specific pre-training. The pre-trained TSFMs, particularly Chronos across all model sizes, exhibit notably higher annualized returns and Sharpe ratios, especially at medium and long window sizes. For instance, Chronos (tiny) achieves Sharpe ratios improving from 1.69 to 3.33 for the 252-day window size and from 2.70 to 4.10 for the 512-day window size, compared with the zero-shot results. Similarly, Chronos (mini) improves from 3.71 to 4.05 at 252 days and from 3.79 to 4.54 at 512 days, while Chronos (small) increases from 0.86 to 4.89 and 0.94 to 5.42 for the same respective window sizes, all relative to their zero-shot results. TimesFM models also benefit from pre-training. TimesFM (8M) achieves Sharpe ratios improving from –0.89 to 2.86 for the 252-day window size and from –0.18 to 3.23 for the 512-day window size, compared with its zero-shot results. Similarly, TimesFM (20M) improves from –0.89 to 3.51 at 252 days and from –0.89 to 3.66 at 512 days, both showing substantial gains relative to their zero-shot results. The changes in annualized returns follow a similar pattern. Moreover, across many model configurations, particularly those with shorter window sizes, the standard deviation of portfolio returns declines, and maximum drawdowns become less severe. Among all TSFMs, the Chronos (small) model exhibits the strongest performance for a window size of 512 days, generating an annualized return of 36.84%, a standard deviation of 6.79%, and a Sharpe ratio of 5.42. Although these values remain below the corresponding benchmark values of 47.25%, 7.31%, and 6.46, respectively, for the same window size, the model substantially narrows the performance gap relative to both the zero-shot and fine-tuned TSFMs. Conversely, we observe a marked improvement in both maximum drawdown and one-day maximum drawdown when moving from the benchmark model to Chronos (Small). For the same window size, these metrics decline from 18.75 and 14.34 to 9.01 and 6.59, respectively. Also, for the same window size, compared with the benchmark (skew = 1.72, kurtosis = 26.79), the Chronos (small) model shows slightly negative

excess returns for selected U.S. stocks (AAPL and MSFT), generated using Chronos TSFMs. The left column displays zero-shot forecasts from the publicly available Chronos model, the middle column shows forecasts based on a fine-tuned, publicly available Chronos model, and the right column reports forecasts from a Chronos model pre-trained on global excess returns. All models use the small configuration. The black lines represent realized excess returns, the shaded blue regions denote 80% predictive intervals, and the dashed blue lines indicate the 10^{th} and 90^{th} percentile boundaries. The vertical dotted line marks the forecast start date. The forecasting horizon spans 30 trading days, from September 19, 2025, to October 31, 2025. The estimation window comprises 512 trading days, covering the period from September 11, 2023, to September 18, 2025. Consistent with Section 5.1.2, the number of generated samples is fixed at 20. For both AAPL and MSFT, the fine-tuned Chronos model appears systematically mis-centered: realized excess returns fall outside the 80% predictive intervals. In contrast, the zero-shot variant produces slightly better forecasts, while the pre-trained version tracks the realized return path more closely. This illustrates why TSFMs typically exhibit weak R^2_{OOS} performance. TSFM-based forecasts tend to align poorly with the realized magnitudes of excess returns, capturing primarily the directional movements rather than their exact values. This discrepancy leads to lower R^2_{OOS} relative to the benchmarks, despite maintaining reasonably strong directional predictive power. This pattern is consistent with our broader finding that domain-specific pre-training materially enhances forecast stability and directional accuracy, whereas fine-tuning on the same dataset, as implemented here, does not. Nonetheless, we do not view fine-tuning as inherently detrimental; more targeted approaches, such as using narrower, asset-specific datasets or optimized hyperparameter configurations, may restore or even improve performance.

skew (–0.29) and lower yet still extreme kurtosis (20.43). This indicates a shift toward mild downside asymmetry and somewhat reduced tail risk. Moreover, consistent with the benchmark, zero-shot, and fine-tuned results, the pre-trained TSFMs also exhibit stronger performance for the long leg than for the short leg.[29]

Figure 4 displays the cumulative long–short returns of the pre-trained TSFMs. A comparison with the zero-shot results in Figure 2 and the fine-tuned results in Figure 3 clearly demonstrates that pre-trained TSFMs deliver substantially improved and more stable portfolio performance. The cumulative return trajectories show that pre-training enables both Chronos and TimesFM models to generate persistent positive excess returns over time, in contrast to the flatter or declining patterns observed in the zero-shot and fine-tuned settings. The improvement is particularly pronounced at longer rolling window sizes (252 and 512 days), where the pre-trained TSFMs display smoother upward trends.[30]

Overall, these results confirm that domain-specific pre-training substantially strengthens TSFM-based portfolio strategies, transforming previously weak or noisy return profiles into robust and economically meaningful performance. These pre-trained TSFMs achieve better alignment between predictive performance and portfolio performance, yielding higher Sharpe ratios, lower drawdowns, and more stable returns compared to their generic off-the-shelf pre-trained counterparts. This improvement is also achieved with substantially smaller model sizes compared to the originally released pre-trained models. Specifically, TimesFM 2 is scaled down from 500 million parameters to 8 million and 20 million parameter variants. In addition, our Chronos models include tiny, mini, and small variants, with even the tiny variant outperforming the largest Chronos model released by the original authors, despite the latter being pre-trained on a more diverse and substantially larger dataset. Another important finding is that our pre-trained models generally outperform not only the Chronos and TimesFM models released by their respective authors, but also the majority of other general-purpose models evaluated in Section 5.1.2, despite differences in their architectures and pre-training datasets.

This performance gap underscores the critical importance of pre-training context: exposure to financial time series enables models to internalize domain-relevant statistical patterns that generic pre-training fails to capture. Consequently, although the benchmark model still delivers the best overall

[29]Table B.12 presents the spread portfolio performance of pre-trained TSFMs. A comparison with the zero-shot results in Table B.3 and the fine-tuned results in Table B.10 shows that domain-specific pre-training substantially improves the performance of TSFM-based portfolios. While zero-shot models exhibit weak or inconsistent return spreads, pre-trained variants generate markedly stronger and more stable performance, particularly at longer window sizes. Although the benchmark model continues to achieve the highest overall returns, pre-training significantly narrows the performance gap.

[30]Figure B.3 presents the cumulative log returns of the long and short portfolio legs for each model. The results are consistent with our previous findings, indicating that pre-training the TSFMs substantially improves portfolio performance in both the long and short portfolio legs when evaluated separately.

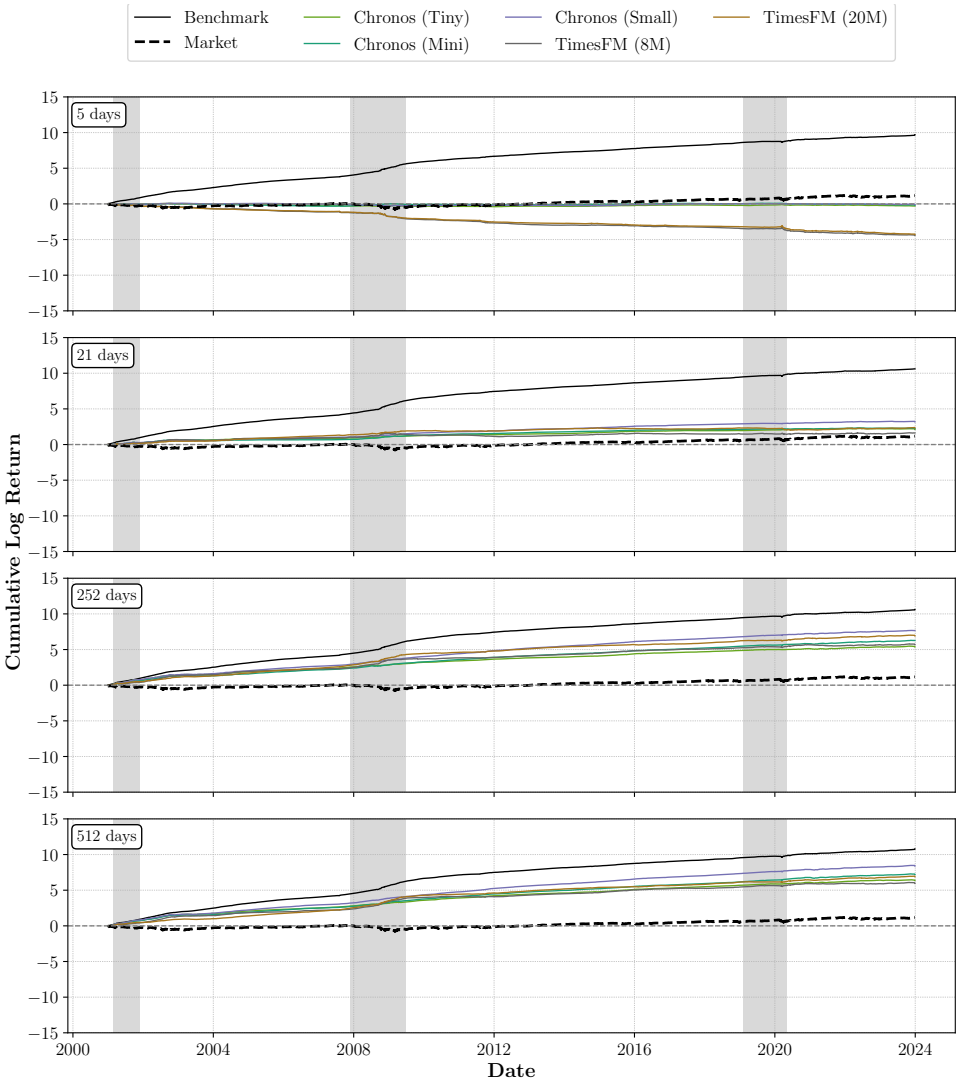
Table 11: Pre-Trained TSFMs - Portfolio Performance

Model	Benchmark				Chronos (Tiny)				Chronos (Mini)			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512
	42.54	46.52	46.50	47.25	-1.14	9.47	23.69	28.00	-0.70	9.84	27.55	31.63
Annualized Return	26.58	28.26	29.00	29.22	6.33	11.63	19.41	21.82	7.31	11.76	21.01	23.80
	15.95	18.26	17.50	18.03	-7.47	-2.17	4.27	6.18	-8.01	-1.92	6.54	7.83
	7.40	6.86	6.85	7.31	5.56	6.36	7.12	6.83	4.89	5.89	6.80	6.96
Standard Deviation	14.14	13.88	13.71	13.81	10.42	11.07	11.77	11.96	11.13	11.40	12.20	11.98
	12.09	12.37	12.31	12.31	14.02	14.40	14.18	13.90	13.67	13.95	13.64	13.67
	5.74	6.78	6.79	6.46	-0.21	1.49	3.33	4.10	-0.14	1.67	4.05	4.54
Sharpe Ratio	1.88	2.04	2.12	2.12	0.61	1.05	1.65	1.82	0.66	1.03	1.72	1.99
	1.32	1.48	1.42	1.46	-0.53	-0.15	0.30	0.44	-0.59	-0.14	0.48	0.57
	16.88	18.46	18.45	18.75	-0.45	3.76	9.40	11.11	-0.28	3.90	10.93	12.55
Daily Return (bps)	10.55	11.22	11.51	11.59	2.51	4.62	7.70	8.66	2.90	4.67	8.34	9.44
	6.33	7.25	6.94	7.15	-2.97	-0.86	1.70	2.45	-3.18	-0.76	2.60	3.11
	12.08	13.10	13.62	14.34	38.04	12.58	10.35	10.68	31.70	12.77	10.07	11.73
Max DD	32.51	32.85	31.86	32.62	34.73	27.08	30.70	28.47	25.49	28.61	30.38	29.23
	30.31	28.43	30.20	27.85	86.71	63.71	46.14	42.83	87.79	60.81	42.73	40.11
	4.86	3.75	4.15	4.93	3.68	7.04	4.36	4.31	4.98	6.13	3.70	5.42
Max DD (1-day)	8.28	8.21	8.19	8.19	6.02	6.47	7.63	7.31	7.13	7.05	8.13	8.14
	5.56	4.49	5.03	5.27	7.44	6.59	6.15	5.80	6.35	5.75	4.93	5.80
	2.24	2.19	2.25	1.72	-0.73	-1.36	-0.08	0.19	-1.16	-1.02	0.17	0.03
Skew	-0.01	0.00	0.02	-0.05	-0.47	-0.32	-0.37	-0.19	-0.41	-0.29	-0.18	-0.16
	0.35	0.38	0.38	0.36	0.07	0.01	0.18	0.16	0.08	0.03	0.23	0.20
	33.34	31.81	32.72	26.79	10.22	24.30	6.63	7.72	18.03	23.67	10.41	12.04
Kurt	11.62	12.31	12.53	12.56	8.05	10.49	14.61	14.09	11.33	13.20	17.99	18.09
	6.17	5.16	5.40	5.53	7.47	6.07	5.46	5.55	6.53	5.01	4.39	5.22

Model	Chronos (Small)				TimesFM (8M)				TimesFM (20M)			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512
	-0.32	13.99	33.61	36.84	-18.77	7.54	25.14	26.31	-18.22	10.50	30.50	30.36
Annualized Return	6.72	14.08	25.27	26.76	-3.71	9.38	19.01	19.80	-3.50	11.10	21.28	21.41
	-7.04	-0.09	8.34	10.08	-15.06	-1.84	6.13	6.51	-14.72	-0.60	9.21	8.95
	4.83	5.47	6.87	6.79	9.87	9.72	8.78	8.15	9.92	8.48	8.68	8.30
Standard Deviation	11.13	11.60	11.72	11.77	11.14	10.85	12.34	12.60	11.25	11.82	13.23	13.33
	13.74	13.76	13.83	13.55	14.36	14.32	12.69	12.19	14.33	13.53	12.24	11.81
	-0.07	2.56	4.89	5.42	-1.90	0.78	2.86	3.23	-1.84	1.24	3.51	3.66
Sharpe Ratio	0.60	1.21	2.16	2.27	-0.33	0.86	1.54	1.57	-0.31	0.94	1.61	1.61
	-0.51	-0.01	0.60	0.74	-1.05	-0.13	0.48	0.53	-1.03	-0.04	0.75	0.76
	-0.13	5.55	13.34	14.62	-7.45	2.99	9.98	10.44	-7.23	4.17	12.10	12.05
Daily Return (bps)	2.67	5.59	10.03	10.62	-1.47	3.72	7.54	7.86	-1.39	4.41	8.45	8.50
	-2.79	-0.03	3.31	4.00	-5.98	-0.73	2.43	2.58	-5.84	-0.24	3.66	3.55
	34.48	9.82	8.17	9.91	98.79	41.78	17.21	14.24	98.63	29.14	10.82	11.02
Max DD	42.31	28.55	29.16	29.28	67.99	23.63	31.63	32.22	66.79	29.18	31.00	30.78
	84.76	47.64	44.34	41.58	97.56	73.99	40.05	34.76	97.34	66.99	33.61	35.44
	4.91	6.95	3.48	6.59	7.19	5.34	7.39	7.22	8.22	5.52	7.60	7.40
Max DD (1-day)	6.74	7.18	8.14	8.50	7.49	5.72	8.37	8.53	7.37	5.73	8.06	7.98
	7.26	6.52	5.73	6.36	8.56	8.64	6.05	6.05	8.90	8.16	6.44	6.44
	-1.67	-1.45	0.22	-0.29	-1.45	-0.47	0.12	0.64	-1.72	-0.73	0.16	0.63
Skew	-0.52	-0.29	-0.25	-0.25	-0.61	-0.33	-0.34	-0.25	-0.55	-0.42	-0.22	-0.14
	0.01	0.04	0.19	0.11	-0.16	0.06	0.22	0.32	-0.23	-0.00	0.17	0.30
	22.02	41.52	8.91	20.43	16.78	8.91	19.46	25.84	19.81	9.43	20.79	23.94
Kurt	9.58	12.62	17.58	18.00	7.45	8.27	16.25	16.37	6.43	6.72	12.32	11.96
	6.96	5.97	5.73	5.93	11.95	8.59	5.87	6.33	12.59	9.17	6.63	7.18

Note: This table reports average yearly portfolio performance metrics across different rolling window sizes (5, 21, 252, and 512 trading days) for each model. The benchmark model is CatBoost, the best-performing model among the benchmarks. The time series foundation models (TSFMs) include Chronos (tiny, mini, and small) and TimesFM (with 8 million and 20 million parameters). Zero-shot inference is performed using the pre-trained models. Each cell displays three values from top to bottom: long–short portfolio, long-only leg, and short-only leg. Metrics include annualized return, standard deviation, Sharpe ratio, daily return (in basis points), maximum drawdown (Max DD), one-day maximum drawdown (Max DD (1-day)), skewness, and kurtosis of portfolio returns. Portfolios are formed using decile sorting based on model forecasts, with equal weighting across stocks.

Figure 4: Cumulative Log Returns of Pre-Trained TSFMs: Long–Short Portfolios



Note: This figure displays the cumulative log returns of long–short portfolios constructed using various forecasting models over rolling windows of 5, 21, 252, and 512 trading days. The benchmark model is CatBoost, the best-performing model among the benchmarks. The time series foundation models (TSFMs) include Chronos (tiny, mini, and small) and TimesFM (with 8 million and 20 million parameters). Zero-shot inference is performed using the pre-trained models. Each subplot corresponds to a specific horizon, as indicated by the text labels in the upper-left corners. The benchmark model (CatBoost) is highlighted in black with bold lines. The dashed black line represents the cumulative log return of the market (S&P 500). Shaded areas indicate U.S. recession periods, as defined by the National Bureau of Economic Research (NBER). All portfolios are equally weighted.

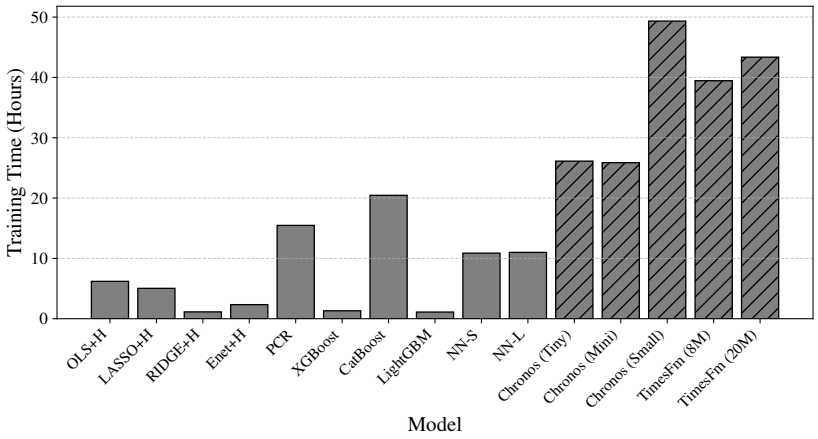
performance, incorporating domain-specific pre-training significantly narrows the gap, suggesting that tailored pre-training represents a crucial step toward unlocking the full potential of TSFMs in financial forecasting and portfolio management. The performance gains from pre-trained TSFMs also mirror prior findings in the LLM literature, where finance-specific models outperform larger general-purpose ones by leveraging domain knowledge (Rahimikia and Drinkall, 2024; He et al., 2025). Our evidence shows that this specialization effect likewise holds for TSFMs.

5.1.5 Training Time Comparison Across Models

Although forecasting and portfolio performance are central to model evaluation, another crucial consideration, particularly for TSFMs, is the computational cost and duration of pre-training. Due to their Transformer-based architectures inherited from LLMs, TSFMs require extensive computational resources and prolonged pre-training times, often spanning several days or weeks on large-scale graphics processing unit (GPU) clusters, whereas benchmark models can be trained efficiently on standard hardware within hours. This disparity underscores the substantial infrastructure demands and energy costs associated with TSFM development, which partly explain the limited number of publicly available pre-trained models. To make these contrasts explicit, we report the training times for benchmark models and the pre-training times for TSFMs, highlighting how computational demands scale with model complexity and the practical trade-offs this introduces.

Figure 5 reports the total training time (in hours) required for each forecasting model. Benchmark models (OLS, Lasso, Ridge, Elastic Net, PCR, XGBoost, CatBoost, LightGBM, and neural networks) are compared with TSFMs (Chronos and TimesFM, indicated by hatched bars). Training times were measured, as far as possible, using identical hardware and software configurations to ensure comparability across models. Reported values correspond to the final model trained for the year 2022, which uses the maximum available U.S. excess return data. All values are approximate, and small differences in training times should not be interpreted as meaningful. With the exception of TSFMs, which were trained on GPU, all other models were trained on central processing unit (CPU). The results show clear contrasts in training durations across models. Benchmark models, including OLS, Lasso, Ridge, and Elastic Net, require approximately 2 to 6 hours of training, whereas PCR takes around 15 hours. In comparison, the neural network models exhibit an average training time of roughly 11 hours. In contrast, TSFMs are far more computationally intensive: Chronos ranges from around 25 hours (tiny and mini) to nearly 50 hours (small), and TimesFM extends to about 40 hours for larger configurations. These results illustrate how model complexity and parameter size substantially increase computational costs.

Figure 5: Training Time by Model



Note: This figure reports the total training time (in hours) required for each forecasting model. Benchmark models (OLS, Lasso, Ridge, Elastic Net, PCR, XGBoost, CatBoost, LightGBM, and neural networks) are compared with time series foundation models (TSFMs, including Chronos and TimesFM, indicated by hatched bars). 'H' indicates that the model is estimated using the Huber loss. Training times were measured, as far as possible, using identical hardware and software configurations to ensure comparability across models. Reported values correspond to the final model trained for the year 2022, which uses the maximum available U.S. excess return data. All values are approximate, and small differences in training times should not be interpreted as meaningful. With the exception of TSFMs, which were trained on GPUs, all other models were trained on CPUs.

The reported values represent comparable durations for the pre-training phase, which constitutes the most computationally intensive stage of model development. During this phase, as described in Section 3.2.3, models learn generalizable representations from large-scale historical datasets through repeated parameter updates and extensive optimization across numerous epochs. The subsequent stage of computational demand occurs during inference, as detailed in Section 3.2.1, when the pre-trained models are applied to out-of-sample data for forecasting. Although not presented here, inference times display a broadly similar pattern to that shown in Figure 5, with TSFMs continuing to require GPU acceleration and longer runtimes compared to benchmark models. However, this disparity is generally less pronounced than during pre-training, as inference involves forward passes rather than gradient-based optimization. Overall, the elevated computational demands of TSFMs at both stages suggest that, given current hardware capabilities, these models cannot yet be regarded as a computationally efficient substitute for conventional forecasting models. However, this gap is expected to narrow over time with continued advancements in hardware and broader access to high-performance computational infrastructure.

5.2 Results with Scaled Data

The data used for pre-training benchmark models and for fine-tuning and pre-training TSFMs in Section 5.1 are U.S. excess returns. One of the claimed potential benefits of using TSFMs is their capability to scale and utilize substantially larger datasets, which traditional and ML models may

lack due to limitations in model architecture, available software, and hardware compatibility. This is consistent with their root models, LLMs, which are also capable of being pre-trained on substantial amounts of textual data at scale. The maximum data size used for testing all models so far has reached 176.96 million observations (see the U.S. panel in Table 1 for the number of observations used in each pre-training year), which is substantially lower than that used for pre-training publicly available TSFMs (see Table D.2 for the number of observations used in publicly available TSFMs). Therefore, it is essential to extend our analysis by scaling the data size and examining how it impacts performance both at the forecasting level and at the portfolio level.

We construct several scaled datasets to address distinct empirical questions. First, we expand the scope of the training data from U.S. excess returns to global excess returns, thereby increasing the total number of observations from approximately 176.96 million to 496.89 million. This expansion enables us to examine how extending the sample from a single-country setting to a cross-country framework encompassing 94 countries affects model performance. Furthermore, a central proposition of TSFMs is their ability to integrate data observed at heterogeneous frequencies within a unified framework to generate forecasts at any desired frequency. To critically evaluate this claim, we extend our global dataset by incorporating monthly JKP factors, which expands the daily dataset size from 496.89 million to approximately 2 billion observations in the TSFMs pre-trained for the year 2022. Finally, many TSFM implementations increase their pre-training data size through the inclusion of synthetic data. To examine this property, we replace the JKP factors with synthetic variables of identical dimensionality, allowing us to isolate whether changes in model performance are driven by the inclusion of specialized financial factors or merely by the expanded data volume. For each dataset, variation, and year, we pre-train all TSFMs from scratch again. Section 5.2.1 reports the performance of benchmark models pre-trained on globally aggregated data. Section 5.2.2 presents the results of TSFMs that were pre-trained and fine-tuned using scaled datasets. Section 5.2.3 extends the analysis by tracking the portfolio performance of various models over time. Section 5.2.4 summarizes the results after accounting for transaction costs. Finally, Section 5.2.5 investigates how hyperparameter tuning affects the performance of these scaled TSFMs.

5.2.1 Scaled Benchmark Results

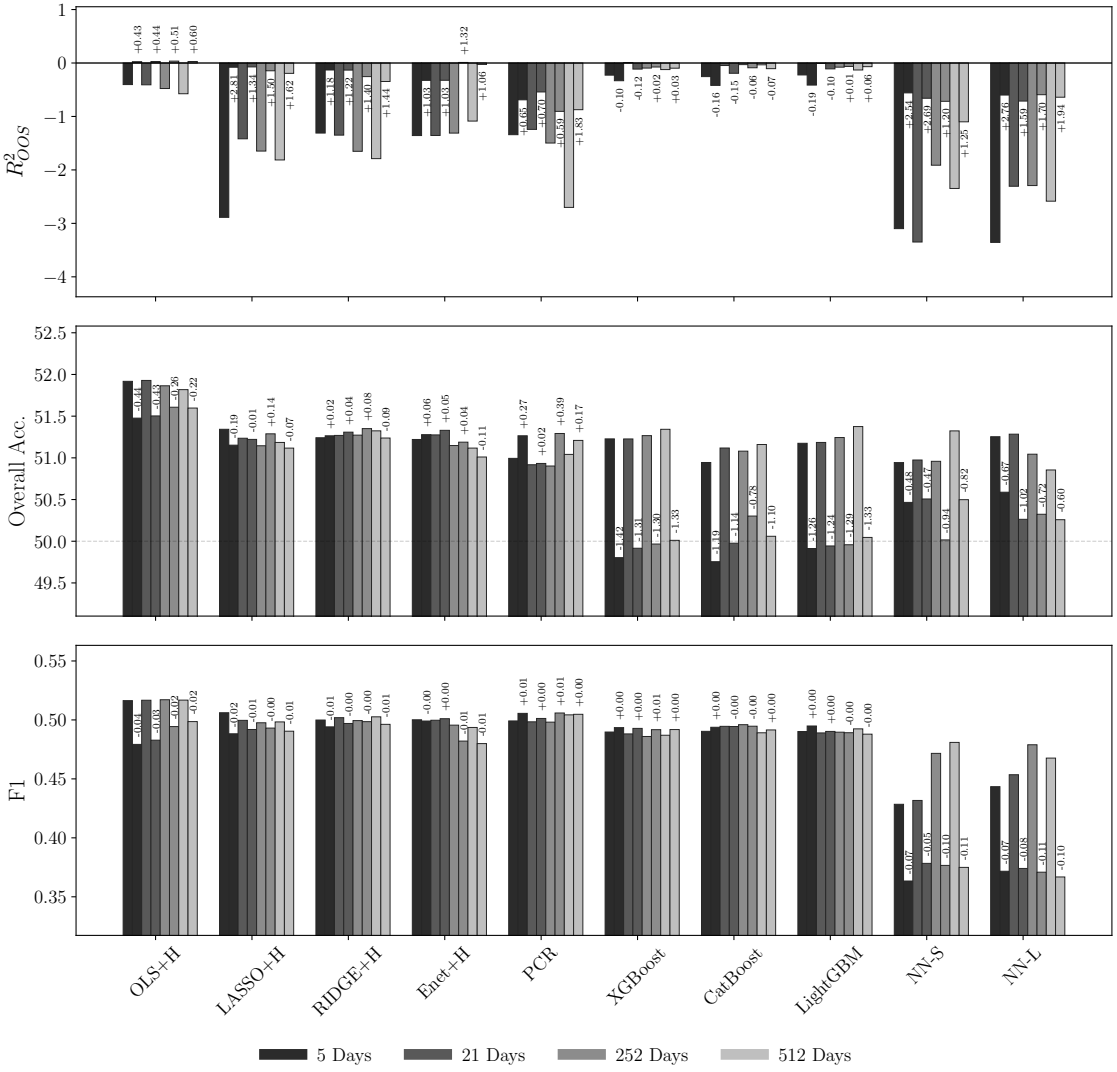
Before analyzing the TSFM results, we first assess how data scaling affects the forecasting and portfolio performance of the benchmark models. This step establishes a baseline for understanding the impact of using scaled (global) data and ensures a fair comparison with the TSFMs. Specifically, we replace the U.S. excess return data with the global excess returns and re-train all benchmark models. For each

model, we train 22 yearly models using the same hyperparameter search environment in Table C.1, which are then employed to generate forecasts of U.S. excess returns for the corresponding year. This procedure ensures direct comparability between the scaled results presented here and those reported in Section 5.1.1. Figure 6 compares forecasting performance metrics (R^2_{OOS}, overall accuracy, and F1) for benchmark models under two regimes: U.S.-only (left bar), and global (right bar) data. Bars of the same color correspond to the same rolling training window size (5, 21, 252, and 512 trading days). Each duplet of adjacent bars illustrates the performance change attributable to expanding the training data from U.S.-only to global. 'Overall Acc.' denotes overall directional accuracy, and 'F1' refers to the macro-averaged F1 score. In the middle panel, the horizontal line indicates the 50% overall accuracy. Figure 7 compares portfolio performance metrics (annualized return, standard deviation, and Sharpe ratio) for the same set of benchmark models and follows the same presentation structure.

Figure 6 shows that expanding the training data from U.S.-only to global excess returns leads to mixed effects on benchmark model performance. For R^2_{OOS}, a general improvement is observable, except for models such as CatBoost. This improvement is particularly pronounced for the OLS model, where R^2_{OOS} shifts from negative to positive values across all window sizes. Across all models, the average changes for the 5-, 21-, 252-, and 512-day window sizes are 1.10%, 0.86%, 0.82%, and 0.98%, respectively. Directional accuracy, however, follows a largely negative pattern, with most models showing small declines that bring performance closer to the 50% threshold. The average changes in overall accuracy are -0.53%, -0.55%, -0.46%, and -0.55% for the 5-, 21-, 252-, and 512-day window sizes, respectively. This suggests that expanding the dataset to a global context weakens the models' ability to capture consistent directional signals, potentially due to differences in market structures, and regional dynamics. Finally, the F1 scores remain largely unchanged or slightly lower across all models. Overall, these results indicate that while global data may offer marginal informational benefits for certain models, scaling the dataset tends to reduce overall forecasting performance.

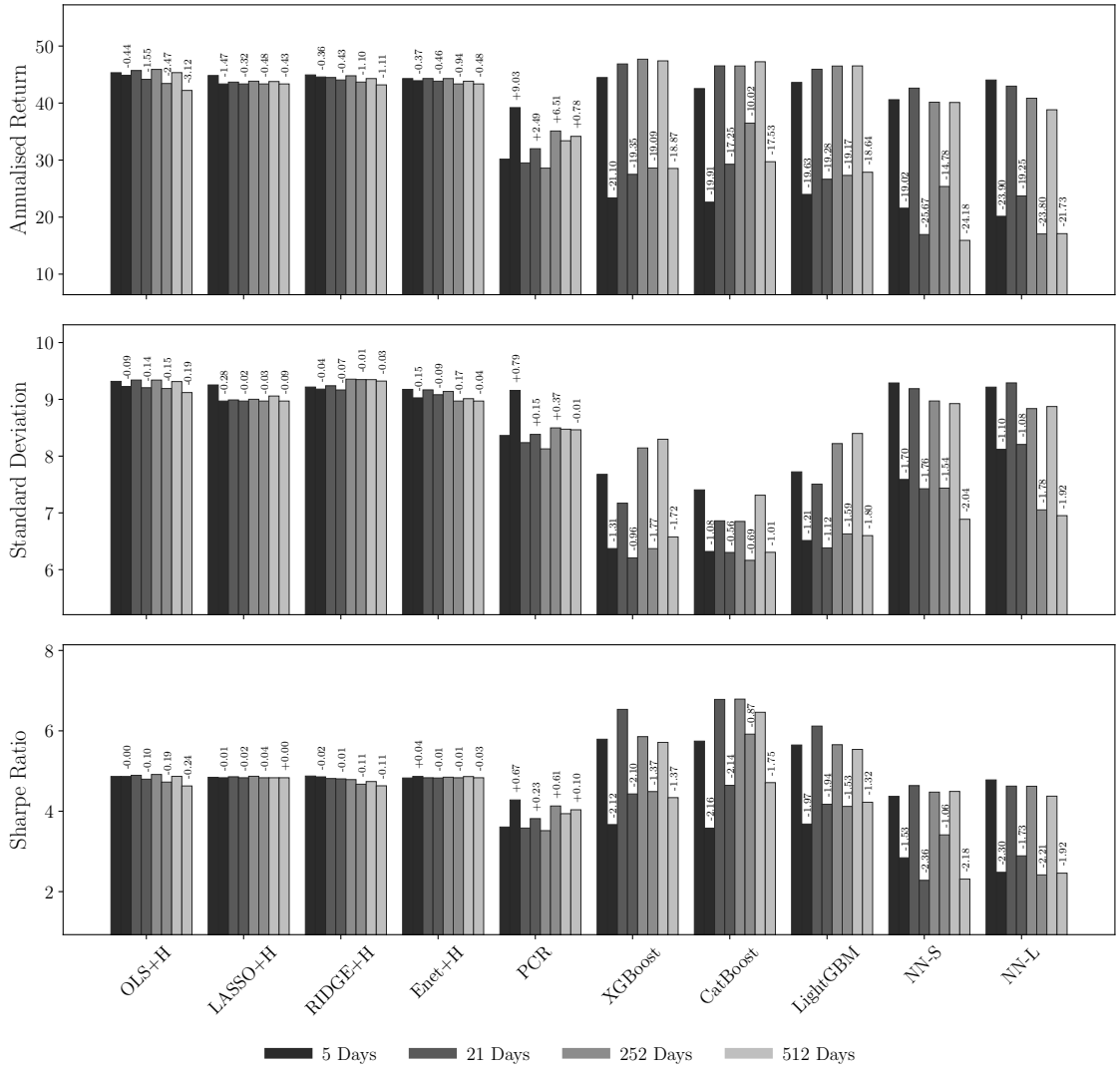
Turning to portfolio performance, Figure 7 indicates that expanding the training data from U.S.-only to global excess returns generally leads to weaker outcomes across benchmark models. Annualized returns decline for nearly all models. Across all models, the average change of annualized returns for the 5-, 21-, 252-, and 512-day window sizes are -9.72%, -10.11%, -8.53%, and -10.53%, respectively. For CatBoost, identified as the best-performing model, the observed reductions are 19.91%, 17.25%, 10.02%, and 17.53% for the respective window sizes. Although portfolio standard deviations decrease slightly, Sharpe ratios also deteriorate for most models, with only marginal or isolated exceptions. On average across all models, Sharpe ratios drop by -0.94, -1.02, -0.68, and -0.88 for the respective window

Figure 6: Impact of Global Data on Forecasting Performance (Benchmarks)



Note: This figure compares forecasting performance metrics (R^2_{OOS}, overall accuracy, and F1) for benchmark models under two regimes: U.S.-only (left bar), and global (right bar) data. Bars of the same color correspond to the same rolling training window size (5, 21, 252, and 512 trading days). Each duplet of adjacent bars illustrates the performance change attributable to expanding the training data from U.S.-only to Global. Benchmark models include linear (OLS, Lasso, Ridge, Elastic Net, and PCR), ensemble (XGBoost, CatBoost, and LightGBM), and neural network (NN-S and NN-L) models. 'Overall Acc.' denotes overall directional accuracy, and 'F1' refers to the macro-averaged F1 score. In the middle panel, the horizontal line indicates the 50% overall accuracy. 'H' indicates that the model is estimated using the Huber loss.

Figure 7: Impact of Global Data on Portfolio Performance (Benchmarks)



Note: This figure compares Portfolio performance metrics (annualized return, standard deviation, and Sharpe ratio) for benchmark models under two regimes: U.S.-only (left bar), and global (right bar). Bars of the same color correspond to the same rolling training window size (5, 21, 252, and 512 trading days). Each duplet of adjacent bars illustrates the performance change attributable to expanding the training data from U.S.-only to Global. Benchmark models include linear (OLS, Lasso, Ridge, Elastic Net, and PCR), ensemble (XGBoost, CatBoost, and LightGBM), and neural network (NN-S and NN-L) models. 'H' indicates that the model is estimated using the Huber loss.

sizes. For CatBoost, the changes in Sharpe ratio correspond to -2.16, -2.14, -0.87, and -1.75. Overall, these results imply that while scaling the dataset expands informational scope, it does not necessarily improve portfolio performance for benchmark models. The inclusion of heterogeneous global data appears to introduce additional noise and cross-market inconsistencies that dilute the strength of predictive signals.[31] [32]

5.2.2 Scaled TSFM Results

We examine how scaling the data influences the performance of TSFMs. First, we expand the pre-training data from U.S. to global dataset. Since TSFMs are designed to handle pre-training on data with mixed frequencies, the next experiment extends the dataset further to include both global data and monthly JKP factors. Finally, we pre-train all models after replacing the JKP factors with synthetic data of identical dimensionality. Figure 8 presents the results for forecasting performance, while Figure 9 reports the corresponding portfolio performance. Here, we have four pre-training regimes: U.S.-only as the left bar in each set (this corresponds to the results presented in Section 5.1.4), global as the second, JKP-augmented as the third, and synthetic-augmented as the fourth. Each group of adjacent bars depicts the change in performance resulting from expansions of the pre-training data, progressing from U.S.-only to global, and subsequently to JKP- and synthetic-augmented variants. JKP factors used in the augmented data are defined in Jensen et al. (2023), and the synthetic data is generated following Ansari et al. (2024).[33] The benchmark group presents the results of the CatBoost

[31] It should be noted that all benchmark models, model complexities, and hyperparameter search environments are kept consistent with those used in Section 5.1.1, ensuring a fair and controlled comparison between the results reported here and those based on U.S.-only data. Although it is possible to further tune each model specifically for the larger and more heterogeneous global dataset, such optimization lies beyond the scope of this study. The objective is to isolate the effect of data scaling itself, rather than to maximize the performance of individual models through additional calibration. This design choice allows for a direct assessment of how increasing the data scale while holding model configurations constant affects both forecasting and portfolio performance, reflecting the performance of each model in an out-of-the-box setting with minimal modification.

[32] This question may raise the concern that larger neural networks could achieve better performance by capturing more complex nonlinearities. To assess this, we extended the network architecture to 128 and 512 hidden units while keeping all other hyperparameters (activation function, optimizer, learning rate, and regularization) unchanged. For ease of comparison, we also report the results for the smaller architectures with 8 and 32 units (NN-S and NN-L), which are the models used throughout this study. The forecasting, portfolio, and spread portfolio results are presented in Table B.13, Table B.14, and Table B.15, respectively. Comparing across model architectures, we find that increasing the number of hidden units beyond 32 yields mixed results. R^2_{OOS} improves slightly, particularly for longest window size and for U.S. data rather than global data, consistent with modest gains in overall accuracy and F1 scores. Similarly, portfolio performance metrics remain largely consistent across the 8-, 32-, 128-, and 512-unit models for both the U.S. and global samples. Although the Sharpe ratio occasionally improves with increased model size, these gains are not monotonic, as performance sometimes declines with higher complexity. Also, in all cases, performance remains below that of the ensemble models. Overall, these findings further highlight the superiority of the ensemble models used as benchmark throughout this study.

[33] Following Ansari et al. (2024), synthetic series are drawn from a zero mean Gaussian process prior on a fixed uniform grid. For each series a small set of base kernels is sampled with replacement from a bank that includes periodic, linear, radial basis, rational quadratic, white noise, and constant components. These kernels are composed by repeatedly applying random sums or products to form a covariance function, from which a single trajectory is sampled.

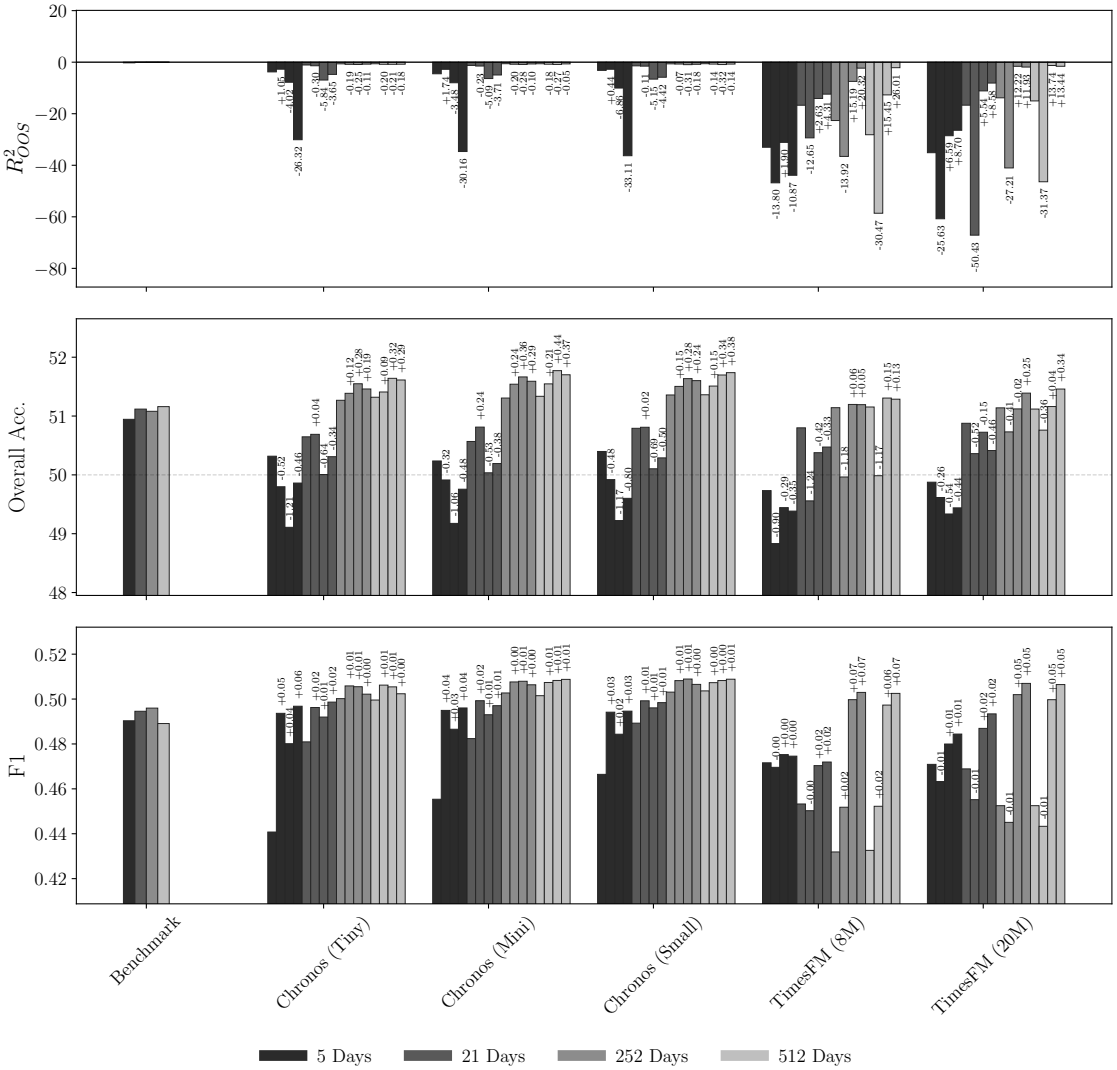
model, trained on U.S. data.[34]

When expanding the dataset from U.S. to global coverage, we observe distinct shifts in both forecasting accuracy and portfolio-level performance. Chronos and TimesFM models generally exhibit a decline in R^2_{OOS} values, suggesting a loss of explanatory power when confronted with more heterogeneous international data. For Chronos, the average changes are 1.07%, -0.21%, -0.15%, and -0.17% for the 5-, 21-, 252-, and 512-day window sizes, respectively. In contrast, the corresponding average changes for the TimesFM models are -19.72%, -31.54%, -20.57%, and -30.92%. An examination of the classification-based metrics indicates that both overall accuracy and F1 scores improve for Chronos, whereas no such improvement is observed for TimesFM models. Averaged across models, the changes in overall accuracy for Chronos are -0.44%, 0.10%, 0.17%, 0.15% for the four window sizes repectively. In contrast, the corresponding changes for TimesFM are -0.58%, -0.88%, -0.79%, -0.76%. The F1 scores exhibit the same pattern as well. This indicates that, despite the decline in explanatory precision for Chronos, the directional accuracy improves in the global setting. From the portfolio performance perspective in Figure 9, the move to global data yields higher annualized returns and Sharpe ratios across Chronos models but not for TimesFM models. For Chronos, the average changes in annualized returns are -4.06%, 5.90%, 6.34%, and 5.73%, while the corresponding changes in standard deviations are -0.41%, -1.15%, -0.08%, and 0.05%. These yield changes in Sharpe ratios of -0.88, 1.67, 0.97, and 0.80 for the 5-, 21-, 252-, and 512-day window sizes, respectively. More specifically, focusing on the Chronos model's Sharpe ratio with a window size of 512, increases of 1.56, 1.05, and 1.36 are observed for the tiny, mini, and small variants, respectively. In all cases, these improvements are associated with higher annualized returns and concurrently lower standard deviations. For TimesFM, although some improvements in standard deviation are clear, the average changes in Sharpe ratios are -0.71, -0.41, 2.76, and -1.63 for the respective window sizes. These findings, particularly for the Chronos models, contrast with earlier results in Section 5.2.1, where data scaling generally had a negative impact on benchmark models. The reversal observed here may stem from the substantially larger number of parameters and greater architectural complexity of TSFMs, which enable them to exploit the richer global data, an effect analogous to performance scaling trends observed in LLMs. These changes are most pronounced for longer window sizes, aligning with the findings in Section 5.1.4, which show that TSFMs tend to outperform when longer input windows are used.

Although the change from U.S. to global coverage is considerable, the major increase in data size occurs when we extend the global excess data with JKP factors. For the Chronos models, R^2_{OOS} values

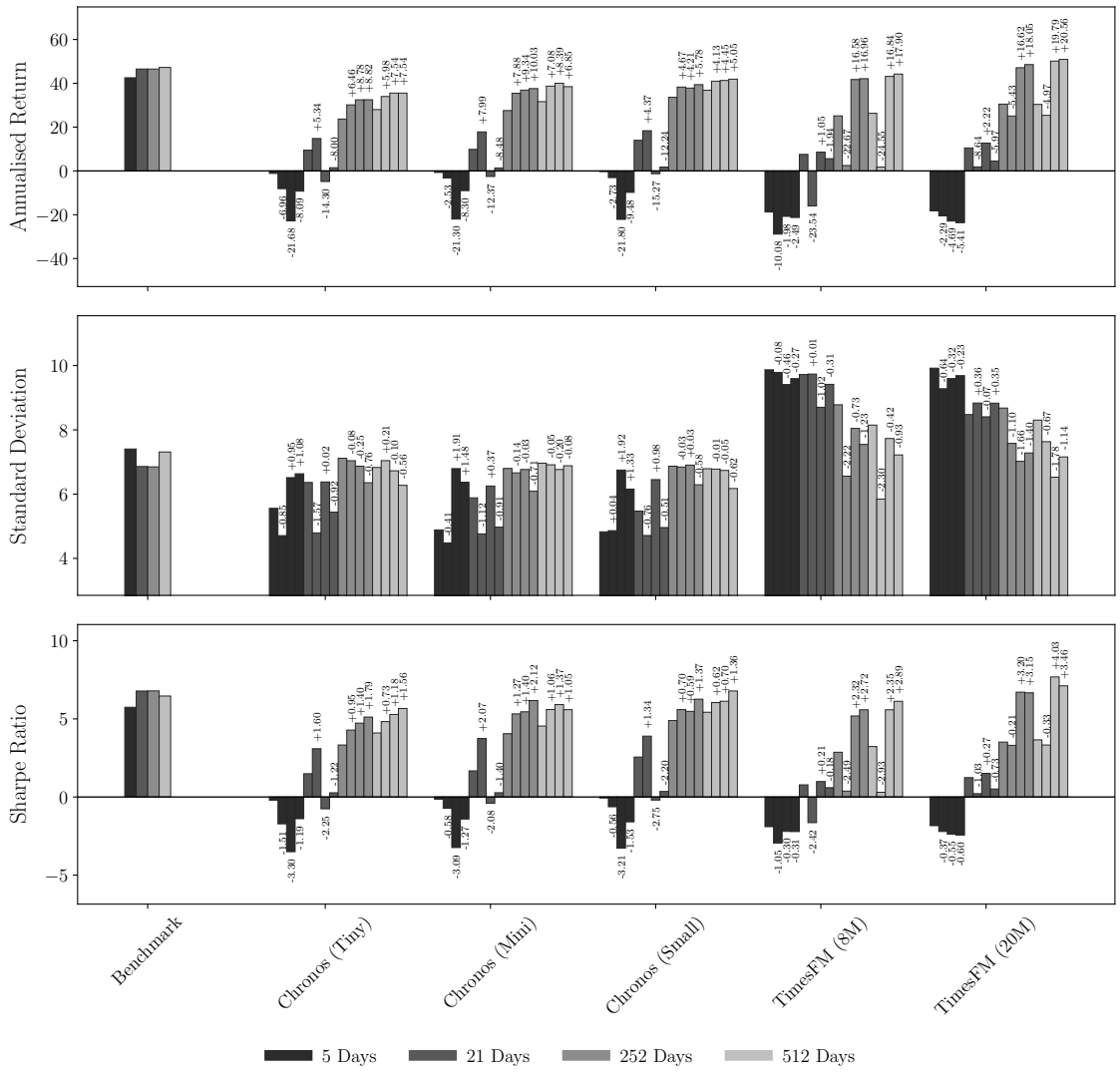
[34]We report the benchmark model (CatBoost) trained on U.S. data rather than global data, as Section 5.2.1 shows that expanding the dataset from U.S. to global generally results in a decline in the performance of benchmark models. Therefore, we retain this best-performing model for comparison.

Figure 8: Impact of Global and Augmented Pre-Training on Forecasting Performance



Note: This figure compares forecasting performance metrics (R^2_{OOS}, overall accuracy, and F1) for five models under four pre-training regimes: U.S.-only as the left bar in each set, Global as the second, JKP-augmented as the third, and Synthetic-augmented as the fourth. Bars of the same color correspond to the same rolling training window size (5, 21, 252, and 512 trading days). Each group of adjacent bars depicts the change in performance resulting from expansions of the pre-training data, progressing from U.S.-only to Global, and subsequently to JKP- and synthetic-augmented variants. Models include Chronos (tiny, mini, small, base, and large) and TimesFM (with 8 million and 20 million parameters). JKP factors used in the augmented data are defined in Jensen et al. (2023), and the synthetic data is generated following Ansari et al. (2024). The benchmark group presents the results of the CatBoost model, trained on U.S. data. 'Overall Acc.' denotes overall directional accuracy, and 'F1' refers to the macro-averaged F1 score. In the middle panel, the horizontal line indicates the 50% overall accuracy.

Figure 9: Impact of Global and Augmented Pre-Training on Portfolio Performance



Note: This figure compares portfolio performance metrics (annualized return, standard deviation, and Sharpe ratio) for five models under four pre-training regimes: U.S.-only as the left bar in each set, Global as the second, JKP-augmented as the third, and Synthetic-Augmented as the fourth. Bars of the same color correspond to the same rolling training window size (5, 21, 252, and 512 trading days). Each group of adjacent bars depicts the change in performance resulting from expansions of the pre-training data, progressing from U.S.-only to Global, and subsequently to JKP- and synthetic-augmented variants. Models include Chronos (tiny, mini, and small) and TimesFM (with 8 million and 20 million parameters). JKP factors used in the augmented data are defined in Jensen et al. (2023), and the synthetic data is generated following Ansari et al. (2024). The benchmark group presents the results of the CatBoost model, trained on U.S. data.

generally decline, whereas the TimesFM models exhibit improved R^2_{OOS}. Classification-based metrics, including accuracy and F1 scores, also show improvements across most models. For Chronos, the average changes in R^2_{OOS} are -4.79%, -5.36%, -0.28%, and -0.27%, and in overall accuracy are -1.15%, -0.62%, 0.31%, and 0.36% for the 5-, 21-, 252-, and 512-day windows, respectively. For TimesFM, the corresponding changes are 4.24%, 4.09%, 13.71%, and 14.60% for R^2_{OOS} and -0.41%, -0.29%, 0.02%, and 0.10% for overall accuracy. The changes in F1 scores follow the same pattern, with TimesFM benefiting more than Chronos from data scaling. From a portfolio standpoint, the JKP-augmented data produces generally positive outcomes. TimesFM models show substantial improvements in performance with this larger dataset, delivering higher annualized returns and Sharpe ratios relative to both the U.S. and global configurations, while Chronos models experience improvements as well. Focusing on the average changes in Sharpe ratio, the Chronos models exhibit shifts of -3.20, -2.36, 1.17, and 1.08, while the TimesFM models show corresponding changes of -0.43, 0.24, 2.76, and 3.19 across the 5-, 21-, 252-, and 512-day windows, respectively, which result from generally higher annualized returns and lower standard deviations. Notably, TSFMs further consolidate its lead over the benchmark model. As an example, for a window size of 512, the TimesFM models with 8M and 20M parameters achieve increases in Sharpe ratio of 2.35 and 4.03, respectively, resulting in Sharpe ratios of 5.58 and 7.69. The larger model thus exceeds the benchmark model's Sharpe ratio of 6.46 under the same window size. This highlights a distinctive difference between conventional models and TSFMs: even with the reduced sizes used here, TSFMs require substantially more data to reach or surpass benchmark performance.[35]

This improvement in performance raises an important question: does the inclusion of these factors, even when it differs in frequency from the primary dataset, provide useful signals that TSFMs can effectively capture? This idea is analogous to the principle underlying LLMs, where exposure to text from diverse topics and sources enables the model to learn from the heterogeneous nature of the pre-training data. To examine this hypothesis, we employ synthetically augmented data in which the JKP factors are replaced with synthetic variables of identical dimensionality. Analysis of Figure 8 reveals that replacing the JKP factors with synthetic variables does not generally degrade performance; in fact, it often leads to improvements across several forecasting metrics. For instance, for the TimesFM (20M) model with a window size of 512, we observe a slightly smaller increase in R^2_{OOS} (13.44% vs.

[35]We also examine how scaling the fine-tuning data affects the performance of TSFMs. Figure B.5 and Figure B.6 illustrate the effects of expanding the fine-tuning dataset from U.S.-only to global and subsequently to JKP-augmented configurations. Moving from U.S. to global fine-tuning and further to JKP-augmented fine-tuning, performance gains are primarily observed for the TimesFM models, most notably in terms of R^2_{OOS}. While TimesFM exhibits measurable improvements, the Chronos models generally show little change or mild degradation in forecasting performance. Portfolio-level outcomes reflect a similar pattern: Sharpe ratios show improvements across models but remain negative in most configurations, suggesting that general-purpose pre-trained TSFMs still struggle to produce economically meaningful forecasts, even after fine-tuning on larger datasets. These findings are consistent with the results reported in Section 5.1.3.

13.74%), a larger gain in directional accuracy (0.34% vs. 0.04%), and an approximately equivalent improvement in the F1 score (0.05). For the Chronos (small) model using the same window size, we observe a slightly smaller decrease in R^2_{OOS} (-0.14% vs. -0.32%), a larger gain in directional accuracy (0.38% vs. 0.34%), and a modest improvement in the F1 score (0.01 vs. 0.00). A similar pattern is also observable for other TSFMs, particularly for longer window sizes, which generally exhibit comparable performance or even improvements when moving from JKP-augmented to synthetic-augmented data.[36]

Turning to the portfolio performance results in Figure 9, we again observe that pre-trained models with JKP and synthetic-augmented data perform comparably, and in some cases, the synthetic data even enhances portfolio performance. Among the TSFMs demonstrating the greatest improvement, the Chronos (small) model with window size of 512 exhibits an increase in its Sharpe ratio from 0.70 to 1.36 when transitioning from the JKP-augmented dataset to the synthetic-augmented dataset. This enhancement is primarily driven by higher annualized returns and a lower standard deviation. For the TimesFM (8M) and TimesFM (20M) models with the same window size, the Sharpe ratios change from 4.03 to 3.46 and from 2.35 to 2.89, respectively, indicating comparable or improved performance. A similar pattern is generally observed across other model sizes and window sizes. These findings suggest that the observed performance gains may not arise solely from meaningful regional signals embedded in the JKP factors. Instead, they may stem from the structural advantages of incorporating broader and more heterogeneous input representations, regardless of whether the additional data originate from real or synthetic sources. Overall, the superior performance achieved with synthetic augmentation indicates that for large-scale TSFMs, the diversity and scale of the input data play a dominant role in enhancing performance rather than the precise alignment of auxiliary data with the main dataset. It is also worth noting that most publicly available pre-trained TSFMs already incorporate synthetic data to expand their pre-training sample size. Therefore, the limited performance of many existing models likely reflects constraints in the quality and domain relevance of their core real-world training data. Replacing generic mixed-domain datasets with more specialized, task-focused data, while continuing to extend them through synthetic augmentation, could further improve model performance, especially given the scarcity of large, high-quality, domain-specific time series datasets for pre-training such large-scale models in finance.

While Kelly et al. (2024) document a 'virtue of complexity' in return prediction, often interpreted

[36]Table B.16 reports the modified DM test results. The table compares pre-trained TSFMs under two augmentation regimes: JKP-augmented (no superscript) and synthetic-augmented (models with a superscript *). Synthetic-augmented variants are broadly comparable to, and in several cases outperform, their JKP-augmented counterparts. Across both models, the synthetic-augmented variant generally exhibits superior performance relative to the JKP-augmented variant when evaluated at the same model size. Moreover, synthetic-augmented models often outperform JKP-augmented models of the same type even when the latter employ a larger scale. However, the benchmark model still demonstrates significantly superior performance compared to even the augmented TSFMs.

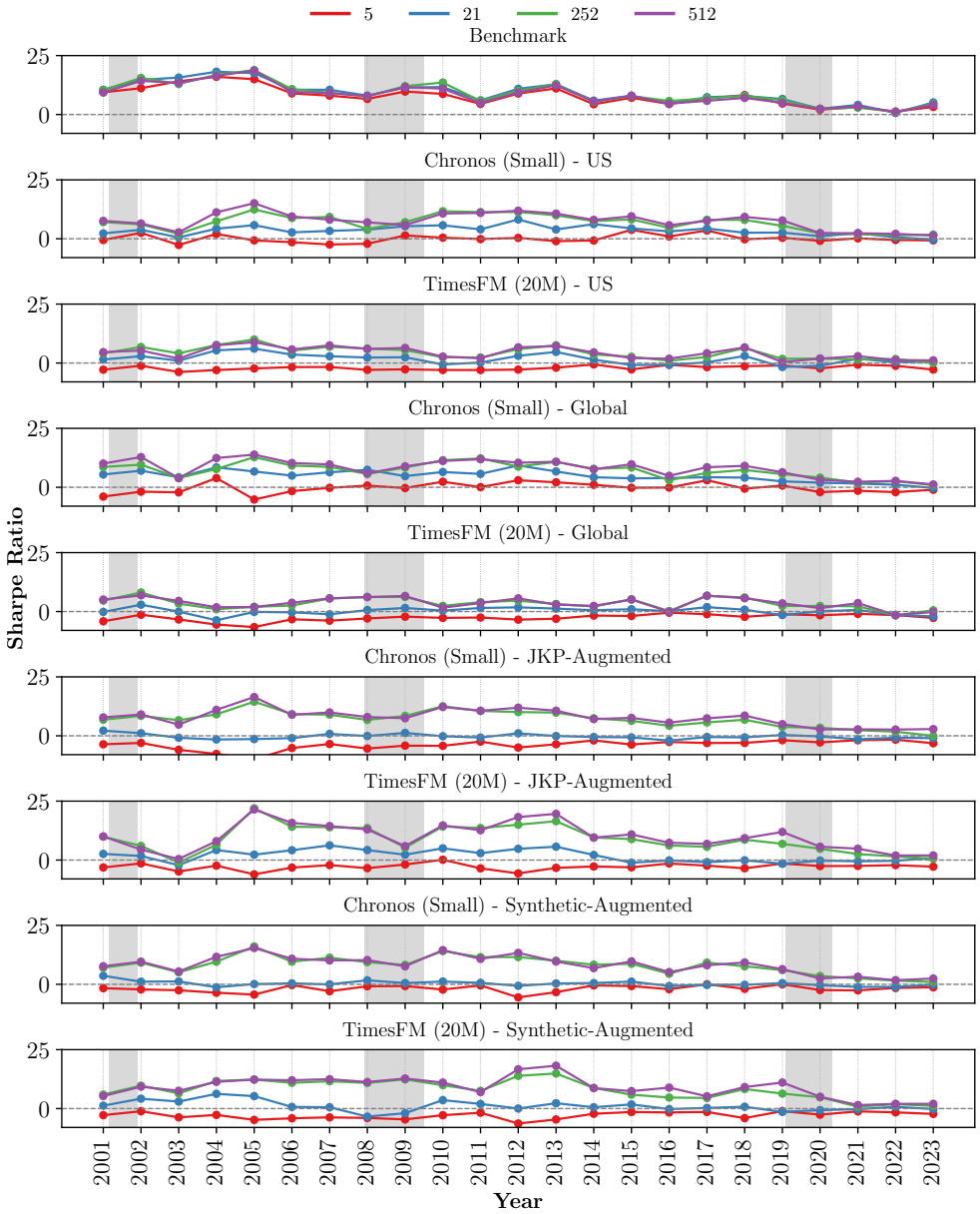
as being chiefly compute or resource limited, our evidence adds a complementary dimension: *data*. Even after training benchmark models and pre-training TSFMs on the entire in-domain excess-return data available at the global level, we find that adding more observations measurably improves both forecasting accuracy and portfolio performance, especially for models that are substantially larger than conventional architectures. This pattern is consistent with compute–data scaling results in Transformer models, where optimal capacity must rise with training tokens (Hoffmann et al., 2022). Yet performance remains distinctly data limited: when moving to substantially larger models, such as TSFMs, and parameter counts outpace the effective sample, marginal returns to complexity flatten. In short, architecture and compute are necessary but insufficient; the frontier for realizing the promised gains from complexity is defined by the scale and alignment of finance-native data. Synthetic data can offer partial relief, but it is no substitute for additional real, task-specific information.

5.2.3 Tracing Portfolio Performance over Time

So far, all presented results have focused on the overall performance of different forecasting models across a broad out-of-sample period from 2001 to 2023. While these aggregate statistics are informative, they do not capture potential temporal variation in model performance. Portfolio profitability can vary substantially over time, particularly in response to changes in market regimes, volatility, or macroeconomic conditions. Therefore, it is crucial to examine how model-driven portfolios evolve through time. To address this, Figure 10 reports yearly Sharpe ratios over the out-of-sample period. Results are presented by model across multiple window sizes. The set of models includes a benchmark, the CatBoost model, that is trained using U.S. excess returns, as well as the largest TSFMs: Chronos and TimesFM with parameter sizes of small and 20M, respectively, both pre-trained from scratch. We also report results for TSFMs pre-trained on different datasets: U.S.-only excess returns, global excess returns, global excess returns augmented with JKP factors, and global excess returns augmented with synthetic data. Shaded regions correspond to U.S. recession periods as identified by the NBER.

The cross-model annual average Sharpe ratio declines materially over time, indicating a broad reduction in risk-adjusted profitability regardless of model complexity. The benchmark also exhibits a clear and persistent negative trend across all window sizes. However, its relative performance remains strongest at shorter horizons (5 to 21 days), where it often matches or exceeds the TSFMs, reflecting its ability to adapt quickly to short-term fluctuations. In contrast, the more complex TSFMs display greater resilience over medium to long window sizes (252 to 512 days), where their performance declines more moderately and often surpasses that of the benchmark. This pattern suggests that while simpler models remain effective for short-term prediction, higher model complexity and architectures designed

Figure 10: Yearly Sharpe Ratios of Long–Short Portfolios



Note: The figure reports annual Sharpe ratios from 2001 to 2023 for long–short equal-weighted portfolios. Results are presented by model across multiple window sizes (5, 21, 252, and 512 trading days). The set of models includes a benchmark, the CatBoost model, that is trained using U.S. excess returns, as well as the largest time series foundation models (TSFMs): Chronos and TimesFM (with parameter sizes of small and 20M, respectively), both trained from scratch. We also report results for TSFMs pre-trained on different datasets: U.S.-only excess returns, global excess returns, global excess returns augmented with JKP factors (as defined in Jensen et al. (2023), denoted Global-Augmented), and global excess returns augmented with synthetic data of the same size as the JKP factors (as defined in Ansari et al. (2024), denoted Synthetic-Augmented). Shaded regions correspond to U.S. recession periods as identified by the National Bureau of Economic Research (NBER).

to capture long-term temporal dependencies help preserve predictive power when short-term signals weaken. During periods of market stress, such as the 2008 financial crisis and the 2020 COVID-19 shock, TSFMs exhibit relatively higher Sharpe ratios compared to the benchmark model. For instance, during the 2008 financial crisis, the best-performing TSFM model, TimesFM (20M) pre-trained with synthetic-augmented data, achieves average Sharpe ratios of 11.63 and 12.01 for window sizes of 252 and 512, respectively, exceeding the corresponding benchmark model ratios of 9.88 and 9.74. During the 2020 COVID-19 shock, the same TimesFM configuration attains Sharpe ratios of 4.81 and 5.02, again surpassing the benchmark model's performance, which records values of 2.44 and 2.40 for the same window sizes. Also, a comparison across TSFMs also reveals that models pre-trained on larger and more diverse datasets, such as global or augmented, generally outperform those pre-trained on narrower U.S.-only data, particularly at longer horizons. Nonetheless, no class of models is immune to the overall downward trend in Sharpe ratio, emphasizing that increasing complexity or data scope mitigates but does not eliminate performance degradation.[37]

5.2.4 Transaction Cost

Table 12 reports annualized Sharpe ratios of equal-weighted long–short portfolios formed using daily forecasts from different forecasting models across rolling window sizes of 5, 21, 252, and 512 trading days. Performance is evaluated net of transaction costs under four scenarios: no costs (0 bps), fixed costs of 20 bps and 40 bps, and a mixed-cost structure where small- and large-cap stocks face distinct estimated trading costs following Frazzini et al. (2012). In the mixed specification, the estimated costs correspond approximately to 21.3 bps for small-cap stocks and 11.2 bps for large-cap stocks. All TSFMs are pre-trained from scratch. The top panel reports results using only U.S. excess return data for pre-training; the second panel uses global data; the third panel combines global data with JKP factors; and the bottom panel combines global data with synthetic data. While earlier tables demonstrated that pre-trained TSFMs deliver strong performance in frictionless environments, real-world portfolio profitability ultimately depends on whether such gains persist once trading costs are introduced. Because our implementation trades daily and spans a broad cross-section of U.S. stocks, implying high turnover and substantial market-wide trading costs, transaction costs have a significant impact on profitability. Therefore, our objective here is to identify which models remain more resilient under varying levels of transaction costs.

[37] We also report the annual Sharpe ratio results for all models evaluated in Section 5.1.1, trained on U.S. excess returns. The corresponding outcomes are presented in Figure B.7. The findings indicate that all models exhibit a similar pattern of performance deterioration over time, consistent with the benchmark model, CatBoost, although some deviations arise in specific years. The superior performance of ensemble models also remains evident. Moreover, regardless of the model employed, the general upward or downward trends appear consistent. While the choice of model does not typically influence the direction of performance changes, it does affect the magnitude of the resulting Sharpe ratios.

As shown in Table 12, net performance deteriorates sharply once trading frictions are introduced. Moving from 0 bps to 20 bps (and beyond) drives all long–short strategies to negative Sharpe ratios across window sizes, regardless of model class. For example, for the benchmark model, the average Sharpe ratio across the four window sizes falls from about 6.44 at 0 bps to –3.62 at 20 bps and –13.59 at 40 bps, so even modest frictions flip performance from strongly positive to strongly negative. Measured by the extent to which the Sharpe ratio remains closer to zero under costs, the most resilient specifications are the larger TSFMs, particularly TimesFM (20M), when pre-trained on scaled datasets that augment global excess returns with either JKP factors or synthetic data. At longer windows (252–512 days), these augmented TimesFM models yield the least negative net Sharpe ratios, outperforming both Chronos variants and the benchmark model. At the highest transaction cost of 40 bps, averaging over the 252- and 512-day windows, TimesFM (20M) with JKP and synthetic augmentation attains average Sharpe ratios of about -11.24 and -11.00, compared with -11.78 for the best Chronos model (Chronos small with global pre-training), -12.87 for global-only TimesFM (20M), and -13.55 for the benchmark. Under the mixed-cost specification, the corresponding long-window averages are roughly -1.91 and -1.90 for the JKP and synthetic-augmented TimesFM (20M) models, versus about -2.94 for the best Chronos model, -4.63 for global-only TimesFM (20M), and -3.33 for the benchmark. Overall, these results highlight that while trading frictions erode performance across the board, larger TSFMs with enriched pre-training remain comparatively more robust.

Global scaling of the data yields little improvement in cost resilience and in some cases even degrades it, whereas JKP or synthetic augmentation substantially mitigates cost-induced performance decay. Chronos models benefit from scale but remain less robust to transaction costs than augmented TimesFM, while the benchmark model is comparatively more resilient at shorter window sizes (5–21 days) yet lags behind augmented TSFMs at medium to long window sizes. For example, across all TSFM specifications and window sizes, and under the highest transaction cost (40 bps), moving from the benchmark model to TimesFM (20M) pre-trained with synthetic-augmented data improves the Sharpe ratio from -13.53, -13.74, -14.46, and -12.64 to -10.88, -8.84, -10.84, and -11.15 for window sizes of 5, 21, 252, and 512, respectively. For Chronos (small), the corresponding Sharpe ratios are -27.36, -28.00, -13.22, and -13.23, indicating comparatively greater resilience at longer window sizes. Under the mixed-cost structure, TimesFM (20M) attains Sharpe ratios of –6.56, –4.07, –1.93, and –1.87, while Chronos (small) achieves –14.03, –13.28, –3.32, and –3.07 compared with benchmark values of –3.73, –3.33, –3.69, and –2.96, collectively indicating again generally greater relative resilience at longer window sizes. Model size also shapes cost resilience within each pre-trained group of TSFMs. Focusing on the 512-day window, JKP-augmented Chronos exhibits modest gains as model size increases: under

the highest fixed cost of 40 bps, the Sharpe ratio improves from approximately –12.63 (tiny) to –11.90 (mini) and –12.03 (small), while under the mixed-cost structure it rises from roughly –3.54 to –2.87 and –2.86. For JKP-augmented TimesFM, the 8M and 20M variants are nearly identical at the 40 bps cost level (–11.47 vs. –11.65) but diverge under the mixed-cost structure, where performance improves from –2.78 to –1.87. For the synthetic-augmented with the same window size, Chronos again achieves its most resilient specification at the mini scale (–11.95 vs. –12.99 and –13.23 under 40 bps, and –3.03 vs. –3.47 and –3.07 under mixed costs). Meanwhile, TimesFM (20M) consistently outperforms TimesFM (8M), improving from –11.96 to –11.15 at 40 bps and from –2.76 to –1.87 under the mixed-cost structure. Therefore, conditional on a given pre-training scheme, moderate increases in model size tend to enhance cost resilience—particularly for TimesFM and, to a lesser extent, Chronos. Overall, three consistent patterns emerge: (i) trading frictions erase frictionless gains for all models, (ii) resilience increases with both model size and the scale and diversity of pre-training data (JKP or synthetic augmentation), and (iii) these benefits are most pronounced at larger window sizes.[38]

[38]Table B.17 presents the corresponding results for the benchmark models. Consistent with the main findings, all benchmark models exhibit substantial deterioration in Sharpe ratios once transaction costs are incorporated. Under both fixed and mixed-cost scenarios, most models deliver negative risk-adjusted performance across all window sizes. Among the benchmark models, ensemble models demonstrate comparatively greater resilience to trading frictions, linear models show moderate robustness, while neural networks experience the most pronounced declines in performance as costs increase. Also, models trained on U.S. excess return data generally attain slightly higher net Sharpe ratios than their globally trained counterparts, suggesting that broader training universes do not necessarily mitigate the adverse impact of transaction costs.

Table 12: Long–Short Portfolio Sharpe Ratios with Transaction Costs

Model	0 bps				20 bps				40 bps				Mixed			
	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
U.S.																
Benchmark	5.74	6.78	6.79	6.46	-3.94	-3.52	-3.90	-3.13	-13.53	-13.74	-14.46	-12.64	-3.73	-3.33	-3.69	-2.96
Chronos (Tiny)	-0.21	1.49	3.33	4.10	-13.74	-8.39	-4.75	-4.76	-27.24	-18.22	-12.76	-13.57	-13.25	-8.15	-4.66	-4.66
Chronos (Mini)	-0.14	1.67	4.05	4.54	-15.68	-9.20	-4.66	-4.33	-31.17	-20.01	-13.29	-13.14	-15.11	-8.94	-4.57	-4.27
Chronos (Small)	-0.07	2.56	4.89	5.42	-16.17	-9.51	-3.80	-3.72	-32.23	-21.47	-12.42	-12.78	-15.60	-9.22	-3.74	-3.68
TimesFM (8M)	-1.90	0.78	2.86	3.23	-5.84	-3.40	-2.37	-2.43	-9.69	-7.52	-7.53	-7.99	-5.72	-3.27	-2.21	-2.24
TimesFM (20M)	-1.84	1.24	3.51	3.66	-6.10	-4.73	-3.26	-3.41	-10.31	-10.54	-9.93	-10.37	-5.98	-4.57	-3.08	-3.22
Global																
Chronos (Tiny)	-1.72	3.09	4.28	4.82	-16.94	-9.98	-3.82	-3.52	-32.06	-22.98	-11.88	-11.84	-16.34	-9.64	-3.71	-3.43
Chronos (Mini)	-0.72	3.74	5.32	5.60	-16.76	-9.67	-3.56	-3.13	-32.77	-23.04	-12.42	-11.80	-16.14	-9.34	-3.46	-3.05
Chronos (Small)	-0.63	3.89	5.59	6.04	-15.78	-9.73	-3.08	-2.94	-30.86	-23.32	-11.70	-11.86	-15.20	-9.42	-3.00	-2.87
TimesFM (8M)	-2.95	-1.64	0.38	0.30	-6.81	-4.26	-3.56	-4.02	-10.59	-6.81	-7.26	-7.95	-6.70	-4.16	-3.42	-3.86
TimesFM (20M)	-2.21	0.21	3.31	3.33	-6.91	-5.73	-4.91	-4.83	-11.54	-11.52	-12.94	-12.80	-6.77	-5.55	-4.66	-4.59
JKP-Augmented																
Chronos (Tiny)	-3.50	-0.76	4.73	5.28	-13.01	-7.85	-3.73	-3.68	-22.40	-14.86	-12.18	-12.63	-12.66	-7.64	-3.58	-3.54
Chronos (Mini)	-3.24	-0.40	5.45	5.91	-12.22	-8.30	-3.24	-2.98	-21.10	-16.09	-11.95	-11.90	-11.89	-8.04	-3.12	-2.87
Chronos (Small)	-3.28	-0.20	5.48	6.13	-12.45	-8.38	-3.15	-2.95	-21.46	-16.43	-11.79	-12.03	-12.12	-8.11	-3.04	-2.86
TimesFM (8M)	-2.21	0.99	5.19	5.58	-6.75	-4.48	-2.82	-2.95	-11.24	-9.84	-10.82	-11.47	-6.62	-4.31	-2.65	-2.78
TimesFM (20M)	-2.39	1.51	6.71	7.68	-6.70	-3.93	-2.09	-2.03	-10.97	-9.29	-10.83	-11.65	-6.58	-3.80	-1.94	-1.87
Synthetic-Augmented																
Chronos (Tiny)	-1.39	0.27	5.12	5.66	-12.57	-11.90	-3.94	-3.65	-23.60	-23.92	-13.01	-12.99	-12.16	-11.34	-3.74	-3.47
Chronos (Mini)	-1.41	0.27	6.17	5.59	-13.53	-13.46	-3.76	-3.17	-25.54	-27.10	-13.74	-11.95	-13.07	-12.86	-3.59	-3.03
Chronos (Small)	-1.59	0.35	6.26	6.78	-14.53	-13.89	-3.47	-3.20	-27.36	-28.00	-13.22	-13.23	-14.03	-13.28	-3.32	-3.07
TimesFM (8M)	-2.22	0.59	5.58	6.12	-6.74	-4.32	-2.85	-2.93	-11.21	-9.14	-11.28	-11.96	-6.61	-4.17	-2.68	-2.76
TimesFM (20M)	-2.44	0.51	6.67	7.11	-6.68	-4.19	-2.10	-2.04	-10.88	-8.84	-10.84	-11.15	-6.56	-4.07	-1.93	-1.87

Note: This table reports annualized Sharpe ratios of equal-weighted long–short portfolios formed using daily forecasts from different forecasting models across rolling window sizes of 5, 21, 252, and 512 trading days. Performance is evaluated net of transaction costs under four scenarios: no costs (0 bps), fixed costs of 20 bps and 40 bps, and a mixed-cost structure where small- and large-cap stocks face different estimated trading costs following Frazzini et al. (2012). In the mixed specification, the estimated costs correspond roughly to 21.3 bps for small-cap stocks and 11.2 bps for large-cap stocks. The benchmark strategy is based on CatBoost, while time series foundation models (TSFMs) include Chronos (tiny, mini, and small) and TimesFM (8M and 20M parameters). All TSFMs are pre-trained from scratch. The top panel reports results using only U.S. excess return data for pre-training; the second panel uses global data; the third panel combines global data with JKP factors as defined in Jensen et al. (2023); and the bottom panel combines global data with synthetic data of the same size as the JKP factors as defined in Ansari et al. (2024).

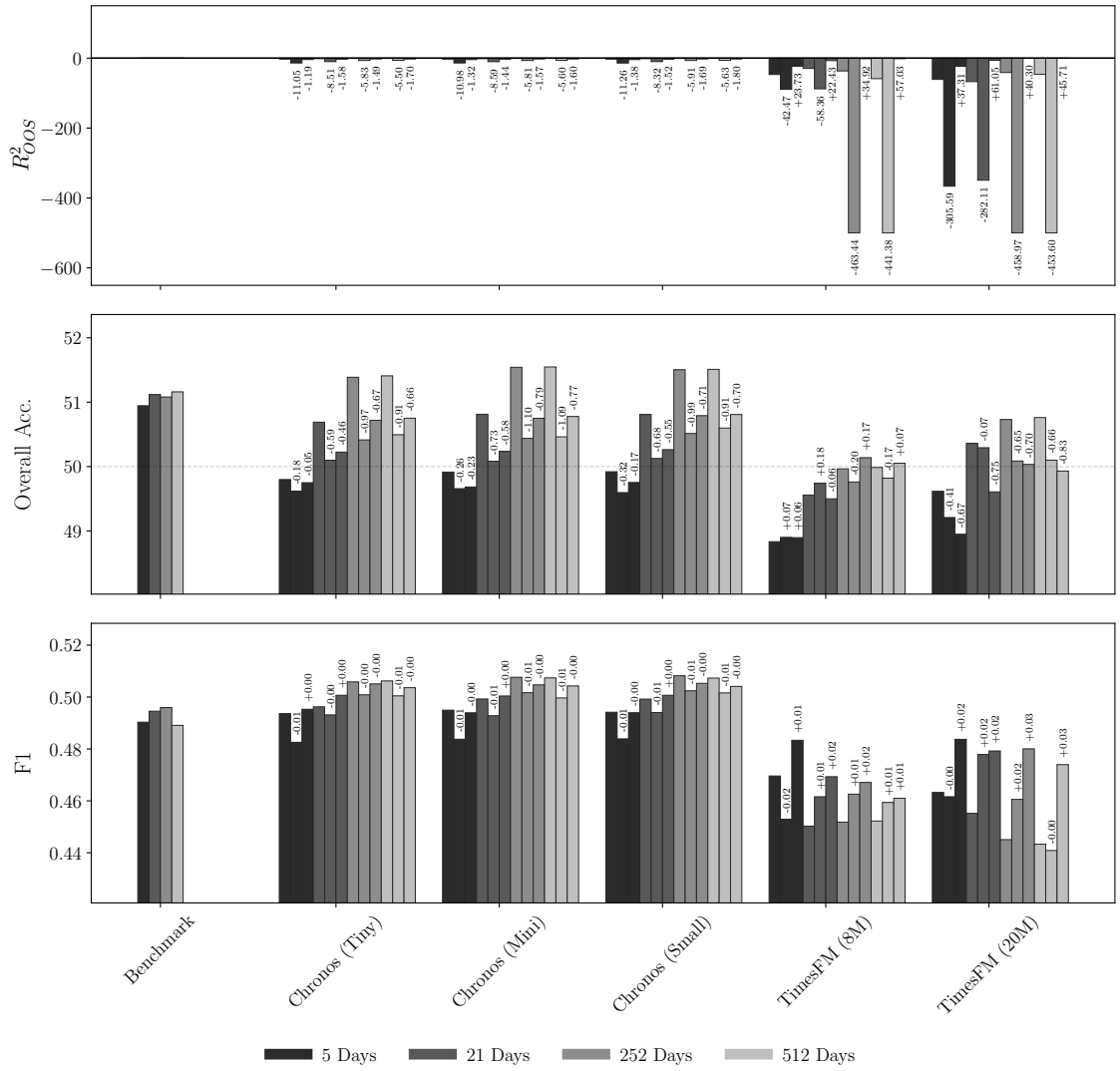
5.2.5 Hyperparameter Impact

So far, all TSFMs have been pre-trained under the default hyperparameter settings specified by the respective authors. In contrast, as discussed in Section 5 and shown in Table C.1, benchmark models underwent explicit hyperparameter tuning during their initial training year, with the resulting configurations reused in subsequent years. This naturally raises the question of how sensitive these large-scale TSFMs are to their hyperparameter configurations. Moreover, many of the original hyperparameter choices were motivated by the goal of developing general-purpose TSFMs pre-trained on substantially large datasets, a context that differs from the more focused setting of this study. Consequently, both the model architecture and, more specifically, the choice of hyperparameters may now emerge as critical factors that warrant systematic evaluation. To investigate this, we conduct a controlled experiment in which the Chronos and TimesFM models are pre-trained under alternative hyperparameter configurations. In this setup, only selected hyperparameters are varied, while all other settings remain fixed at their default values. This design enables us to isolate and quantify the impact of individual hyperparameter choices on overall model performance.

In particular, for Chronos, as described in Section 2.2, we modify the tokenizer's quantization range, comparing the default interval of $[-15, 15]$ with a restricted range of $[-2, 2]$, as well as a dynamic quantization strategy in which the bounds are recalibrated annually to the 5^{th} and 95^{th} percentiles of the pre-training data. For TimesFM, as described in Section 2.3, we alter the input patch length, reducing it from the default value of 32 to 8 in one configuration and expanding it to 128 in another. The resulting outcomes are summarized in Figure 11 and Figure 12, which compare the effects of these hyperparameter adjustments on forecasting and portfolio performance, respectively. For Chronos, the left bar reports results for models pre-trained on global data with the default quantization range. The middle bar presents results obtained by restricting the tokenizer's quantization range from its default interval of $[-15, 15]$ to $[-2, 2]$. The right bar represents a dynamic strategy, where the quantization bounds are recalibrated each year based on the distribution of excess returns. For TimesFM, the left bar again reports results for models pre-trained on global data with default hyperparameters. The middle bar presents results obtained by reducing the default input patch length from 32 to 8, while the right bar shows results when increasing the input patch length to 128. Bars of the same color correspond to the same training window size.

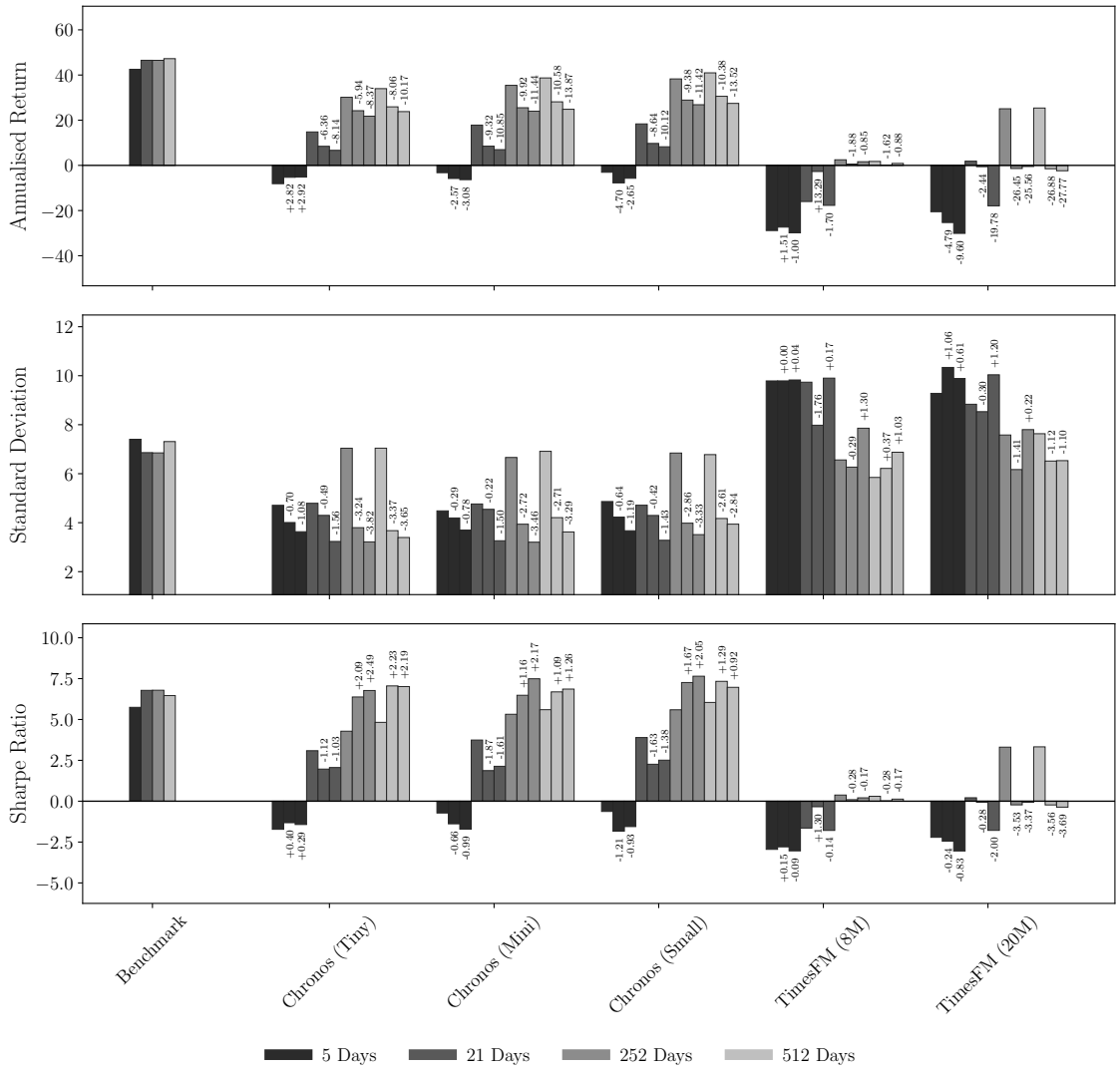
Across both Chronos and TimesFM, the impact of hyperparameter changes is notable. For Chronos, narrowing the quantization range or applying a dynamic quantization strategy results in no improvement in forecasting metrics and, in some cases, a slight degradation relative to the default configu-

Figure 11: Impact of Hyperparameter Choice on Forecasting Performance



Note: This figure compares forecasting performance metrics (R^2_{OOS}, overall accuracy, and F1) for five models under three pre-training regimes. For Chronos, the left bar reports results for models pre-trained on global data with the default quantization range. The middle bar presents results obtained by restricting the tokenizer's quantization range from its default interval of $[-15, 15]$ to $[-2, 2]$, thereby allowing us to evaluate how a tighter binning interval affects model performance. The right bar corresponds to a dynamic strategy, in which the quantization bounds are reset each year to the 5th and 95th percentiles of the excess returns used for pre-training. For TimesFM, the left bar again reports results for models pre-trained on global data with default hyperparameters. The middle bar presents results obtained by reducing the default input patch length from 32 to 8, while the right bar shows results when increasing the input patch length to 128. Bars of the same color correspond to the same rolling training window size (5, 21, 252, and 512 trading days). Models include Chronos (tiny, mini, small, base, and large) and TimesFM (version 1 with 200 million and version 2 with 500 million parameters). The benchmark group presents the results of the CatBoost model, trained on U.S. data, evaluated over trading window sizes of 5, 21, 252, and 512 days, respectively. 'Overall Acc.' denotes overall directional accuracy, and 'F1' refers to the macro-averaged F1 score. In the middle panel, the horizontal line indicates the 50% overall accuracy. To enhance interpretability of the plots, R^2_{OOS} values were truncated at –500 in order to mitigate the influence of extreme negative outliers on the visual scale.

Figure 12: Impact of Hyperparameter Choice on Portfolio Performance



Note: This figure compares portfolio performance metrics (annualized return, standard deviation, and Sharpe ratio) for five models under three pre-training regimes. For Chronos, the left bar reports results for models pre-trained on global data with the default quantization range. The middle bar presents results obtained by restricting the tokenizer's quantization range from its default interval of $[-15, 15]$ to $[-2, 2]$, thereby allowing us to evaluate how a tighter binning interval affects model performance. The right bar corresponds to a dynamic strategy, in which the quantization bounds are reset each year to the 5th and 95th percentiles of the excess returns used for pre-training. For TimesFM, the left bar again reports results for models pre-trained on global data with default hyperparameters. The middle bar presents results obtained by reducing the default input patch length from 32 to 8, while the right bar shows results when increasing the input patch length to 128. Bars of the same color correspond to the same rolling training window size (5, 21, 252, and 512 trading days). Models include Chronos (tiny, mini, small, base, and large) and TimesFM (version 1 with 200 million and version 2 with 500 million parameters). The benchmark group presents the results of the CatBoost model, trained on U.S. data, evaluated over trading window sizes of 5, 21, 252, and 512 days, respectively.

ration. For Chronos, tightening the quantization range leads to average changes in R^2_{OOS} of -3.73%, -8.47%, -5.85%, and -5.58% for window sizes of 5, 21, 252, and 512, respectively. The corresponding changes in overall accuracy are -0.25%, -0.67%, -1.02%, and -0.97%. Similarly, when switching to dynamic quantization, the average changes in R^2_{OOS} are -1.30%, -1.51%, -1.58%, and -1.70%, with associated changes in overall accuracy of -0.15%, -0.53%, -0.72%, and -0.71%. A similar pattern is observed for the F scores, which generally show either no change or a deterioration in performance.

However, the portfolio performance results reveal a clear improvement in Sharpe ratios, particularly for longer window sizes (252 and 512 days). This improvement in the Sharpe ratio is primarily driven by a pronounced decrease in portfolio standard deviation. Tightening the quantization range results in average changes in annualized returns of -1.48%, -8.11%, -8.41%, and -9.67% for window sizes of 5, 21, 252, and 512, respectively. The corresponding changes in standard deviation are -0.54%, -0.38%, -2.94%, and -2.90%, which together yield average changes in the Sharpe ratio of -0.49, -1.54, 1.64, and 1.54. When switching to dynamic quantization, we observe average changes of -0.94%, -9.70%, -10.41%, and -12.52% in annualized returns, -1.02%, -1.50%, -3.54%, and -3.26% in standard deviation, and corresponding changes of -0.54, -1.34, 2.24, and 1.46 in the average Sharpe ratio. As an example, within the Chronos model using a 512-day window, tightening the quantization range yields Sharpe ratio improvements of 2.23, 1.09, and 1.29, while the dynamic quantization setup produces corresponding improvements of 2.19, 1.26, and 0.92, respectively, for the tiny, mini, and small model variants. These gains suggest that although tighter or adaptive quantization does not enhance predictive accuracy, it contributes to economically meaningful portfolio outcomes, with Sharpe ratios in several cases exceeding those of the benchmark model.

For TimesFM, reducing the input patch length from 32 to 8 leads to a deterioration in out-of-sample performance, as measured by R^2_{OOS}, whereas extending the patch length to 128 improves it. The effects on overall accuracy and the F1 score, however, are less uniform. When the input patch length is shortened to 8, the average changes in overall accuracy are -0.17%, 0.06%, -0.43%, and -0.42%. When it is increased to 128, the corresponding changes are -0.31%, -0.41%, -0.27%, and -0.38%. Although TimesFM (8M) exhibits a slight improvement, the TimesFM (20M) configuration shows a noticeable decline in performance. Similarly to the overall accuracy results, the findings for the F1 scores are mixed, though they generally indicate slight improvements for both TimesFM model sizes. Turning to portfolio performance, variation in patch length generally exerts a negative effect, with most configurations exhibiting lower Sharpe ratios relative to the default specification. When the input patch size is increased and the window size is set to 512, the Sharpe ratios for TimesFM (8M) and TimesFM (20M) decrease by 0.17 and 3.69, respectively. For shorter input patch sizes, the

corresponding declines are 0.28 and 3.56. A similar pattern emerges for a window size of 252: the reductions in Sharpe ratios amount to 0.17 and 3.37 under longer patch sizes, and 0.28 and 3.53 under shorter patch sizes. In general, these declines stem from a combination of lower annualized returns and higher standard deviations. These findings indicate that, unlike Chronos, TimesFM does not benefit from changes in its input patch length, and its performance tends to deteriorate when deviating from the default configuration.

Taken together, these results highlight the sensitivity of pre-trained TSFMs to hyperparameter choices. In the case of Chronos, even without scaling the data beyond the global sample, appropriate hyperparameter adjustments enable the model to outperform benchmark models in several configurations. The hyperparameter settings adopted by model developers are also typically motivated by the goal of ensuring generalization across heterogeneous datasets with mixed frequencies. Narrowing this objective to more specialized model classes while concurrently reducing the size and scope of the TSFMs, as undertaken in this study, creates opportunities for deeper examination of the role of hyperparameter selection.

5.3 International Results

All preceding out-of-sample results focus on the U.S. market; however, extending the analysis to international settings is essential. Accordingly, we evaluate model performance across seven major markets: Hong Kong, Taiwan, South Korea, Germany, the United Kingdom, India, and Australia, selected based on market size, data reliability, and the permissibility of short selling. The benchmark models are trained on global data and then applied to these individual markets. We subsequently report results for the TSFMs, presented separately for different pre-training datasets, including global, JKP-augmented, and synthetic-augmented variants. Section 5.3.1 presents the international results for the benchmark models, and Section 5.3.2 presents the corresponding results for the TSFMs.[39]

5.3.1 Benchmark Results

Table 13 reports R^2_{OOS} for benchmark models trained on global data and evaluated across seven major markets (HKG, TWN, KOR, DEU, GBR, IND, and AUS). Ensemble models consistently lead performance. Averaging across all markets and window sizes, ensemble models (XGBoost, CatBoost, and LightGBM) deliver an average of R^2_{OOS} of about 1.78%, clearly outperforming neural networks at

[39]Table B.18 summarizes the scope and coverage of the international out-of-sample evaluation. The United Kingdom, Germany, and India provide the largest datasets, comprising approximately 59 million, 44 million, and 40 million excess return observations, respectively. All other markets contain fewer than 30 million observations. Across all markets, the evaluation period spans 2001–2023, with the number of securities generally ranging between 2,000 and 11,500 per year. The time span aligns with the corresponding out-of-sample period used for the U.S. market results.

approximately 0.35% and linear models at approximately -0.12%. Averaged across ensemble models and window sizes, Germany, India, and Australia exhibit the largest out-of-sample improvements, with average R^2_{OOS} values of approximately 5.80%, 3.83%, and 2.45%, respectively. Hong Kong delivers more modest yet consistently positive gains, with an average of about 0.91%. The U.K. shows only slight improvement, with an average near 0.12%, whereas Taiwan and Korea remain difficult environments for forecasting, yielding negative average R^2_{OOS} values of roughly -0.10% and -0.51%. Linear models yield clearly positive average R^2_{OOS} only in Germany (2.40%) and Australia (0.80%), while their average values in the remaining markets remain close to zero or become negative. PCR performs even worse, producing negative average R^2_{OOS} in five of the seven markets and only positive averages in Germany (2.74%) and Australia (0.69%). Moreover, when averaging R^2_{OOS} across all markets and models within each family, ensemble models display clear gains as the window size lengthens: the average R^2_{OOS} increases from 1.24% at the 5-day window to 1.90% and 2.05% at the 21- and 252-day windows, and remains elevated at 1.94% for the 512-day window. By contrast, the corresponding averages for the linear models are -0.13%, -0.10%, -0.14%, and -0.13%, and for the neural networks are 0.48%, 0.37%, 0.29%, and 0.27% at the 5-, 21-, 252-, and 512-day windows, respectively, indicating no systematic improvement in predictive performance as the window lengthens.

Table 14 also presents the corresponding Sharpe ratios for the same set of benchmark models. Averaging across all markets, window sizes, and model specifications within each family, ensemble models again deliver the strongest outcomes, achieving an average Sharpe ratio of about 3.17, compared with roughly 1.57 for linear models and 0.99 for neural networks. Averaged across ensemble models and window sizes, Germany, India, and Australia exhibit the highest risk-adjusted performance, with average Sharpe ratios of approximately 5.85, 5.76, and 6.06, respectively. Hong Kong also shows consistent, albeit smaller, improvements (around 1.80 on average), while the U.K. exhibits moderate gains (about 1.67). In contrast, Taiwan and Korea remain challenging environments, with ensemble Sharpe ratios averaging only about 0.88 and 0.20, respectively. Linear models yield clearly positive average Sharpe ratios only in Germany and Australia (around 4.90 and 5.13), and more modest gains in Hong Kong, the U.K., and India, while performance in Taiwan remains strongly negative (about -1.46) and is close to zero in Korea (about 0.06). Within the linear family, PCR continues to underperform the other linear specifications: its average Sharpe ratio is roughly 1.21 across all markets and window sizes, and values above 3 are confined to Germany and Australia. Neural networks trail the ensemble models, with lower and less stable Sharpe ratios across markets. Moreover, when averaging Sharpe ratios across all markets and models within each family for a given window size, ensemble models display the most favorable pattern as the window size lengthens: their average Sharpe ratio increases

from about 2.92 at the 5-day window to 3.33 and 3.31 at the 21- and 252-day windows, and remains elevated at 3.14 for the 512-day window. By contrast, the corresponding averages for the linear models are 1.73, 1.53, 1.53, and 1.49, and for the neural networks are 1.03, 1.04, 0.99, and 0.91 at the 5-, 21-, 252-, and 512-day windows, respectively, indicating no systematic improvement in portfolio performance as the window lengthens. Overall, the Sharpe ratio results reinforce the superiority of ensemble models and confirm that statistical improvements in predictive power correspond to tangible economic gains in several, but not all, international markets.

Table 13: Benchmark Models - R^2_{OOS} (International)

Model	**OLS+H**				**LASSO+H**				**RIDGE+H**				**Enet+H**				**PCR**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
HKG	0.24	0.27	0.27	0.30	0.28	0.30	0.28	0.28	0.29	0.33	0.28	0.25	0.18	0.20	0.32	0.33	-0.05	-0.21	-0.08	-0.17
TWN	-0.81	-0.79	-0.89	-0.94	-1.60	-1.61	-1.78	-1.87	-1.87	-1.82	-2.14	-2.29	-2.23	-2.24	-1.28	-1.37	-2.73	-1.89	-2.98	-2.74
KOR	-0.64	-0.64	-0.72	-0.78	-1.24	-1.27	-1.37	-1.44	-1.46	-1.53	-1.70	-1.75	-1.72	-1.75	-0.99	-1.06	-2.05	-1.23	-2.60	-2.08
DEU	1.24	1.26	1.39	1.44	2.42	2.44	2.57	2.64	2.68	2.70	2.98	3.12	2.95	2.98	2.09	2.19	3.16	2.16	2.73	2.90
GBR	-0.22	-0.22	-0.24	-0.26	-0.62	-0.61	-0.71	-0.78	-0.70	-0.71	-0.81	-0.92	-0.94	-0.94	-0.45	-0.53	-1.31	-0.99	-1.35	-1.41
IND	-0.20	-0.19	-0.19	-0.19	-0.16	-0.17	-0.24	-0.25	-0.29	-0.27	-0.31	-0.35	-0.46	-0.48	-0.06	-0.09	-0.90	-0.50	-0.62	-0.64
AUS	0.48	0.49	0.52	0.54	0.85	0.86	0.87	0.88	1.03	1.04	1.05	1.06	1.04	1.05	0.76	0.78	0.98	0.58	0.61	0.57
Model	**XGBoost**				**CatBoost**				**LightGBM**				**NN-S**				**NN-L**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
HKG	0.59	0.88	0.98	0.95	0.65	1.07	1.11	1.08	0.54	1.03	0.99	1.10	0.73	0.57	0.12	0.14	0.53	0.65	0.78	0.91
TWN	-0.50	0.05	-0.03	-0.21	-0.35	0.15	0.27	0.16	-0.62	0.04	-0.07	-0.12	-1.33	-1.26	-1.69	-1.97	-1.39	-1.67	-1.30	-1.31
KOR	-0.59	-0.44	-0.36	-0.51	-0.70	-0.60	-0.30	-0.24	-0.99	-0.49	-0.48	-0.45	-0.82	-1.03	-1.53	-1.19	-1.35	-1.18	-1.18	-0.84
DEU	4.59	5.84	6.22	6.11	4.77	6.06	6.56	6.30	4.50	5.98	6.32	6.31	3.00	3.02	2.93	2.41	3.39	3.15	3.53	3.41
GBR	-0.19	0.14	0.22	0.12	-0.23	0.25	0.40	0.28	-0.26	0.21	0.23	0.22	-0.61	-0.76	-0.84	-1.08	-0.68	-0.72	-0.75	-0.67
IND	2.70	3.99	4.11	3.68	3.11	4.23	4.71	4.23	2.89	4.09	4.14	4.03	1.33	1.15	0.40	0.69	1.40	0.96	1.40	1.29
AUS	2.06	2.45	2.65	2.50	1.98	2.47	2.66	2.54	2.04	2.55	2.74	2.71	1.17	1.16	1.04	0.80	1.30	1.17	1.19	1.16

Note: This table reports the out-of-sample R^2 (R^2_{OOS}) from forecasts of benchmark models trained on global data evaluated across rolling window sizes of 5, 21, 252, and 512 trading days from 2001 to 2023. Benchmark models include linear (OLS, Lasso, Ridge, Elastic Net, and PCR), ensemble (XGBoost, CatBoost, and LightGBM), and neural network (NN-S and NN-L) models. 'H' indicates that the model is estimated using the Huber loss. The set of countries is determined by three criteria: the relative size of the equity market (measured by total market capitalization), the allowance of short selling, and the availability of reliable data. The ordering of countries is arbitrary.

Table 14: Benchmark Models - Sharpe Ratio (International)

Model	**OLS+H**				**LASSO+H**				**RIDGE+H**				**Enet+H**				**PCR**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
HKG	1.71	1.52	1.39	1.30	1.02	1.02	1.02	1.02	1.62	1.36	1.28	1.28	1.20	1.24	1.02	1.02	1.20	0.50	0.99	0.87
TWN	-1.51	-1.24	-1.44	-1.37	-1.52	-1.52	-1.52	-1.52	-1.58	-1.37	-1.47	-1.41	-1.70	-1.65	-1.52	-1.52	-1.35	-1.39	-1.22	-1.38
KOR	-0.01	-0.05	-0.02	0.07	0.15	0.15	0.15	0.15	0.01	-0.05	-0.00	-0.00	0.12	0.12	0.15	0.15	0.16	-0.04	0.05	-0.14
DEU	5.93	5.48	5.17	5.03	4.78	4.78	4.78	4.78	5.91	5.50	5.30	5.32	5.21	5.29	4.78	4.78	4.76	3.03	3.64	3.70
GBR	0.94	0.93	0.86	0.75	0.57	0.57	0.57	0.57	0.90	0.76	0.97	0.88	0.78	0.73	0.57	0.57	0.75	0.24	0.61	0.42
IND	0.54	0.84	0.81	0.91	0.30	0.30	0.30	0.30	0.58	0.70	0.76	0.59	0.11	0.06	0.30	0.30	0.25	0.42	0.70	0.76
AUS	6.42	5.55	5.64	5.27	4.93	4.93	4.93	4.93	6.42	5.81	5.31	5.27	5.46	5.49	4.93	4.93	5.61	3.45	3.71	3.59
Model	**XGBoost**				**CatBoost**				**LightGBM**				**NN-S**				**NN-L**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
HKG	1.63	1.81	1.88	1.74	1.75	1.86	1.93	1.92	1.69	1.77	1.79	1.77	0.66	0.83	0.78	0.68	1.17	0.95	0.44	0.42
TWN	0.25	0.96	1.12	0.48	0.42	1.50	1.59	1.15	0.47	1.28	0.79	0.59	-0.98	-0.64	-1.12	-0.95	-1.03	-1.01	-0.67	-0.89
KOR	0.06	0.19	0.41	0.16	0.19	0.30	0.28	0.29	0.12	0.10	0.17	0.15	0.10	-0.04	-0.08	-0.08	0.36	0.14	-0.01	0.01
DEU	5.69	5.95	5.85	5.73	5.59	6.08	5.92	5.90	5.61	5.91	5.93	5.99	2.86	2.56	3.16	2.45	2.62	2.78	2.37	2.60
GBR	1.31	1.63	1.67	1.52	1.27	1.94	2.35	1.88	1.40	1.80	1.57	1.67	0.48	0.41	0.59	0.11	0.13	0.37	0.52	0.56
IND	5.34	6.00	5.58	5.17	5.49	6.44	6.74	6.16	5.22	6.13	5.58	5.31	1.18	1.21	1.28	1.43	1.13	1.47	1.39	1.65
AUS	5.88	6.03	6.15	6.10	5.89	6.16	6.02	6.16	5.97	6.02	6.14	6.16	2.87	2.29	2.79	2.14	2.81	3.29	2.44	2.62

Note: This table reports the out-of-sample Sharpe ratios of long–short portfolios constructed from forecasts of benchmark models trained on global data, evaluated across rolling window sizes of 5, 21, 252, and 512 trading days from 2001 to 2023. Benchmark models include linear (OLS, Lasso, Ridge, Elastic Net, and PCR), ensemble (XGBoost, CatBoost, and LightGBM), and neural network (NN-S and NN-L) models. 'H' indicates that the model is estimated using the Huber loss. Portfolios are formed using decile sorting based on model forecasts, with equal weighting across stocks. The set of countries is determined by three criteria: the relative size of the equity market (measured by total market capitalization), the allowance of short selling, and the availability of reliable data. The ordering of countries is arbitrary.

5.3.2 TSFM Results

Table 15 reports R^2_{OOS} from forecasts of the TSFMs pre-trained on global data (top panel), JKP-augmented data (middle panel), and synthetic-augmented data (bottom panel). On average across all models, markets, and window sizes, global-only pre-training delivers strongly negative performance, with average R^2_{OOS} of about -13.09% (and window-specific averages of -15.71%, -14.43%, -10.19%, and -12.03% at the 5-, 21-, 252-, and 512-day windows, respectively), far below the ensemble models in Section 5.3.1. Augmenting the pre-training data with JKP factors substantially mitigates these losses: the average R^2_{OOS} rises to roughly -4.36%, with window-level gains of about 4.11%, 9.61%, 10.05%, and 11.16% relative to global-only TSFMs at the 5-, 21-, 252-, and 512-day windows. Synthetic augmentation yields an intermediate pattern, with average R^2_{OOS} around -7.16% overall and changes of -8.14%, 9.32%, 10.35%, and 12.18% at the same window sizes (the 5-day window deteriorates, whereas longer windows improve sharply). Focusing on the longer 252- and 512-day windows, the cross-market average R^2_{OOS} improves from roughly -11.11% under global-only pre-training to about -0.51% with JKP augmentation and turns slightly positive at 0.15% under synthetic augmentation, bringing TSFMs close to break-even but still below the ensemble averages at comparable window sizes. Gains are broad-based yet uneven across markets: they are most pronounced in Germany and India, with Australia also improving noticeably, while Hong Kong and the U.K. move closer to break-even but often remain slightly negative, and Taiwan and Korea remain the most challenging environments. By TSFM family, augmentation benefits TimesFM more than Chronos: starting from very weak results under global-only pre-training, TimesFM (20M) improves from an average of about -29.14% to 1.39% (JKP) and 0.79% (synthetic) at the longer windows, whereas Chronos models move only from values around zero to a narrow range between roughly -0.24% and 0.30%. Finally, model size effects are secondary but consistent under augmentation at long window sizes: TimesFM (20M) tends to exceed TimesFM (8M), and Chronos (small) generally edges out the mini and tiny variants. For Chronos, this advantage is most pronounced under synthetic augmentation; under JKP augmentation the three sizes perform more similarly, with the tiny model occasionally slightly ahead.

Table 16 translates TSFM forecasts into portfolio performance and is broadly consistent with Table 15. Under global pre-training, averaging across models, the longer-window (252- and 512-day) Sharpe ratios are already high in Germany, India, and Australia, with cross-model averages of about 2.70, 4.01, and 2.27, respectively. Hong Kong and the U.K. achieve more moderate but still positive longer-window Sharpe ratios of roughly 0.82 and 1.12, whereas Taiwan and Korea remain challenging, with averages of only 0.67 and 0.45. Augmenting the pre-training data with JKP factors or synthetic series raises performance further, particularly at longer windows: the average longer-window Sharpe

ratio across all markets and models increases from about 1.72 under global pre-training to 2.44 with JKP augmentation and 2.34 with synthetic augmentation. The gains are largest and most reliable in Germany, India, and Australia, where longer-window Sharpe ratios rise to 3.82 and 4.00 (Germany), 5.36 and 4.86 (India), and 3.55 and 3.47 (Australia) under JKP and synthetic augmentation, implying improvements of roughly 1.12–1.35 relative to the global baseline. Hong Kong, the U.K., and Taiwan also benefit, with longer-window averages increasing to around 1.15–1.20, 1.45–1.56, and 1.01–1.13, respectively, while Korea remains comparatively hard to exploit, with Sharpe ratios generally below 1 even after augmentation. For TimesFM in Taiwan, JKP and synthetic augmentation are particularly effective: at the 252-day window, Sharpe ratios move from -0.08 and -0.42 (TimesFM 8M and 20M under global pre-training) to 0.01 and 1.13 with JKP augmentation and to 0.41 and 0.85 with synthetic augmentation; at the 512-day window, they move from 0.20 and -0.58 to 0.16 and 1.16 (JKP) and to 0.44 and 0.96 (synthetic). Averaging across markets and the longer windows, Chronos models dominate under global pre-training, with average Sharpe ratios of about 2.32 versus 0.82 for TimesFM, whereas with JKP and synthetic augmentation the average advantage shifts to TimesFM at larger scale (2.84 and 2.74 for TimesFM, compared with 2.18 and 2.08 for Chronos). Overall, the portfolio results confirm that the statistical gains from JKP and synthetic augmentation for TSFMs translate into substantial improvements in risk-adjusted performance.

The international TSFM results closely mirror the patterns observed for the U.S. in Section 5.2.2. Ensemble models set a high bar internationally: across markets and window sizes they deliver consistently positive and stable R^2_{OOS} and Sharpe ratios, with particularly strong performance in Germany, India, and Australia and only modest gains in Hong Kong and the U.K.; Taiwan and Korea remain difficult environments for return forecasting. By contrast, TSFMs pre-trained only on global data generate strongly negative average R^2_{OOS}, even though longer-window Sharpe ratios are already competitive in several markets. Augmenting the pre-training data with JKP factors or synthetic financial series substantially improves TSFM performance: cross-market R^2_{OOS} at longer windows shifts from clearly negative under global-only pre-training to values near zero, and in some cases slightly above zero, while Sharpe ratios increase in nearly all markets, with the largest improvements again in Germany, India, and Australia. Model scale and architecture matter: larger TimesFM variants benefit most from augmentation and, at long window sizes in markets with stronger predictability, often approach the best ensemble models, whereas Chronos models improve but generally remain behind. Overall, the international evidence confirms the U.S. pattern that scale and financial-domain specialization jointly determine TSFM efficacy; at the same time, conventional ML models, especially ensemble models, remain the most robust and reliable performers across global markets.

Table 15: Pre-Trained TSFMs - R^2_{OOS} (International)

	Global																			
Model	**Chronos (Tiny)**				**Chronos (Mini)**				**Chronos (Small)**				**TimesFM (8M)**				**TimesFM (20M)**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
HKG	-0.78	-0.40	-0.75	-1.14	-0.65	-0.67	-0.84	-1.13	-1.11	-0.43	-0.48	-0.88	-31.64	-23.15	-31.64	-46.53	-43.24	-51.53	-32.18	-34.42
TWN	-1.04	-1.27	-2.68	-2.91	-1.36	-1.44	-3.66	-4.07	-1.65	-1.60	-4.67	-4.74	-30.96	-22.34	-27.66	-39.14	-42.39	-50.72	-34.00	-38.71
KOR	-1.07	-0.69	-0.82	-1.04	-1.13	-0.38	-1.73	-2.15	-1.40	-1.39	-1.67	-2.30	-30.69	-22.74	-21.64	-31.83	-40.20	-46.77	-26.77	-37.65
DEU	-0.66	0.96	1.50	1.08	-0.91	1.28	2.52	2.24	-0.73	1.82	3.17	2.91	-36.70	-22.84	-23.12	-22.39	-46.81	-49.51	-18.79	-17.75
GBR	-1.20	-0.64	-0.59	-0.74	-1.25	-0.59	-0.47	-0.63	-1.14	-0.60	-0.59	-0.77	-32.31	-22.44	-24.49	-30.96	-43.15	-49.79	-25.59	-30.12
IND	0.38	2.06	2.44	2.11	0.24	3.19	4.17	3.95	0.50	3.66	4.55	4.41	-31.69	-22.63	-17.11	-17.98	-41.00	-48.93	-29.95	-27.73
AUS	-0.87	0.24	0.27	0.49	-1.10	0.34	0.67	0.69	-0.97	0.32	0.75	0.72	-35.06	-22.98	-18.45	-13.95	-46.18	-52.62	-26.34	-28.02
	JKP-Augmented																			
HKG	-4.67	-2.69	-0.75	-0.91	-3.83	-2.62	-0.93	-1.38	-5.59	-2.16	-0.96	-1.26	-22.57	-10.28	-3.53	-7.43	-19.95	-6.61	-0.06	0.13
TWN	-4.81	-3.05	-2.39	-2.09	-4.89	-3.21	-3.66	-4.09	-5.64	-3.40	-4.38	-4.63	-23.15	-12.74	-5.82	-10.33	-20.53	-10.50	-1.83	-1.54
KOR	-4.33	-1.63	-0.93	-0.93	-4.90	-2.08	-1.22	-1.19	-4.69	-2.84	-1.83	-2.62	-23.27	-13.76	-5.28	-9.84	-20.73	-11.87	-1.58	-2.22
DEU	-5.43	-1.52	1.84	1.80	-5.54	-1.47	2.37	2.26	-5.44	-0.95	3.90	3.53	-23.70	-6.57	3.70	3.39	-22.07	-1.75	7.29	7.69
GBR	-4.92	-3.36	-0.52	-0.69	-4.85	-3.64	-0.56	-0.75	-6.20	-3.85	-0.69	-0.74	-22.95	-11.69	-3.94	-7.66	-21.04	-8.68	-1.09	-0.97
IND	-3.70	-0.84	2.79	2.42	-3.90	-0.22	3.52	2.67	-3.74	-0.56	4.01	3.60	-20.86	-9.49	0.12	-3.49	-18.36	-5.61	4.29	4.65
AUS	-4.48	-1.26	0.34	0.50	-5.17	-2.04	0.39	0.70	-5.05	-1.45	0.79	0.95	-23.62	-9.26	-0.57	-2.56	-21.57	-5.12	2.16	2.49
	Synthetic-Augmented																			
HKG	-18.86	-2.27	-0.62	-0.91	-21.23	-3.12	-0.59	-0.62	-23.89	-3.48	-0.45	-0.46	-32.41	-8.46	-0.94	-0.98	-18.09	-4.92	-0.06	-0.05
TWN	-24.41	-7.01	-2.68	-2.55	-25.19	-6.04	-2.68	-2.59	-23.57	-5.80	-2.61	-2.78	-33.29	-10.95	-3.54	-3.12	-19.64	-7.74	-2.45	-2.31
KOR	-20.53	-6.04	-0.77	-1.00	-23.28	-9.44	-0.99	-0.85	-26.36	-6.08	-0.83	-1.07	-31.94	-12.03	-3.93	-3.97	-20.12	-9.25	-3.09	-3.39
DEU	-17.74	-1.05	1.49	1.48	-20.81	-0.86	2.12	2.19	-23.48	-1.69	3.01	2.56	-33.70	-5.73	5.37	5.88	-20.39	-0.80	7.20	7.49
GBR	-21.44	-2.29	-0.62	-0.97	-25.95	-3.12	-0.56	-0.73	-29.62	-3.30	-0.58	-0.69	-32.35	-9.66	-1.97	-2.01	-19.69	-6.33	-1.53	-1.54
IND	-18.77	-4.52	1.65	1.81	-19.37	-4.15	2.98	2.31	-19.95	-3.74	3.71	3.07	-29.72	-7.74	2.15	2.24	-16.77	-4.07	3.23	3.52
AUS	-18.73	-2.05	0.15	0.25	-22.34	-1.99	0.31	0.75	-27.64	-1.69	0.45	0.80	-33.38	-7.86	1.23	1.45	-20.10	-3.64	1.93	2.07

Note: This table reports the out-of-sample R^2 (R^2_{OOS}) from forecasts of the time series foundation models (TSFMs) pre-trained on global data (top panel), JKP-augmented data (middle panel), and synthetic-augmented data (bottom panel), evaluated across rolling window sizes of 5, 21, 252, and 512 trading days from 2001 to 2023. JKP factors used in the augmented data are defined in Jensen et al. (2023), and the synthetic data is generated following Ansari et al. (2024). The TSFMs include Chronos (tiny, mini, and small) and TimesFM (with 8 million and 20 million parameters). Zero-shot inference is performed using the pre-trained models. The set of countries is determined by three criteria: the relative size of the equity market (measured by total market capitalization), the allowance of short selling, and the availability of reliable data. The ordering of countries is arbitrary.

Table 16: Pre-Trained TSFMs - Sharpe Ratio (International)

	Global																			
Model	**Chronos (Tiny)**				**Chronos (Mini)**				**Chronos (Small)**				**TimesFM (8M)**				**TimesFM (20M)**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
HKG	0.23	0.73	1.10	0.79	0.19	0.73	1.11	0.91	0.22	0.81	1.41	1.37	-1.40	-1.15	-0.20	0.10	-0.78	0.12	0.80	0.82
TWN	0.31	0.45	1.19	0.99	0.26	1.07	1.57	1.23	0.42	0.63	1.29	1.30	0.49	0.40	-0.08	0.20	0.36	0.07	-0.42	-0.58
KOR	-0.25	0.32	0.77	0.30	-0.04	0.29	0.38	0.68	-0.07	0.21	0.64	0.75	-0.22	-0.15	0.35	0.06	-0.08	-0.11	0.35	0.21
DEU	-0.27	1.81	2.53	2.75	0.13	2.02	3.11	3.24	0.29	2.19	3.42	3.40	-3.53	-1.14	1.02	0.97	-2.70	0.69	3.11	3.48
GBR	-0.18	0.34	1.39	1.54	-0.26	0.82	1.33	1.57	-0.18	0.81	1.47	1.77	-1.12	-0.98	0.16	0.23	-0.85	-0.19	0.89	0.86
IND	2.81	4.79	5.63	5.09	2.92	5.57	6.21	6.08	3.08	5.60	6.71	6.40	0.42	0.38	0.76	0.55	1.00	1.44	1.12	1.58
AUS	0.02	1.81	2.30	2.63	-0.31	2.17	2.77	2.67	0.17	2.26	2.65	2.99	-4.86	-2.52	0.53	0.89	-3.84	-0.12	2.49	2.78
	JKP-Augmented																			
HKG	-0.46	0.37	1.09	0.96	-0.25	0.35	1.10	0.93	-0.13	0.27	1.09	1.16	-0.98	-0.05	0.97	1.24	-0.96	0.21	1.42	1.58
TWN	0.52	0.78	0.86	1.28	0.48	0.70	1.57	1.45	0.53	1.14	1.77	1.88	0.32	0.25	0.01	0.16	0.17	0.07	1.13	1.16
KOR	-0.25	0.39	0.73	0.27	-0.12	0.40	0.72	0.48	-0.24	0.22	0.92	0.32	-0.17	0.03	0.21	0.03	-0.16	-0.08	0.69	0.66
DEU	-2.16	0.24	2.73	2.69	-2.41	0.27	3.01	3.14	-2.19	0.58	3.20	3.27	-2.79	1.04	4.26	4.56	-2.63	2.35	5.56	5.77
GBR	-1.03	-0.03	1.37	1.62	-0.90	0.10	1.32	1.61	-0.96	-0.04	1.41	1.55	-0.87	-0.13	1.59	1.46	-0.98	0.18	1.89	1.81
IND	1.27	3.48	4.71	4.27	1.33	3.29	5.48	5.11	1.74	3.47	6.11	5.77	0.57	1.76	4.76	4.73	0.72	3.14	6.31	6.39
AUS	-2.52	-0.18	2.18	2.46	-2.74	0.03	2.28	2.21	-2.32	0.54	2.59	2.77	-4.31	0.14	4.82	4.99	-4.75	1.29	5.39	5.84
	Synthetic-Augmented																			
HKG	-0.44	0.27	0.85	1.27	-0.49	0.54	1.12	1.18	-0.42	0.26	1.12	1.33	-0.87	-0.27	1.03	1.12	-1.09	0.27	1.28	1.72
TWN	0.34	0.57	1.01	1.26	0.70	0.88	1.14	1.34	0.82	0.82	1.29	1.37	0.50	0.34	0.41	0.44	0.38	0.58	0.85	0.96
KOR	-0.11	-0.01	0.73	0.37	-0.13	0.00	0.83	0.33	-0.44	-0.22	0.65	0.53	-0.08	0.03	0.11	0.14	-0.14	-0.22	0.23	0.32
DEU	-0.87	0.47	2.71	2.75	-1.25	0.46	3.22	3.36	-1.81	0.70	3.10	3.20	-2.88	0.26	4.91	5.20	-3.48	1.45	5.65	5.89
GBR	-0.76	-0.20	0.80	1.16	-0.87	0.20	1.58	1.81	-0.79	0.30	1.46	1.88	-0.66	-0.13	1.10	1.40	-0.74	0.23	1.73	1.61
IND	2.07	2.65	4.03	3.83	1.70	3.03	5.26	4.97	1.63	3.18	5.53	5.34	0.88	1.46	4.73	4.76	0.81	2.79	5.24	4.87
AUS	-1.61	0.45	2.05	2.16	-2.08	0.13	2.11	2.45	-2.41	0.45	2.35	2.55	-4.30	-0.73	4.84	5.28	-4.56	0.33	5.31	5.57

Note: This table reports the out-of-sample Sharpe ratios of long–short portfolios constructed from forecasts of the time series foundation models (TSFMs) pre-trained on global data (top panel), JKP-augmented data (middle panel), and synthetic-augmented data (bottom panel), evaluated across rolling window sizes of 5, 21, 252, and 512 trading days from 2001 to 2023. JKP factors used in the augmented data are defined in Jensen et al. (2023), and the synthetic data is generated following Ansari et al. (2024). The TSFMs include Chronos (tiny, mini, and small) and TimesFM (with 8 million and 20 million parameters). Zero-shot inference is performed using the pre-trained models. Portfolios are formed using decile sorting based on model forecasts, with equal weighting across stocks. The set of countries is determined by three criteria: the relative size of the equity market (measured by total market capitalization), the allowance of short selling, and the availability of reliable data. The ordering of countries is arbitrary.

6 Conclusion

This paper offers the first empirical assessment of TSFMs for excess return forecasting across U.S. and global markets. Off-the-shelf, generic pre-trained TSFMs underperform strong benchmarks, especially tree-based ensembles. Fine-tuning improves performance modestly relative to zero-shot settings but remains insufficient to close the gap with benchmark models. In contrast, TSFMs pre-trained from scratch on financial data deliver substantial gains in both predictive and economic performance and are most competitive with longer input windows; conventional models dominate at shorter window sizes. Nonetheless, TSFMs remain weaker in terms of goodness-of-fit: zero-shot models yield poor R^2, fine-tuning offers modest improvements, and finance-native pre-training does not fully close the gap.

TSFMs pre-trained from scratch on smaller yet domain-specific datasets, and with considerably fewer parameters, can outperform most off-the-shelf TSFMs in both forecasting accuracy and portfolio performance. While benchmark models tend to degrade when trained on heterogeneous global returns, TSFMs benefit from broader and more diverse pre-training sets. Additional gains emerge when the data are augmented with monthly factors or synthetic variables. Across international markets, tree-based ensembles remain the most reliable baseline; however, scaled, finance-specialized TSFMs become increasingly competitive. Their performance, though, is sensitive to pre-training composition, window length, and hyperparameter choices. TSFMs also display slower performance decay over time, suggesting greater temporal robustness. These advantages come with substantial computational costs, particularly for from-scratch pre-training, underscoring the need for more efficient, finance-oriented architectures. Overall, TSFMs represent a promising paradigm for financial forecasting when domain alignment and computational feasibility are jointly managed. Also, some off-the-shelf TSFMs exhibit encouraging signs of generalization, achieving strong results with minimal exposure to financial data during pre-training. Nonetheless, the findings indicate that practitioners should avoid relying solely on generic TSFMs and instead prioritize domain-specific pre-training to achieve more reliable and robust financial forecasts.

Building on these findings, several avenues for future research emerge. TSFMs are inherently capable of generating multiple forecasts that form a full predictive distribution, allowing future research to explore higher-order statistics. Extending the analysis to longer forecasting horizons would further illuminate how predictive signals evolve and decay over time. Furthermore, a natural progression is to test multivariate extensions of TSFMs that capture cross-asset dependencies, market factors, and macro-financial linkages. Beyond returns, applying TSFMs to other forms of financial data such as volatility, order flow, or limit order book information could enhance their applicability to risk

management and market microstructure analysis. Furthermore, future research should conduct a more comprehensive analysis of how hyperparameter configurations and architectural design choices influence model performance, with the objective of developing finance-oriented TSFMs specifically optimized for distinct predictive and decision-making applications.

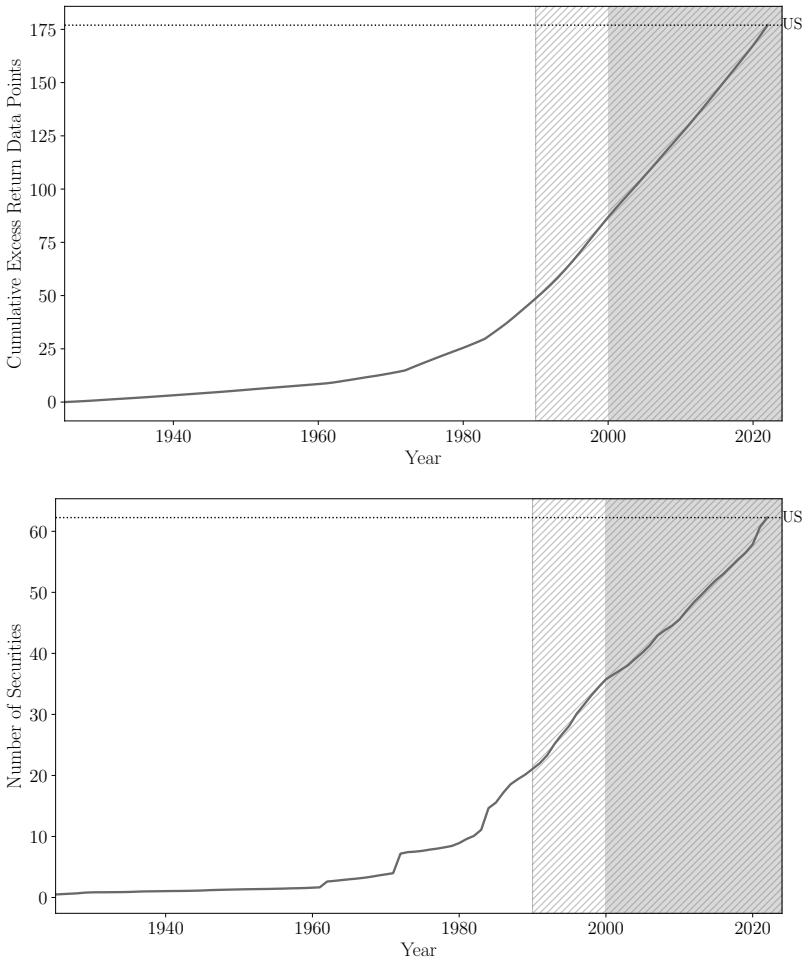
A Appendix: Data Overview

Table A.1: Mapping Codes and Corresponding Country/Region Names

Code	Country/Region	Code	Country/Region	Code	Country/Region	Code	Country/Region
ARG	Argentina	FIN	Finland	LKA	Sri Lanka	RUS	Russia
ARE	United Arab Emirates	FRA	France	LTU	Lithuania	SAU	Saudi Arabia
AUS	Australia	GGY	Guernsey	LUX	Luxembourg	SGP	Singapore
AUT	Austria	GHA	Ghana	LVA	Latvia	SRB	Serbia
BGD	Bangladesh	GRC	Greece	MAR	Morocco	SVN	Slovenia
BEL	Belgium	GBR	United Kingdom	MEX	Mexico	SVK	Slovakia
BHR	Bahrain	HRV	Croatia	MLT	Malta	SWE	Sweden
BMU	Bermuda	HKG	Hong Kong	MWI	Malawi	THA	Thailand
BRA	Brazil	HUN	Hungary	MUS	Mauritius	TTO	Trinidad and Tobago
BWA	Botswana	IDN	Indonesia	MYS	Malaysia	TUN	Tunisia
CAN	Canada	IND	India	NAM	Namibia	TUR	Turkey
CHE	Switzerland	IRL	Ireland	NLD	Netherlands	TWN	Taiwan
CHL	Chile	IRN	Iran	NGA	Nigeria	TZA	Tanzania
CHN	China	ISL	Iceland	NOR	Norway	UGA	Uganda
CIV	Côte d'Ivoire	ISR	Israel	NZL	New Zealand	UKR	Ukraine
COL	Colombia	ITA	Italy	OMN	Oman	US (USA)	United States
CZE	Czech Republic	JAM	Jamaica	PAK	Pakistan	VEN	Venezuela
CYP	Cyprus	JOR	Jordan	PER	Peru	VNM	Vietnam
DEU	Germany	JPN	Japan	PHL	Philippines	ZAF	South Africa
DNK	Denmark	KAZ	Kazakhstan	POL	Poland	ZMB	Zambia
ECU	Ecuador	KEN	Kenya	PRT	Portugal	ZWE	Zimbabwe
EGY	Egypt	KOR	South Korea	PSE	Palestine	MUL	Multiple Countries
ESP	Spain	KWT	Kuwait	QAT	Qatar		
EST	Estonia	LBN	Lebanon	ROU	Romania		

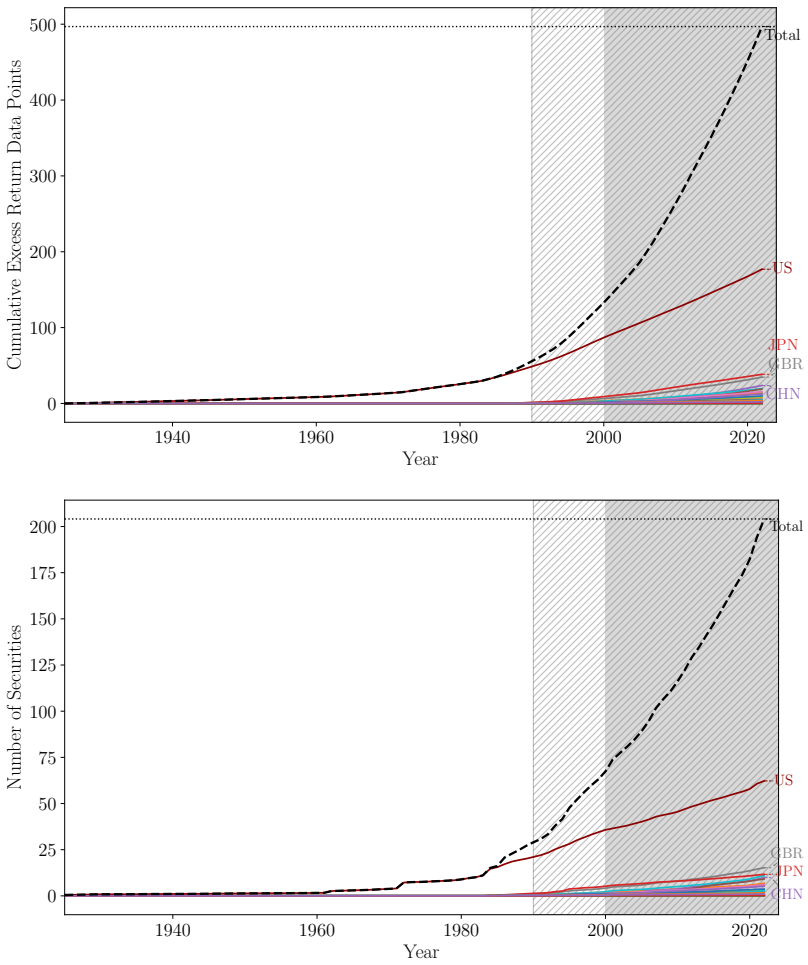
Note: This table presents the full country/region names corresponding to the abbreviated country/region codes used throughout the paper. These codes follow the standard ISO 3166-1 alpha-3 format where applicable. The entries are sorted alphabetically by code. The code 'MUL' refers to cases where a company's securities are listed across multiple countries or exchanges, or on a multilateral trading facility (MTF) that does not correspond to a single national market.

Figure A.1: Cumulative Number of Observations and Unique Securities (U.S.)



Note: The top figure displays the cumulative number of excess return observations (in millions) for the U.S. from 1925 to 2022. A dashed horizontal line indicates the maximum in 2022, corresponding to 176.96 million observations. The bottom figure presents the cumulative number of unique securities (in thousands) for the U.S. over the same period. Similarly, the dashed horizontal line marks the maximum in 2022, corresponding to 62.25 thousand securities. In both figures, the light-shaded region represents the sample period starting in 1990 (expanding window) used to train the predictive models, while the dark-shaded region denotes the period from 2001 to 2022, during which the predictive models are trained on a yearly basis.

Figure A.2: Cumulative Number of Observations and Unique Securities by Market



Note: The top figure displays the cumulative number of excess return observations (in millions) from 1925 to 2022. The dashed line represents the total number of observations aggregated across all markets, with the four markets containing the largest counts highlighted. A dashed horizontal line marks the maximum in 2022, corresponding to 496.89 million observations. The bottom figure presents the number of unique securities (in thousands) over the same period. Similarly, the dashed horizontal line indicates the maximum in 2022, corresponding to 204.09 thousand securities. In both figures, the light-shaded region represents the sample period starting in 1990 (expanding window) used to train the predictive models, while the dark-shaded region denotes the period from 2001 to 2022, during which the predictive models are trained on a yearly basis.

Table A.2: Annual Breakdown of Cross-Market Excess Returns

Year	U.S.	JPN	GBR	CHN	DEU	IND	KOR	HKG	CAN	AUS	TWN	THA	MYS	FRA	SGP	SWE
2000	86.84	8.97	6.43	0.51	3.11	1.23	1.78	1.91	3.55	1.16	1.13	1.79	1.61	1.12	0.69	0.50
2001	90.79	9.98	7.19	0.68	3.55	1.44	2.07	2.14	3.94	1.37	1.30	2.00	1.79	1.31	0.78	0.58
2002	94.64	11.01	7.96	0.93	4.04	1.67	2.41	2.37	4.33	1.60	1.49	2.21	1.97	1.50	0.89	0.66
2003	98.42	12.04	8.73	1.21	4.49	1.89	2.75	2.62	4.74	1.88	1.71	2.42	2.15	1.70	1.00	0.75
2004	102.23	13.07	9.50	1.50	4.95	2.13	3.10	2.93	5.18	2.21	1.96	2.63	2.36	1.89	1.13	0.84
2005	106.08	14.10	10.27	2.03	5.45	2.37	3.52	3.26	5.66	2.52	2.24	2.87	2.57	2.09	1.26	0.93
2006	109.93	15.54	11.41	2.89	6.08	2.69	4.03	3.72	6.16	3.01	2.63	3.21	2.84	2.34	1.46	1.03
2007	113.86	17.03	12.67	3.79	6.72	3.04	4.61	4.23	6.69	3.54	3.11	3.56	3.15	2.61	1.68	1.15
2008	117.84	18.54	14.02	4.74	7.43	3.53	5.22	4.78	7.21	4.14	3.67	3.89	3.49	2.87	1.92	1.29
2009	121.74	20.03	15.37	5.69	8.20	4.09	5.91	5.34	7.73	4.75	4.25	4.22	3.82	3.14	2.15	1.45
2010	125.63	21.52	16.74	6.68	8.98	4.76	6.61	5.92	8.26	5.37	4.84	4.59	4.14	3.45	2.39	1.61
2011	129.57	22.98	18.11	7.76	9.77	5.55	7.34	6.52	8.79	6.01	5.45	4.98	4.43	3.71	2.63	1.78
2012	133.65	24.47	19.55	9.02	10.62	6.47	8.10	7.17	9.33	6.68	6.10	5.38	4.74	3.99	2.89	1.96
2013	137.75	25.86	20.92	10.27	11.41	7.44	8.82	7.78	9.87	7.32	6.71	5.76	5.04	4.27	3.12	2.14
2014	141.93	27.17	22.21	11.49	12.16	8.50	9.49	8.37	10.40	7.92	7.29	6.12	5.35	4.56	3.34	2.32
2015	146.18	28.51	23.57	12.74	12.95	9.62	10.16	9.00	10.94	8.56	7.89	6.50	5.66	4.88	3.56	2.51
2016	150.42	29.88	25.00	14.03	13.76	10.84	10.85	9.68	11.49	9.22	8.50	6.92	5.98	5.22	3.79	2.72
2017	154.63	31.26	26.49	15.42	14.62	12.09	11.56	10.41	12.01	9.93	9.14	7.37	6.31	5.57	4.03	2.95
2018	158.87	32.69	28.05	16.91	15.51	13.41	12.31	11.20	12.54	10.66	9.79	7.85	6.63	5.92	4.27	3.21
2019	163.15	34.09	29.60	18.46	16.41	14.71	13.06	12.01	13.06	11.41	10.43	8.33	6.94	6.27	4.52	3.50
2020	167.49	35.51	31.22	20.10	17.36	16.08	13.84	12.87	13.59	12.18	11.08	8.82	7.25	6.64	4.78	3.82
2021	172.12	36.96	32.92	21.83	18.40	17.52	14.66	13.84	14.12	12.99	11.77	9.33	7.58	7.04	5.03	4.17
2022	176.99	38.41	34.71	23.69	19.56	19.02	15.52	14.86	14.65	13.87	12.49	9.84	7.91	7.46	5.29	4.52

Year	POL	ITA	ZAF	BRA	CHE	IDN	ISR	TUR	VNM	PAK	MUL*	GRC	ESP	NLD	PHL	NOR
2000	0.26	0.78	0.66	0.46	0.68	0.58	0.28	0.38	0.01	0.32	0.54	0.19	0.31	0.52	0.49	0.23
2001	0.30	0.86	0.76	0.53	0.76	0.66	0.33	0.44	0.01	0.37	0.57	0.24	0.36	0.58	0.55	0.27
2002	0.34	0.94	0.86	0.62	0.84	0.74	0.39	0.51	0.02	0.42	0.63	0.31	0.40	0.63	0.61	0.31
2003	0.38	1.03	0.96	0.71	0.92	0.82	0.45	0.59	0.02	0.47	0.69	0.38	0.45	0.69	0.67	0.35
2004	0.42	1.12	1.06	0.80	1.00	0.90	0.51	0.66	0.02	0.52	0.76	0.45	0.49	0.74	0.72	0.39
2005	0.47	1.20	1.15	0.90	1.08	0.98	0.57	0.74	0.02	0.59	0.82	0.52	0.54	0.80	0.77	0.44
2006	0.53	1.33	1.28	1.04	1.18	1.07	0.65	0.85	0.03	0.68	0.90	0.62	0.60	0.86	0.83	0.49
2007	0.60	1.46	1.41	1.18	1.30	1.16	0.74	0.96	0.03	0.76	0.98	0.72	0.67	0.93	0.89	0.57
2008	0.69	1.59	1.54	1.32	1.43	1.26	0.85	1.06	0.06	0.84	1.07	0.82	0.74	1.00	0.96	0.65
2009	0.80	1.73	1.67	1.47	1.55	1.36	0.95	1.17	0.09	0.93	1.16	0.92	0.82	1.06	1.02	0.72
2010	0.94	1.87	1.80	1.62	1.68	1.46	1.07	1.27	0.16	1.03	1.25	1.02	0.90	1.13	1.08	0.79
2011	1.10	2.00	1.93	1.77	1.81	1.57	1.18	1.37	0.29	1.13	1.33	1.11	0.99	1.19	1.15	0.87
2012	1.29	2.14	2.07	1.93	1.94	1.68	1.33	1.49	0.47	1.24	1.42	1.21	1.08	1.25	1.21	0.94
2013	1.48	2.27	2.21	2.09	2.07	1.81	1.48	1.61	0.68	1.34	1.49	1.30	1.16	1.32	1.27	1.03
2014	1.67	2.40	2.34	2.23	2.19	1.93	1.63	1.72	0.89	1.45	1.56	1.38	1.24	1.38	1.34	1.11
2015	1.87	2.53	2.48	2.38	2.32	2.06	1.79	1.85	1.09	1.58	1.63	1.46	1.33	1.44	1.41	1.19
2016	2.08	2.67	2.62	2.53	2.46	2.21	1.95	1.98	1.30	1.71	1.70	1.55	1.41	1.51	1.48	1.27
2017	2.31	2.83	2.77	2.69	2.60	2.37	2.11	2.11	1.52	1.85	1.77	1.63	1.50	1.57	1.56	1.36
2018	2.55	2.99	2.92	2.85	2.74	2.53	2.28	2.25	1.74	1.98	1.84	1.71	1.59	1.64	1.63	1.44
2019	2.83	3.14	3.08	3.01	2.87	2.70	2.45	2.39	1.98	2.12	1.90	1.80	1.69	1.70	1.70	1.53
2020	3.15	3.31	3.23	3.16	3.02	2.88	2.62	2.53	2.24	2.26	1.97	1.88	1.79	1.78	1.78	1.62
2021	3.47	3.49	3.39	3.33	3.19	3.08	2.80	2.68	2.52	2.40	2.04	1.96	1.90	1.86	1.85	1.72
2022	3.80	3.68	3.54	3.50	3.40	3.30	3.00	2.82	2.79	2.54	2.11	2.05	2.02	1.95	1.92	1.82

Year	MEX	RUS	LKA	DNK	BEL	BGD	CHL	JOR	EGY	AUT	FIN	NZL	KWT	SAU	NGA	ARG
2000	0.42	0.18	0.20	0.28	0.31	0.01	0.28	0.11	0.07	0.33	0.25	0.25	0.00	-	0.04	0.25
2001	0.46	0.21	0.25	0.32	0.35	0.01	0.32	0.13	0.09	0.36	0.28	0.27	0.00	0.00	0.04	0.28
2002	0.50	0.24	0.29	0.37	0.39	0.01	0.36	0.15	0.12	0.39	0.32	0.30	0.02	0.00	0.06	0.30
2003	0.55	0.26	0.34	0.41	0.44	0.02	0.40	0.17	0.15	0.43	0.36	0.32	0.03	0.02	0.07	0.33
2004	0.59	0.29	0.39	0.46	0.48	0.03	0.44	0.19	0.18	0.46	0.40	0.35	0.05	0.04	0.08	0.36
2005	0.63	0.32	0.43	0.50	0.54	0.05	0.48	0.21	0.21	0.49	0.44	0.38	0.07	0.06	0.09	0.39
2006	0.68	0.36	0.50	0.56	0.59	0.08	0.53	0.24	0.24	0.54	0.48	0.42	0.10	0.08	0.11	0.41
2007	0.72	0.42	0.56	0.62	0.66	0.11	0.57	0.28	0.27	0.58	0.53	0.47	0.16	0.11	0.13	0.44
2008	0.78	0.48	0.62	0.68	0.71	0.15	0.61	0.35	0.32	0.62	0.58	0.51	0.22	0.15	0.17	0.47
2009	0.84	0.55	0.68	0.75	0.78	0.18	0.66	0.42	0.38	0.67	0.63	0.56	0.29	0.20	0.22	0.50
2010	0.91	0.62	0.73	0.81	0.84	0.23	0.71	0.50	0.44	0.70	0.67	0.61	0.36	0.26	0.28	0.52
2011	0.97	0.71	0.79	0.88	0.91	0.32	0.75	0.58	0.51	0.74	0.71	0.65	0.43	0.32	0.34	0.55
2012	1.04	0.80	0.85	0.94	0.98	0.43	0.80	0.66	0.58	0.78	0.75	0.70	0.51	0.38	0.40	0.58
2013	1.11	0.89	0.93	1.00	1.04	0.52	0.85	0.74	0.65	0.82	0.79	0.75	0.58	0.44	0.46	0.61
2014	1.17	0.98	1.00	1.06	1.09	0.60	0.90	0.81	0.72	0.87	0.84	0.80	0.65	0.50	0.51	0.64
2015	1.24	1.08	1.08	1.12	1.15	0.69	0.95	0.87	0.79	0.91	0.88	0.85	0.71	0.56	0.57	0.67
2016	1.31	1.17	1.15	1.18	1.21	0.78	1.01	0.94	0.86	0.95	0.92	0.90	0.77	0.63	0.62	0.70
2017	1.38	1.27	1.23	1.25	1.27	0.89	1.06	1.01	0.94	1.00	0.97	0.96	0.84	0.70	0.68	0.74
2018	1.46	1.37	1.31	1.31	1.33	1.01	1.11	1.08	1.01	1.05	1.02	1.02	0.90	0.77	0.73	0.78
2019	1.53	1.48	1.39	1.39	1.39	1.12	1.17	1.14	1.08	1.10	1.08	1.08	0.97	0.84	0.79	0.82
2020	1.62	1.58	1.47	1.46	1.45	1.24	1.22	1.20	1.15	1.16	1.14	1.13	1.03	0.91	0.84	0.85
2021	1.71	1.69	1.55	1.54	1.51	1.36	1.28	1.26	1.23	1.23	1.20	1.19	1.09	0.99	0.89	0.89
2022	1.81	1.79	1.63	1.62	1.58	1.48	1.34	1.32	1.31	1.30	1.26	1.25	1.16	1.07	0.94	0.93

Note: This table presents the cumulative number of valid excess return observations by year and market, with values scaled in millions. Although the dataset begins in 1925, coverage for many non-U.S. markets is limited in the earlier decades. Columns are ordered by their 2022 values. Blank cells denote missing data for the respective year.

* When a company's securities are listed across multiple countries or exchanges, or on a multilateral trading facility (MTF) that doesn't correspond to a single national market.

Table A.3: Annual Breakdown of Cross-Market Excess Returns (continued)

Year	BGR	LUX	PER	ARE	HRV	OMN	ROU	IRL	CYP	PRT	MAR	ZWE	TUN	CZE	HUN	COL
2000	0.02	0.11	0.08	0.01	0.03	0.00	0.04	0.12	0.06	0.09	0.04	0.06	0.01	0.08	0.06	0.06
2001	0.02	0.13	0.09	0.01	0.04	0.01	0.04	0.14	0.07	0.11	0.05	0.07	0.02	0.09	0.07	0.07
2002	0.03	0.14	0.10	0.02	0.05	0.01	0.05	0.15	0.09	0.12	0.06	0.09	0.03	0.11	0.08	0.08
2003	0.03	0.16	0.12	0.02	0.05	0.02	0.05	0.16	0.11	0.14	0.07	0.10	0.04	0.12	0.09	0.09
2004	0.03	0.17	0.14	0.03	0.06	0.03	0.06	0.18	0.13	0.16	0.07	0.11	0.04	0.13	0.10	0.10
2005	0.03	0.19	0.16	0.03	0.06	0.05	0.07	0.19	0.14	0.17	0.08	0.12	0.05	0.15	0.11	0.10
2006	0.03	0.20	0.18	0.04	0.07	0.06	0.08	0.21	0.16	0.19	0.09	0.13	0.06	0.16	0.12	0.11
2007	0.05	0.22	0.20	0.06	0.09	0.08	0.09	0.23	0.18	0.20	0.11	0.15	0.08	0.17	0.13	0.12
2008	0.09	0.24	0.22	0.10	0.11	0.10	0.10	0.25	0.20	0.22	0.12	0.16	0.09	0.19	0.14	0.13
2009	0.13	0.27	0.25	0.14	0.14	0.12	0.11	0.27	0.22	0.23	0.15	0.18	0.11	0.20	0.15	0.15
2010	0.17	0.29	0.28	0.18	0.18	0.14	0.13	0.28	0.25	0.25	0.17	0.20	0.13	0.21	0.16	0.16
2011	0.22	0.32	0.31	0.22	0.22	0.17	0.16	0.30	0.27	0.26	0.20	0.21	0.15	0.22	0.17	0.17
2012	0.27	0.35	0.34	0.26	0.25	0.19	0.19	0.32	0.29	0.28	0.23	0.23	0.17	0.23	0.19	0.19
2013	0.32	0.38	0.37	0.30	0.28	0.23	0.23	0.34	0.31	0.30	0.25	0.24	0.18	0.24	0.20	0.20
2014	0.36	0.42	0.40	0.34	0.32	0.26	0.26	0.36	0.33	0.31	0.27	0.26	0.20	0.26	0.22	0.22
2015	0.41	0.45	0.44	0.38	0.36	0.29	0.29	0.38	0.35	0.33	0.29	0.27	0.22	0.27	0.23	0.23
2016	0.47	0.49	0.47	0.42	0.39	0.34	0.33	0.40	0.36	0.35	0.31	0.29	0.24	0.29	0.25	0.25
2017	0.52	0.52	0.51	0.46	0.43	0.38	0.38	0.42	0.38	0.36	0.34	0.30	0.26	0.30	0.26	0.26
2018	0.58	0.56	0.55	0.50	0.46	0.42	0.42	0.44	0.39	0.38	0.36	0.32	0.29	0.32	0.28	0.27
2019	0.63	0.60	0.58	0.55	0.49	0.47	0.46	0.46	0.41	0.40	0.38	0.34	0.31	0.33	0.30	0.28
2020	0.68	0.64	0.61	0.59	0.53	0.51	0.50	0.49	0.43	0.43	0.40	0.36	0.34	0.34	0.32	0.29
2021	0.74	0.68	0.65	0.64	0.56	0.56	0.55	0.51	0.46	0.45	0.42	0.38	0.36	0.36	0.34	0.31
2022	0.83	0.73	0.69	0.68	0.60	0.60	0.59	0.53	0.50	0.47	0.44	0.40	0.39	0.37	0.36	0.33

Year	KEN	UKR	QAT	SVN	MUS	VEN	LTU	JAM	BHR	SRB	EST	CIV	LVA	KAZ	PSE	ISL
2000	0.07	0.01	0.00	0.02	0.00	0.10	0.04	0.01	0.01	0.00	0.03	-	0.01	0.00	-	0.00
2001	0.08	0.01	0.00	0.03	0.01	0.11	0.04	0.02	0.01	0.00	0.04	-	0.01	0.01	-	0.01
2002	0.08	0.01	0.00	0.04	0.01	0.12	0.05	0.02	0.01	0.00	0.04	-	0.01	0.01	-	0.01
2003	0.09	0.01	0.00	0.04	0.01	0.12	0.06	0.02	0.01	0.00	0.05	0.00	0.01	0.01	-	0.01
2004	0.10	0.01	0.01	0.05	0.02	0.13	0.07	0.03	0.02	0.00	0.05	0.00	0.02	0.01	-	0.02
2005	0.11	0.01	0.02	0.06	0.02	0.14	0.07	0.03	0.03	0.00	0.05	0.00	0.02	0.01	-	0.02
2006	0.11	0.01	0.02	0.07	0.03	0.15	0.08	0.04	0.04	0.00	0.05	0.01	0.02	0.02	0.00	0.02
2007	0.12	0.01	0.04	0.08	0.04	0.16	0.09	0.04	0.05	0.00	0.06	0.01	0.03	0.02	0.00	0.03
2008	0.13	0.01	0.05	0.09	0.04	0.16	0.10	0.05	0.06	0.00	0.07	0.01	0.03	0.02	0.00	0.03
2009	0.14	0.02	0.06	0.10	0.05	0.17	0.10	0.06	0.07	0.00	0.07	0.02	0.04	0.02	0.00	0.04
2010	0.15	0.03	0.08	0.12	0.06	0.17	0.11	0.06	0.08	0.00	0.08	0.02	0.05	0.03	0.01	0.04
2011	0.17	0.04	0.10	0.13	0.07	0.18	0.12	0.07	0.09	0.01	0.08	0.03	0.05	0.03	0.01	0.05
2012	0.18	0.06	0.12	0.15	0.08	0.18	0.13	0.08	0.10	0.02	0.09	0.04	0.06	0.04	0.01	0.05
2013	0.19	0.08	0.14	0.16	0.09	0.19	0.14	0.09	0.11	0.02	0.10	0.05	0.07	0.05	0.02	0.06
2014	0.20	0.10	0.15	0.18	0.11	0.20	0.15	0.10	0.12	0.03	0.10	0.06	0.08	0.05	0.03	0.06
2015	0.22	0.12	0.17	0.19	0.12	0.20	0.16	0.11	0.13	0.05	0.10	0.07	0.08	0.06	0.05	0.07
2016	0.23	0.15	0.19	0.20	0.14	0.21	0.18	0.11	0.14	0.06	0.11	0.07	0.09	0.07	0.06	0.08
2017	0.25	0.18	0.20	0.21	0.16	0.21	0.18	0.12	0.15	0.08	0.11	0.08	0.10	0.08	0.07	0.08
2018	0.26	0.21	0.22	0.22	0.18	0.22	0.19	0.14	0.17	0.09	0.12	0.09	0.10	0.09	0.08	0.09
2019	0.28	0.23	0.24	0.24	0.20	0.22	0.20	0.15	0.18	0.11	0.12	0.11	0.11	0.10	0.09	0.10
2020	0.29	0.25	0.25	0.25	0.22	0.23	0.21	0.17	0.19	0.13	0.13	0.12	0.12	0.11	0.10	0.10
2021	0.31	0.27	0.27	0.26	0.25	0.24	0.22	0.19	0.20	0.16	0.14	0.13	0.12	0.11	0.11	0.11
2022	0.32	0.29	0.29	0.28	0.27	0.24	0.24	0.23	0.21	0.20	0.14	0.14	0.13	0.13	0.12	0.12

Year	BWA	MLT	LBN	NAM	ZMB	SVK	GHA	TTO	BMU	TZA	ECU	MWI	IRN	UGA	GGY
2000	0.01	0.03	0.01	0.02	0.02	0.00	0.00	-	-	-	0.00	0.00	-	-	0.00
2001	0.01	0.03	0.01	0.02	0.02	0.00	0.01	0.00	0.00	-	0.00	0.00	-	-	0.00
2002	0.02	0.03	0.02	0.02	0.03	0.00	0.01	0.00	0.00	-	0.00	0.00	-	-	0.00
2003	0.02	0.03	0.02	0.02	0.03	0.00	0.01	0.00	0.00	-	0.00	0.00	0.00	-	0.00
2004	0.02	0.04	0.03	0.02	0.04	0.01	0.01	0.00	0.00	-	0.00	0.00	0.00	-	0.00
2005	0.02	0.04	0.03	0.02	0.04	0.01	0.01	0.00	0.00	-	0.01	0.00	0.00	-	0.00
2006	0.02	0.05	0.03	0.03	0.05	0.01	0.01	0.00	0.00	-	0.01	0.00	0.01	-	0.00
2007	0.03	0.05	0.04	0.03	0.05	0.01	0.02	0.01	0.00	-	0.01	0.00	0.01	-	0.00
2008	0.03	0.06	0.04	0.03	0.05	0.02	0.02	0.01	0.01	-	0.01	0.00	0.01	-	0.00
2009	0.03	0.06	0.04	0.04	0.06	0.02	0.02	0.01	0.01	0.00	0.01	0.00	0.01	-	0.00
2010	0.04	0.07	0.04	0.04	0.06	0.02	0.03	0.02	0.01	0.00	0.01	0.00	0.01	-	0.00
2011	0.04	0.07	0.05	0.04	0.06	0.03	0.03	0.02	0.02	0.00	0.01	0.00	0.01	0.00	0.00
2012	0.05	0.07	0.05	0.04	0.07	0.03	0.04	0.03	0.02	0.00	0.01	0.00	0.01	0.00	0.00
2013	0.05	0.08	0.06	0.05	0.07	0.04	0.04	0.03	0.02	0.00	0.01	0.00	0.02	0.00	0.00
2014	0.06	0.08	0.06	0.05	0.07	0.04	0.05	0.04	0.03	0.00	0.02	0.01	0.02	0.01	0.00
2015	0.06	0.08	0.06	0.06	0.08	0.05	0.05	0.04	0.03	0.01	0.02	0.01	0.02	0.01	0.01
2016	0.07	0.09	0.07	0.06	0.08	0.06	0.06	0.05	0.03	0.01	0.02	0.01	0.02	0.01	0.01
2017	0.07	0.09	0.07	0.07	0.08	0.07	0.07	0.06	0.04	0.01	0.03	0.01	0.02	0.01	0.01
2018	0.08	0.10	0.08	0.08	0.09	0.07	0.07	0.07	0.04	0.02	0.03	0.02	0.03	0.02	0.01
2019	0.09	0.10	0.09	0.09	0.09	0.08	0.08	0.08	0.04	0.02	0.03	0.02	0.03	0.02	0.01
2020	0.10	0.10	0.09	0.09	0.10	0.09	0.09	0.08	0.04	0.03	0.04	0.03	0.03	0.02	0.01
2021	0.11	0.11	0.10	0.10	0.10	0.10	0.10	0.09	0.05	0.04	0.04	0.03	0.03	0.02	0.01
2022	0.11	0.11	0.11	0.11	0.10	0.10	0.10	0.10	0.05	0.05	0.05	0.04	0.03	0.03	0.01

Note: This table presents the cumulative number of valid excess return observations by year and market, with values scaled in millions. Although the dataset begins in 1925, coverage for many non-U.S. markets is limited in the earlier decades. Columns are ordered by their 2022 values. Blank cells denote missing data for the respective year.

Table A.4: Annual Breakdown of Cross-Market Securities

Year	U.S.	GBR	JPN	DEU	CHN	IND	HKG	AUS	KOR	CAN	TWN	THA	FRA	MYS	SWE	SGP
2000	35.76	4.16	5.22	2.25	0.45	1.07	1.19	0.92	1.26	1.96	0.78	1.20	0.87	1.01	0.39	0.46
2001	36.52	4.83	5.69	2.76	0.82	1.18	1.36	1.25	1.69	2.07	1.01	1.27	1.15	1.13	0.51	0.59
2002	37.28	5.08	5.96	2.99	0.90	1.28	1.47	1.39	1.78	2.19	1.16	1.29	1.23	1.19	0.56	0.68
2003	38.05	5.33	6.20	3.18	1.08	1.32	1.76	1.55	1.84	2.32	1.26	1.33	1.29	1.28	0.59	0.74
2004	39.13	5.52	6.42	3.33	1.26	1.44	1.88	1.72	1.97	2.55	1.51	1.45	1.34	1.39	0.62	0.80
2005	40.15	5.83	6.66	3.51	2.27	1.53	1.99	1.88	2.09	2.78	1.61	1.57	1.40	1.48	0.64	0.87
2006	41.35	6.36	6.96	3.76	2.82	1.68	2.15	2.14	2.30	2.99	1.78	1.69	1.49	1.60	0.68	0.96
2007	42.89	6.96	7.37	4.15	3.08	1.96	2.37	2.42	2.49	3.19	2.23	1.80	1.62	1.71	0.76	1.08
2008	43.78	7.37	7.57	4.48	3.25	2.31	2.55	2.63	2.84	3.35	2.35	1.84	1.73	1.77	0.84	1.15
2009	44.53	7.69	7.76	4.77	3.49	2.54	2.69	2.81	2.98	3.53	2.53	1.92	1.80	1.81	0.88	1.22
2010	45.54	8.04	8.00	5.04	3.69	3.04	2.84	3.00	3.15	3.77	2.69	2.00	1.87	1.89	0.96	1.29
2011	47.01	8.44	8.16	5.30	4.28	3.53	3.05	3.20	3.35	4.03	2.86	2.07	1.94	1.96	1.03	1.36
2012	48.40	9.02	8.46	5.72	4.83	4.07	3.28	3.45	3.56	4.22	3.06	2.18	2.05	2.07	1.16	1.43
2013	49.58	9.39	8.69	5.98	5.17	4.69	3.43	3.59	3.70	4.39	3.21	2.24	2.15	2.13	1.21	1.49
2014	50.84	9.88	9.04	6.39	5.51	5.28	3.70	3.85	3.87	4.60	3.38	2.38	2.31	2.22	1.29	1.58
2015	52.02	10.43	9.31	6.76	5.70	5.89	3.93	4.05	4.05	4.77	3.58	2.54	2.45	2.30	1.38	1.66
2016	53.02	11.06	9.61	7.20	6.13	6.32	4.29	4.38	4.26	4.93	3.82	2.69	2.57	2.40	1.51	1.75
2017	54.19	11.71	9.94	7.62	6.65	6.66	4.63	4.66	4.53	5.11	4.03	2.89	2.71	2.47	1.67	1.83
2018	55.38	12.29	10.26	8.09	7.19	7.06	4.92	4.92	4.78	5.25	4.23	3.03	2.82	2.54	1.89	1.92
2019	56.46	12.84	10.50	8.54	7.71	7.43	5.25	5.19	5.02	5.38	4.40	3.18	2.95	2.61	2.08	2.01
2020	57.87	13.58	10.86	9.15	8.17	7.95	5.75	5.49	5.29	5.55	4.63	3.30	3.13	2.68	2.28	2.09
2021	60.71	14.48	11.27	10.10	8.96	8.62	6.27	5.94	5.68	5.70	4.97	3.53	3.42	2.80	2.49	2.19
2022	62.25	15.23	11.65	10.92	9.89	9.10	6.72	6.67	6.04	5.79	5.24	3.64	3.60	2.91	2.69	2.29

Year	POL	CHE	ITA	IDN	ZAF	BRA	VNM	ISR	TUR	PAK	MEX	NOR	ESP	MUL*	NLD	RUS
2000	0.22	0.46	0.44	0.43	0.55	0.39	0.01	0.25	0.28	0.25	0.32	0.20	0.23	0.35	0.35	0.15
2001	0.25	0.52	0.53	0.48	0.66	0.47	0.01	0.27	0.32	0.29	0.35	0.24	0.28	0.52	0.38	0.16
2002	0.27	0.56	0.56	0.51	0.69	0.56	0.01	0.33	0.43	0.30	0.37	0.27	0.29	0.55	0.40	0.16
2003	0.28	0.59	0.59	0.52	0.72	0.58	0.01	0.34	0.44	0.31	0.39	0.30	0.30	0.57	0.41	0.17
2004	0.30	0.61	0.61	0.54	0.74	0.60	0.02	0.35	0.46	0.35	0.41	0.31	0.31	0.58	0.43	0.18
2005	0.34	0.63	0.64	0.56	0.76	0.67	0.02	0.37	0.47	0.40	0.43	0.34	0.33	0.59	0.44	0.20
2006	0.37	0.67	0.69	0.59	0.79	0.71	0.02	0.39	0.51	0.44	0.45	0.38	0.35	0.61	0.46	0.22
2007	0.45	0.73	0.74	0.61	0.85	0.76	0.07	0.47	0.54	0.47	0.48	0.43	0.38	0.64	0.49	0.30
2008	0.52	0.77	0.79	0.64	0.89	0.82	0.08	0.49	0.56	0.51	0.51	0.47	0.41	0.66	0.52	0.33
2009	0.61	0.80	0.82	0.69	0.94	0.85	0.11	0.54	0.57	0.53	0.54	0.50	0.44	0.68	0.53	0.36
2010	0.73	0.85	0.87	0.73	0.97	0.89	0.31	0.61	0.60	0.57	0.57	0.53	0.48	0.71	0.56	0.39
2011	0.82	0.88	0.89	0.77	1.01	0.93	0.47	0.66	0.64	0.60	0.59	0.57	0.51	0.72	0.58	0.45
2012	0.93	0.94	0.95	0.82	1.06	0.98	0.70	0.77	0.70	0.63	0.64	0.62	0.56	0.74	0.60	0.48
2013	1.03	0.99	0.98	0.86	1.11	1.05	0.74	0.84	0.72	0.73	0.68	0.66	0.57	0.76	0.63	0.51
2014	1.12	1.05	1.04	0.92	1.15	1.09	0.80	0.87	0.78	0.76	0.71	0.69	0.60	0.78	0.64	0.54
2015	1.22	1.11	1.11	0.99	1.20	1.15	0.87	0.94	0.84	0.84	0.74	0.73	0.64	0.81	0.67	0.58
2016	1.33	1.18	1.18	1.06	1.25	1.20	0.95	1.00	0.89	0.87	0.78	0.77	0.68	0.82	0.70	0.61
2017	1.45	1.22	1.26	1.13	1.32	1.25	1.03	1.06	0.93	0.91	0.82	0.82	0.72	0.85	0.72	0.65
2018	1.60	1.28	1.33	1.18	1.37	1.30	1.12	1.11	0.97	0.94	0.85	0.86	0.76	0.87	0.75	0.70
2019	1.86	1.34	1.39	1.25	1.42	1.33	1.23	1.16	1.01	0.98	0.88	0.90	0.80	0.89	0.79	0.74
2020	2.03	1.46	1.48	1.35	1.48	1.38	1.35	1.23	1.05	1.03	0.94	0.97	0.85	0.91	0.82	0.77
2021	2.15	1.66	1.63	1.51	1.56	1.46	1.44	1.33	1.11	1.08	1.04	1.04	0.91	0.94	0.88	0.80
2022	2.25	1.78	1.75	1.65	1.61	1.56	1.53	1.50	1.17	1.13	1.12	1.11	0.97	0.96	0.94	0.84

Year	DNK	GRC	AUT	PHL	BGD	LKA	BEL	NZL	FIN	BGR	CHL	EGY	SAU	JOR	KWT	ARG
2000	0.21	0.21	0.25	0.35	0.01	0.25	0.24	0.20	0.19	0.03	0.21	0.09	-	0.08	0.01	0.19
2001	0.26	0.30	0.27	0.37	0.01	0.27	0.29	0.22	0.23	0.03	0.23	0.14	0.00	0.09	0.03	0.21
2002	0.28	0.37	0.28	0.38	0.01	0.29	0.30	0.23	0.24	0.04	0.24	0.14	0.00	0.09	0.06	0.21
2003	0.31	0.41	0.30	0.39	0.07	0.30	0.32	0.23	0.26	0.04	0.25	0.15	0.08	0.10	0.07	0.23
2004	0.32	0.42	0.31	0.39	0.08	0.30	0.33	0.26	0.27	0.04	0.26	0.16	0.08	0.11	0.09	0.24
2005	0.33	0.44	0.32	0.40	0.08	0.32	0.34	0.27	0.28	0.04	0.27	0.17	0.09	0.11	0.10	0.24
2006	0.34	0.45	0.34	0.43	0.10	0.34	0.35	0.29	0.30	0.04	0.27	0.19	0.14	0.14	0.16	0.25
2007	0.37	0.48	0.36	0.44	0.11	0.35	0.39	0.33	0.32	0.16	0.29	0.21	0.16	0.22	0.21	0.26
2008	0.39	0.50	0.38	0.46	0.12	0.36	0.42	0.34	0.33	0.17	0.30	0.22	0.19	0.25	0.23	0.27
2009	0.41	0.51	0.39	0.47	0.13	0.37	0.44	0.35	0.34	0.18	0.31	0.27	0.22	0.28	0.25	0.27
2010	0.43	0.54	0.41	0.49	0.24	0.40	0.46	0.36	0.35	0.22	0.33	0.29	0.25	0.31	0.27	0.29
2011	0.44	0.55	0.43	0.50	0.35	0.42	0.48	0.38	0.37	0.23	0.34	0.30	0.27	0.34	0.29	0.30
2012	0.47	0.58	0.44	0.53	0.37	0.47	0.50	0.41	0.38	0.24	0.36	0.33	0.29	0.36	0.31	0.32
2013	0.50	0.59	0.46	0.55	0.39	0.51	0.51	0.43	0.40	0.26	0.39	0.38	0.31	0.37	0.32	0.33
2014	0.52	0.61	0.48	0.58	0.42	0.53	0.53	0.45	0.41	0.28	0.41	0.40	0.32	0.38	0.34	0.34
2015	0.54	0.63	0.50	0.60	0.45	0.56	0.54	0.47	0.44	0.31	0.42	0.41	0.34	0.40	0.35	0.36
2016	0.57	0.65	0.53	0.62	0.52	0.59	0.56	0.50	0.45	0.33	0.44	0.43	0.37	0.42	0.38	0.37
2017	0.61	0.67	0.55	0.65	0.57	0.62	0.59	0.53	0.48	0.36	0.45	0.44	0.38	0.43	0.39	0.38
2018	0.63	0.70	0.58	0.66	0.60	0.64	0.60	0.55	0.51	0.37	0.47	0.46	0.41	0.44	0.41	0.39
2019	0.66	0.72	0.62	0.68	0.62	0.66	0.63	0.56	0.53	0.39	0.49	0.48	0.43	0.45	0.42	0.41
2020	0.70	0.74	0.65	0.70	0.66	0.68	0.65	0.59	0.56	0.41	0.51	0.51	0.46	0.46	0.43	0.43
2021	0.74	0.77	0.71	0.72	0.71	0.71	0.69	0.62	0.59	0.52	0.54	0.54	0.49	0.48	0.45	0.44
2022	0.80	0.79	0.78	0.76	0.74	0.73	0.72	0.63	0.62	0.58	0.57	0.57	0.52	0.49	0.47	0.46

Note: This table presents the number of securities by year and market, with values scaled in thousands. The numbers are presented cumulatively, showing the number of unique securities present up to and including each year. Although the dataset begins in 1925, coverage for many non-U.S. markets is limited in the earlier decades. Columns are ordered by their 2022 values. Blank cells denote missing data for the respective year. * When a company's securities are listed across multiple countries or exchanges, or on a multilateral trading facility (MTF) that doesn't correspond to a single national market.

Table A.5: Annual Breakdown of Cross-Market Securities (continued)

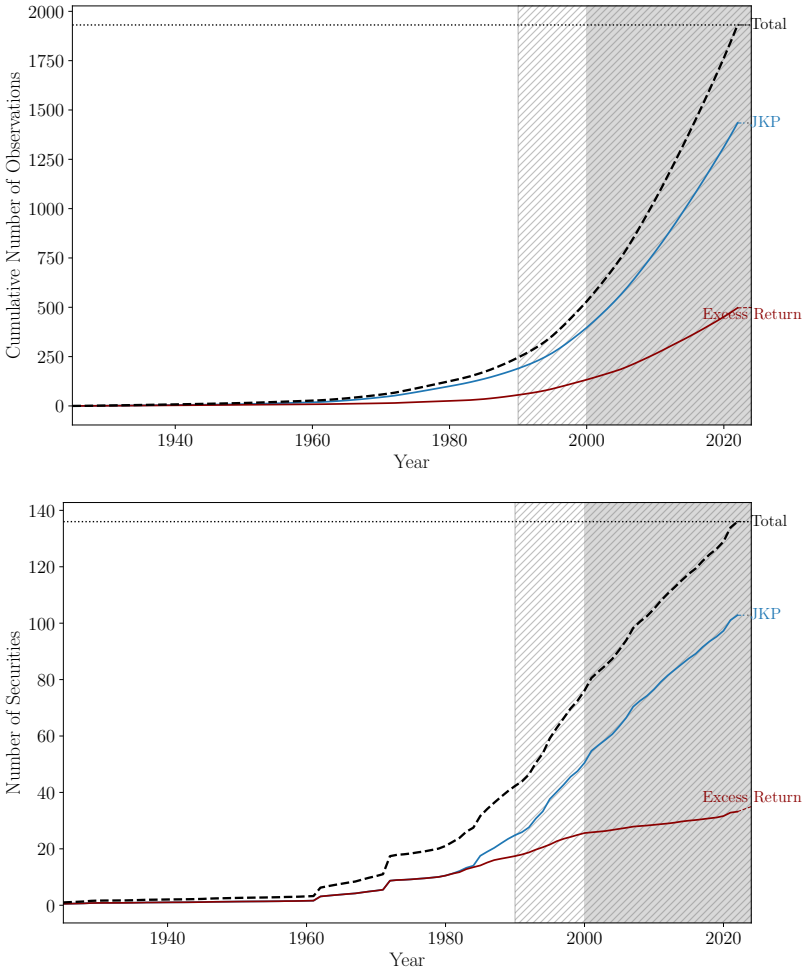
Year	LUX	NGA	ROU	PER	HRV	IRL	ARE	CYP	PRT	OMN	CZE	SRB	HUN	MUS	JAM	COL
2000	0.10	0.03	0.03	0.06	0.03	0.10	0.01	0.06	0.09	0.00	0.09	0.00	0.06	0.00	0.01	0.05
2001	0.11	0.05	0.03	0.07	0.03	0.12	0.01	0.09	0.10	0.01	0.10	0.00	0.06	0.02	0.02	0.06
2002	0.12	0.06	0.04	0.10	0.04	0.12	0.02	0.09	0.11	0.04	0.10	0.00	0.07	0.02	0.02	0.06
2003	0.13	0.06	0.04	0.11	0.04	0.13	0.03	0.10	0.12	0.05	0.11	0.00	0.07	0.02	0.02	0.06
2004	0.14	0.07	0.05	0.12	0.04	0.14	0.03	0.10	0.12	0.05	0.12	0.00	0.08	0.02	0.02	0.07
2005	0.15	0.07	0.06	0.12	0.05	0.14	0.03	0.10	0.13	0.06	0.12	0.00	0.08	0.03	0.02	0.07
2006	0.16	0.07	0.06	0.13	0.06	0.15	0.06	0.11	0.13	0.08	0.12	0.00	0.08	0.03	0.03	0.08
2007	0.17	0.13	0.07	0.15	0.09	0.17	0.12	0.11	0.14	0.08	0.13	0.00	0.09	0.03	0.03	0.08
2008	0.18	0.18	0.07	0.17	0.13	0.18	0.14	0.12	0.15	0.08	0.13	0.01	0.10	0.04	0.03	0.09
2009	0.19	0.20	0.08	0.18	0.14	0.19	0.15	0.13	0.15	0.09	0.15	0.01	0.10	0.04	0.04	0.10
2010	0.21	0.23	0.12	0.19	0.16	0.19	0.17	0.14	0.16	0.10	0.16	0.01	0.11	0.06	0.04	0.11
2011	0.23	0.24	0.15	0.20	0.17	0.20	0.18	0.14	0.17	0.12	0.16	0.02	0.12	0.06	0.05	0.11
2012	0.25	0.25	0.16	0.21	0.18	0.21	0.20	0.15	0.18	0.13	0.17	0.02	0.13	0.07	0.05	0.12
2013	0.27	0.27	0.18	0.23	0.19	0.22	0.20	0.17	0.18	0.14	0.18	0.04	0.14	0.08	0.06	0.12
2014	0.28	0.28	0.19	0.24	0.20	0.22	0.21	0.17	0.19	0.15	0.18	0.06	0.15	0.09	0.06	0.13
2015	0.30	0.29	0.22	0.25	0.21	0.23	0.22	0.17	0.20	0.16	0.19	0.07	0.16	0.09	0.07	0.14
2016	0.32	0.31	0.25	0.27	0.23	0.25	0.24	0.18	0.21	0.18	0.19	0.08	0.16	0.10	0.07	0.14
2017	0.33	0.33	0.27	0.28	0.24	0.26	0.26	0.18	0.22	0.20	0.20	0.09	0.17	0.11	0.08	0.15
2018	0.35	0.34	0.29	0.29	0.25	0.27	0.27	0.19	0.23	0.21	0.20	0.10	0.18	0.14	0.09	0.16
2019	0.37	0.35	0.31	0.29	0.27	0.28	0.28	0.20	0.23	0.21	0.21	0.10	0.19	0.14	0.10	0.16
2020	0.38	0.36	0.33	0.30	0.29	0.29	0.29	0.21	0.24	0.22	0.22	0.13	0.20	0.15	0.11	0.17
2021	0.42	0.38	0.36	0.36	0.31	0.30	0.31	0.26	0.26	0.24	0.23	0.20	0.21	0.17	0.14	0.19
2022	0.45	0.40	0.39	0.38	0.34	0.32	0.32	0.28	0.26	0.25	0.25	0.22	0.22	0.20	0.20	0.20

Year	UKR	TUN	MAR	SVN	KEN	ZWE	QAT	LTU	VEN	BHR	KAZ	ISL	EST	SVK	CIV	NAM
2000	0.01	0.03	0.02	0.03	0.05	0.05	0.00	0.04	0.06	0.01	0.01	0.01	0.03	0.03	-	0.02
2001	0.01	0.03	0.04	0.04	0.05	0.06	0.01	0.04	0.07	0.01	0.01	0.01	0.04	0.03	-	0.02
2002	0.01	0.03	0.04	0.04	0.06	0.06	0.01	0.05	0.08	0.01	0.01	0.02	0.04	0.03	-	0.02
2003	0.01	0.05	0.04	0.05	0.06	0.06	0.01	0.05	0.08	0.04	0.01	0.02	0.04	0.04	0.00	0.02
2004	0.01	0.05	0.05	0.05	0.06	0.07	0.03	0.05	0.08	0.04	0.01	0.02	0.04	0.04	0.01	0.02
2005	0.01	0.05	0.05	0.05	0.06	0.07	0.03	0.05	0.08	0.04	0.01	0.02	0.04	0.04	0.01	0.02
2006	0.01	0.06	0.06	0.07	0.07	0.08	0.06	0.06	0.08	0.04	0.01	0.03	0.04	0.04	0.02	0.03
2007	0.01	0.07	0.08	0.07	0.08	0.10	0.06	0.06	0.09	0.05	0.02	0.03	0.05	0.05	0.02	0.03
2008	0.02	0.07	0.09	0.08	0.09	0.10	0.07	0.07	0.09	0.05	0.02	0.03	0.05	0.05	0.02	0.03
2009	0.02	0.08	0.11	0.09	0.09	0.10	0.08	0.07	0.09	0.06	0.02	0.04	0.05	0.05	0.02	0.03
2010	0.06	0.09	0.12	0.09	0.09	0.11	0.08	0.07	0.09	0.06	0.03	0.04	0.06	0.05	0.03	0.03
2011	0.07	0.10	0.12	0.10	0.10	0.11	0.08	0.07	0.10	0.07	0.03	0.04	0.06	0.05	0.04	0.04
2012	0.08	0.10	0.12	0.11	0.10	0.12	0.09	0.08	0.10	0.07	0.04	0.04	0.06	0.06	0.04	0.04
2013	0.09	0.11	0.13	0.11	0.12	0.12	0.09	0.08	0.10	0.07	0.04	0.05	0.06	0.06	0.05	0.04
2014	0.10	0.12	0.14	0.11	0.13	0.13	0.10	0.09	0.10	0.07	0.04	0.06	0.06	0.07	0.05	0.05
2015	0.11	0.13	0.14	0.12	0.13	0.13	0.10	0.09	0.11	0.08	0.05	0.06	0.07	0.07	0.06	0.05
2016	0.14	0.14	0.15	0.12	0.14	0.14	0.10	0.09	0.11	0.08	0.06	0.07	0.07	0.08	0.06	0.06
2017	0.16	0.15	0.17	0.13	0.14	0.14	0.11	0.10	0.11	0.08	0.07	0.07	0.07	0.08	0.07	0.06
2018	0.16	0.16	0.17	0.13	0.15	0.14	0.11	0.10	0.11	0.08	0.07	0.07	0.08	0.08	0.07	0.06
2019	0.17	0.17	0.17	0.13	0.15	0.15	0.11	0.10	0.11	0.09	0.07	0.08	0.08	0.08	0.07	0.06
2020	0.17	0.17	0.18	0.14	0.16	0.15	0.12	0.11	0.11	0.09	0.08	0.09	0.08	0.08	0.08	0.06
2021	0.18	0.19	0.18	0.15	0.16	0.16	0.12	0.12	0.12	0.10	0.10	0.10	0.09	0.09	0.08	0.07
2022	0.19	0.19	0.19	0.18	0.16	0.16	0.12	0.12	0.12	0.11	0.10	0.10	0.10	0.09	0.08	0.08

Year	BWA	MLT	GHA	PSE	TTO	ZMB	TZA	LVA	LBN	IRN	MWI	ECU	UGA	BMU	GGY
2000	0.01	0.02	0.01	-	-	0.00	-	0.01	0.02	-	0.00	0.00	-	-	0.00
2001	0.01	0.02	0.01	-	0.00	0.01	-	0.01	0.02	-	0.00	0.00	-	0.00	0.00
2002	0.02	0.02	0.01	-	0.00	0.01	-	0.01	0.02	-	0.00	0.00	-	0.00	0.00
2003	0.02	0.02	0.01	-	0.00	0.01	-	0.01	0.02	0.01	0.00	0.00	-	0.00	0.00
2004	0.02	0.02	0.02	-	0.00	0.01	-	0.01	0.02	0.01	0.00	0.00	-	0.00	0.00
2005	0.02	0.02	0.02	-	0.00	0.01	-	0.01	0.02	0.01	0.00	0.00	-	0.00	0.00
2006	0.02	0.02	0.02	0.00	0.01	0.01	-	0.01	0.03	0.01	0.00	0.00	-	0.00	0.00
2007	0.02	0.02	0.02	0.00	0.01	0.01	-	0.02	0.03	0.02	0.00	0.00	-	0.01	0.00
2008	0.02	0.03	0.02	0.01	0.01	0.01	-	0.03	0.03	0.02	0.00	0.00	-	0.01	0.00
2009	0.02	0.03	0.02	0.01	0.02	0.01	0.00	0.03	0.03	0.02	0.00	0.01	-	0.01	0.01
2010	0.03	0.03	0.03	0.01	0.02	0.02	0.00	0.03	0.03	0.02	0.00	0.01	-	0.02	0.01
2011	0.03	0.03	0.03	0.01	0.02	0.02	0.00	0.04	0.03	0.03	0.00	0.01	0.00	0.02	0.01
2012	0.04	0.03	0.03	0.03	0.03	0.02	0.00	0.04	0.03	0.03	0.00	0.01	0.01	0.02	0.01
2013	0.04	0.03	0.04	0.04	0.03	0.03	0.01	0.04	0.04	0.03	0.02	0.02	0.01	0.02	0.01
2014	0.04	0.04	0.04	0.05	0.03	0.03	0.01	0.04	0.04	0.03	0.02	0.02	0.01	0.02	0.01
2015	0.04	0.04	0.04	0.05	0.03	0.03	0.01	0.04	0.04	0.03	0.02	0.02	0.02	0.02	0.01
2016	0.04	0.05	0.04	0.05	0.04	0.03	0.01	0.04	0.04	0.04	0.02	0.02	0.02	0.02	0.01
2017	0.05	0.05	0.05	0.05	0.04	0.04	0.02	0.05	0.04	0.04	0.02	0.02	0.02	0.02	0.01
2018	0.05	0.05	0.05	0.05	0.05	0.04	0.02	0.05	0.04	0.04	0.02	0.02	0.02	0.02	0.01
2019	0.06	0.05	0.05	0.05	0.05	0.04	0.03	0.05	0.04	0.04	0.02	0.02	0.02	0.02	0.01
2020	0.06	0.05	0.06	0.06	0.05	0.04	0.04	0.05	0.04	0.04	0.03	0.02	0.02	0.02	0.01
2021	0.07	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.04	0.04	0.03	0.03	0.02	0.02	0.01
2022	0.07	0.07	0.06	0.06	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.03	0.02	0.02	0.01

Note: This table presents the number of securities by year and market, with values scaled in thousands. The numbers are presented cumulatively, showing the number of unique securities present up to and including each year. Although the dataset begins in 1925, coverage for many non-U.S. markets is limited in the earlier decades. Columns are ordered by their 2022 values. Blank cells denote missing data for the respective year.

Figure A.3: Cumulative Number of Observations and Unique Securities: Excess Returns vs. JKP



Note: This top figure displays the cumulative number of excess return observations, JKP observations as defined by Jensen et al. (2023), and the total number of observations spanning the period from 1925 to 2022. All numbers are reported across all markets combined. The dashed line indicates the aggregate total of observations compiled from all data sources. The maximum number of observations recorded in 2022 is indicated by a dashed horizontal line, corresponding to 1930.95 million observations. The bottom figure displays the number of unique securities (in thousands) over the same period. The maximum number of unique securities in 2022 is similarly shown, corresponding to 135.99 thousand securities. In both figures, the light-shaded region represents the sample period starting in 1990 (expanding window) used to train the predictive models, while the dark-shaded region denotes the period from 2001 to 2022, during which the predictive models are trained on a yearly basis.

Table A.6: Annual Breakdown of JKP Observations

Year	U.S.	JPN	CHN	IND	CAN	GBR	TWN	HKG	AUS	KOR	MYS	DEU	FRA	SGP	THA	SWE	POL
2000	282.76	29.50	0.73	1.64	15.95	14.92	1.71	3.45	4.17	2.94	4.00	3.89	4.91	2.09	2.14	1.54	0.37
2001	293.08	33.69	1.90	2.08	17.74	16.99	2.13	4.17	4.96	3.65	4.79	4.83	5.85	2.52	2.49	1.92	0.50
2002	302.70	38.17	3.37	2.63	19.62	19.09	2.65	5.02	6.04	4.46	5.67	5.84	6.81	3.04	2.86	2.31	0.63
2003	311.72	42.79	5.03	3.35	21.59	21.15	3.36	6.06	7.42	5.36	6.66	6.81	7.74	3.61	3.27	2.71	0.78
2004	320.51	47.61	6.88	4.26	23.69	23.25	4.42	7.23	8.97	6.34	7.80	7.78	8.66	4.26	3.73	3.12	0.96
2005	329.28	52.62	8.87	5.33	25.90	25.50	5.67	8.49	10.72	7.37	9.07	8.77	9.59	5.01	4.28	3.54	1.17
2006	337.93	57.81	10.90	6.60	28.26	27.99	7.17	9.87	12.66	8.51	10.45	9.82	10.56	5.86	4.89	4.00	1.42
2007	346.43	63.12	13.04	8.26	30.72	30.64	8.97	11.44	14.79	9.79	11.89	10.99	11.61	6.82	5.55	4.52	1.75
2008	354.69	68.44	15.35	10.44	33.19	33.29	10.93	13.14	17.00	11.29	13.25	12.23	12.68	7.80	6.23	5.07	2.17
2009	362.55	73.65	17.79	12.99	35.58	35.72	13.06	14.90	19.19	12.88	14.57	13.46	13.72	8.76	6.93	5.64	2.69
2010	370.15	78.77	20.53	15.92	37.97	38.04	15.37	16.79	21.46	14.51	15.94	14.67	14.76	9.76	7.66	6.24	3.27
2011	377.56	83.83	23.78	19.29	40.40	40.29	17.85	18.80	23.74	16.24	17.33	15.90	15.79	10.75	8.41	6.86	3.96
2012	384.81	88.82	27.41	23.30	42.84	42.47	20.46	20.86	25.98	18.12	18.69	17.10	16.80	11.74	9.17	7.48	4.79
2013	391.92	93.85	31.19	27.59	45.40	44.62	23.16	23.04	28.18	20.17	20.04	18.25	17.78	12.74	9.99	8.13	5.69
2014	399.11	98.91	35.01	32.28	47.99	46.81	25.96	25.40	30.37	22.37	21.41	19.36	18.77	13.74	10.84	8.82	6.72
2015	406.39	104.05	38.85	37.01	50.51	49.00	28.85	27.91	32.54	24.68	22.78	20.43	19.79	14.72	11.75	9.59	7.78
2016	413.51	109.23	43.09	41.72	52.91	51.14	31.83	30.59	34.74	27.13	24.13	21.45	20.82	15.66	12.70	10.47	8.84
2017	420.51	114.48	47.98	46.52	55.21	53.22	34.92	33.44	36.99	29.74	25.50	22.45	21.85	16.57	13.68	11.46	9.91
2018	427.48	119.80	53.40	51.42	57.44	55.31	38.04	36.50	39.27	32.52	26.85	23.43	22.87	17.44	14.71	12.39	10.92
2019	434.45	125.18	59.25	56.38	59.58	57.35	41.19	39.72	41.48	35.40	28.21	24.36	23.89	18.27	15.78	13.37	11.89
2020	441.40	130.63	65.53	61.34	61.61	59.31	44.42	43.11	43.72	38.37	29.58	25.31	24.90	19.06	16.88	14.39	12.87
2021	448.76	136.14	72.51	66.72	63.60	61.28	47.71	46.55	46.10	41.46	31.00	26.26	25.90	19.86	18.05	15.50	13.88
2022	456.62	141.68	80.17	72.39	65.56	63.26	51.07	49.98	48.58	44.68	32.41	27.20	26.91	20.62	19.26	16.72	14.90

Year	IDN	ITA	ZAF	TUR	ISR	PAK	CHE	VNM	NOR	GRC	PHL	NLD	ESP	DNK	LKA	FIN	BEL
2000	1.12	1.75	2.04	0.53	0.33	0.46	1.54	-	0.98	0.47	0.65	1.90	1.36	1.03	0.25	0.76	0.90
2001	1.37	2.07	2.42	0.66	0.44	0.60	1.83	-	1.16	0.67	0.79	2.16	1.55	1.20	0.30	0.93	1.07
2002	1.63	2.42	2.79	0.84	0.55	0.75	2.15	-	1.36	0.94	0.93	2.41	1.73	1.39	0.35	1.10	1.24
2003	1.93	2.76	3.15	1.02	0.69	0.90	2.48	-	1.56	1.22	1.09	2.64	1.91	1.59	0.41	1.28	1.41
2004	2.24	3.10	3.50	1.25	0.84	1.09	2.80	-	1.76	1.52	1.25	2.87	2.09	1.79	0.49	1.45	1.58
2005	2.56	3.46	3.87	1.51	1.03	1.31	3.13	-	1.99	1.84	1.44	3.09	2.28	2.01	0.59	1.64	1.77
2006	2.89	3.83	4.24	1.84	1.28	1.55	3.48	0.00	2.26	2.19	1.65	3.31	2.47	2.23	0.72	1.84	1.96
2007	3.26	4.24	4.64	2.21	1.57	1.83	3.84	0.03	2.56	2.56	1.87	3.53	2.67	2.48	0.89	2.04	2.16
2008	3.64	4.67	5.06	2.59	1.90	2.09	4.21	0.11	2.88	2.93	2.10	3.74	2.88	2.74	1.10	2.25	2.36
2009	4.02	5.08	5.49	2.97	2.28	2.42	4.58	0.38	3.19	3.31	2.33	3.94	3.08	2.99	1.34	2.45	2.56
2010	4.45	5.49	5.93	3.39	2.73	2.81	4.94	0.79	3.49	3.67	2.58	4.13	3.29	3.24	1.61	2.65	2.76
2011	4.94	5.90	6.37	3.85	3.25	3.20	5.30	1.32	3.80	4.00	2.86	4.32	3.49	3.49	1.92	2.84	2.96
2012	5.46	6.29	6.81	4.36	3.80	3.65	5.65	1.96	4.10	4.30	3.15	4.50	3.70	3.73	2.26	3.03	3.16
2013	6.04	6.68	7.24	4.91	4.35	4.17	6.00	2.63	4.39	4.59	3.45	4.67	3.89	3.96	2.63	3.23	3.35
2014	6.65	7.09	7.66	5.50	4.91	4.72	6.35	3.34	4.69	4.87	3.76	4.84	4.09	4.18	3.00	3.43	3.54
2015	7.27	7.52	8.07	6.11	5.46	5.28	6.70	4.04	4.98	5.09	4.08	5.00	4.31	4.40	3.37	3.63	3.74
2016	7.92	7.96	8.48	6.71	6.01	5.84	7.04	4.73	5.28	5.31	4.41	5.17	4.54	4.61	3.74	3.85	3.93
2017	8.62	8.43	8.89	7.31	6.59	6.42	7.37	5.39	5.58	5.55	4.75	5.33	4.78	4.82	4.10	4.08	4.13
2018	9.38	8.91	9.30	7.92	7.16	6.98	7.71	6.05	5.91	5.78	5.09	5.50	5.04	5.03	4.47	4.31	4.32
2019	10.21	9.42	9.70	8.52	7.74	7.55	8.06	6.74	6.24	6.00	5.43	5.66	5.31	5.25	4.83	4.55	4.51
2020	11.09	9.95	10.08	9.12	8.30	8.12	8.40	7.44	6.60	6.21	5.77	5.82	5.58	5.46	5.15	4.79	4.69
2021	12.05	10.51	10.45	9.74	8.92	8.69	8.75	8.19	7.02	6.42	6.10	5.98	5.85	5.70	5.52	5.04	4.88
2022	13.10	11.09	10.79	10.42	9.61	9.26	9.09	8.96	7.50	6.63	6.45	6.14	6.12	5.95	5.89	5.31	5.06

Year	BGD	BRA	CHL	NZL	SAU	EGY	MEX	JOR	KWT	AUT	PER	PRT	OMN	ROU	ARE	HRV	MAR
2000	0.00	0.14	0.52	0.63	0.00	0.04	0.57	0.07	0.00	0.59	0.12	0.41	0.00	0.02	0.00	0.02	0.03
2001	0.00	0.19	0.66	0.74	0.00	0.07	0.66	0.10	0.01	0.69	0.16	0.47	0.01	0.03	0.00	0.02	0.05
2002	0.01	0.23	0.79	0.86	0.03	0.11	0.75	0.15	0.03	0.78	0.20	0.54	0.04	0.04	0.01	0.04	0.08
2003	0.02	0.28	0.93	0.98	0.08	0.15	0.86	0.20	0.06	0.86	0.25	0.60	0.08	0.05	0.01	0.05	0.10
2004	0.04	0.34	1.08	1.11	0.13	0.20	0.97	0.26	0.11	0.95	0.31	0.67	0.13	0.06	0.03	0.07	0.14
2005	0.07	0.41	1.25	1.26	0.20	0.27	1.08	0.34	0.16	1.04	0.37	0.74	0.19	0.09	0.04	0.09	0.18
2006	0.10	0.51	1.42	1.40	0.29	0.36	1.21	0.44	0.27	1.14	0.44	0.81	0.26	0.11	0.09	0.12	0.24
2007	0.14	0.65	1.60	1.56	0.40	0.47	1.33	0.59	0.41	1.25	0.52	0.88	0.33	0.15	0.16	0.19	0.31
2008	0.21	0.83	1.77	1.71	0.54	0.62	1.46	0.77	0.58	1.37	0.61	0.96	0.41	0.18	0.25	0.29	0.40
2009	0.32	1.03	1.95	1.86	0.70	0.76	1.59	0.97	0.75	1.48	0.69	1.03	0.49	0.26	0.34	0.40	0.49
2010	0.51	1.25	2.14	2.01	0.88	0.88	1.73	1.16	0.93	1.59	0.78	1.10	0.58	0.36	0.42	0.50	0.58
2011	0.73	1.48	2.33	2.16	1.07	0.99	1.87	1.33	1.10	1.70	0.87	1.18	0.67	0.49	0.51	0.61	0.67
2012	0.98	1.73	2.53	2.31	1.27	1.16	2.01	1.49	1.28	1.80	0.97	1.25	0.78	0.61	0.61	0.71	0.77
2013	1.27	1.99	2.73	2.47	1.48	1.38	2.16	1.68	1.48	1.90	1.06	1.32	0.89	0.76	0.70	0.82	0.85
2014	1.58	2.26	2.93	2.64	1.71	1.61	2.31	1.88	1.69	2.00	1.15	1.39	1.00	0.92	0.80	0.93	0.94
2015	1.93	2.53	3.13	2.82	1.94	1.86	2.47	2.09	1.90	2.10	1.23	1.47	1.10	1.06	0.90	1.04	1.03
2016	2.29	2.80	3.32	3.00	2.19	2.11	2.63	2.29	2.09	2.19	1.33	1.54	1.20	1.16	1.00	1.14	1.12
2017	2.65	3.08	3.52	3.18	2.44	2.37	2.81	2.49	2.28	2.28	1.43	1.61	1.30	1.26	1.11	1.24	1.21
2018	3.04	3.37	3.74	3.36	2.71	2.63	2.98	2.69	2.47	2.37	1.54	1.68	1.41	1.36	1.23	1.33	1.30
2019	3.46	3.66	3.96	3.53	2.98	2.90	3.15	2.87	2.66	2.46	1.67	1.75	1.53	1.46	1.34	1.41	1.38
2020	3.83	3.96	4.18	3.70	3.26	3.17	3.33	3.02	2.85	2.54	1.79	1.81	1.66	1.56	1.45	1.50	1.47
2021	4.26	4.32	4.40	3.87	3.56	3.44	3.50	3.19	3.04	2.63	1.92	1.87	1.78	1.66	1.58	1.58	1.56
2022	4.71	4.71	4.61	4.05	3.90	3.70	3.68	3.35	3.23	2.72	2.04	1.93	1.91	1.77	1.71	1.66	1.65

Note: This table presents the cumulative number of valid JKP observations by year and market, with all values scaled in millions. JKP factors are calculated based on the definitions provided in Jensen et al. (2023). Although the dataset begins in 1925, coverage for many non-U.S. markets is limited in the earlier decades. To improve clarity, data prior to 2000 are excluded from the table. Columns are ordered by their 2022 values. Blank cells denote missing data for the respective year.

Table A.7: Annual Breakdown of JKP Observations (continued)

Year	IRL	TUN	KEN	COL	MUS	HUN	QAT	LTU	SVN	CZE	LVA	CIV	EST	BHR	ISL	SRB	KAZ
2000	0.41	0.02	0.04	0.08	0.01	0.11	-	0.02	0.04	0.30	0.01	0.00	0.03	0.00	0.00	-	-
2001	0.47	0.03	0.05	0.09	0.01	0.15	0.00	0.03	0.06	0.34	0.02	0.00	0.04	0.00	0.01	-	-
2002	0.54	0.04	0.06	0.11	0.02	0.18	0.00	0.04	0.09	0.36	0.03	0.00	0.05	0.01	0.03	-	-
2003	0.60	0.05	0.08	0.13	0.03	0.21	0.01	0.06	0.11	0.39	0.04	0.00	0.06	0.02	0.05	-	-
2004	0.66	0.07	0.10	0.16	0.05	0.25	0.01	0.09	0.14	0.40	0.05	0.00	0.08	0.04	0.07	-	-
2005	0.72	0.09	0.12	0.19	0.06	0.29	0.03	0.13	0.17	0.43	0.07	0.00	0.10	0.06	0.09	-	-
2006	0.79	0.12	0.15	0.22	0.08	0.32	0.06	0.17	0.21	0.45	0.09	0.00	0.12	0.08	0.11	-	-
2007	0.86	0.16	0.19	0.25	0.10	0.36	0.09	0.22	0.26	0.46	0.12	0.01	0.14	0.11	0.13	-	-
2008	0.93	0.21	0.23	0.29	0.13	0.39	0.13	0.27	0.30	0.48	0.16	0.02	0.17	0.14	0.15	-	-
2009	0.99	0.26	0.27	0.33	0.16	0.43	0.18	0.32	0.34	0.50	0.19	0.03	0.19	0.16	0.16	0.00	0.00
2010	1.06	0.32	0.33	0.37	0.20	0.47	0.22	0.37	0.38	0.51	0.23	0.05	0.22	0.19	0.17	0.01	0.01
2011	1.11	0.38	0.38	0.42	0.24	0.51	0.28	0.42	0.43	0.53	0.26	0.07	0.24	0.22	0.17	0.02	0.02
2012	1.17	0.45	0.45	0.47	0.29	0.54	0.33	0.46	0.47	0.55	0.29	0.09	0.26	0.24	0.18	0.04	0.03
2013	1.21	0.52	0.52	0.52	0.33	0.58	0.39	0.51	0.50	0.57	0.33	0.13	0.28	0.27	0.20	0.07	0.05
2014	1.26	0.60	0.59	0.57	0.38	0.61	0.44	0.55	0.54	0.59	0.36	0.16	0.30	0.29	0.22	0.10	0.06
2015	1.31	0.70	0.66	0.61	0.43	0.65	0.50	0.59	0.58	0.61	0.39	0.21	0.33	0.32	0.24	0.12	0.07
2016	1.36	0.79	0.73	0.67	0.50	0.68	0.56	0.63	0.61	0.63	0.41	0.25	0.35	0.35	0.26	0.15	0.08
2017	1.41	0.89	0.80	0.72	0.57	0.72	0.62	0.67	0.64	0.65	0.44	0.30	0.37	0.37	0.29	0.18	0.10
2018	1.45	0.99	0.87	0.77	0.64	0.77	0.68	0.71	0.67	0.67	0.46	0.34	0.40	0.39	0.31	0.21	0.11
2019	1.50	1.08	0.94	0.83	0.71	0.81	0.74	0.74	0.70	0.68	0.49	0.39	0.42	0.42	0.34	0.23	0.13
2020	1.54	1.17	1.01	0.90	0.79	0.85	0.80	0.78	0.72	0.70	0.51	0.44	0.44	0.45	0.37	0.26	0.14
2021	1.58	1.26	1.07	0.96	0.87	0.89	0.86	0.82	0.75	0.72	0.54	0.49	0.47	0.47	0.40	0.28	0.16
2022	1.61	1.35	1.14	1.02	0.94	0.94	0.92	0.86	0.78	0.74	0.55	0.54	0.51	0.50	0.44	0.30	0.17

Note: This table presents the cumulative number of valid JKP observations by year and market, with all values scaled in millions. JKP factors are calculated based on the definitions provided in Jensen et al. (2023). Although the dataset begins in 1925, coverage for many non-U.S. markets is limited in the earlier decades. To improve clarity, data prior to 2000 are excluded from the table. Columns are ordered by their 2022 values. Blank cells denote missing data for the respective year.

Table A.8: Annual Breakdown of JKP Securities

Year	U.S.	JPN	CHN	IND	CAN	GBR	TWN	HKG	AUS	KOR	MYS	DEU	FRA	SGP	THA	SWE	POL
2000	25.60	3.54	3.35	3.60	0.77	0.18	0.91	0.85	0.66	0.48	1.17	0.89	0.42	0.69	0.17	0.35	0.52
2001	25.85	3.68	3.46	3.83	0.85	1.09	1.27	1.26	0.75	0.73	1.24	0.99	0.46	0.74	0.18	0.50	0.53
2002	26.08	3.83	3.56	4.00	0.89	1.18	1.67	1.28	0.99	0.78	1.26	1.01	0.47	0.86	0.19	0.54	0.54
2003	26.32	3.99	3.66	4.14	1.02	1.26	1.77	1.42	1.08	0.97	1.28	1.02	0.48	0.93	0.20	0.57	0.58
2004	26.69	4.21	3.89	4.32	1.09	1.39	1.92	1.46	1.15	1.03	1.30	1.04	0.49	0.99	0.23	0.66	0.63
2005	27.08	4.48	4.26	4.50	1.20	1.44	2.11	1.63	1.23	1.14	1.36	1.08	0.52	1.09	0.26	0.73	0.69
2006	27.45	4.73	4.58	4.71	1.33	1.49	2.29	1.76	1.29	1.43	1.45	1.20	0.57	1.12	0.30	0.79	0.70
2007	27.87	4.98	4.87	4.84	1.78	1.61	2.52	2.13	1.39	1.58	1.59	1.36	0.65	1.15	0.39	0.86	0.72
2008	28.09	5.14	4.96	4.90	2.12	1.68	2.61	2.19	1.42	1.73	1.62	1.43	0.67	1.17	0.50	0.89	0.74
2009	28.27	5.26	5.00	4.92	2.30	1.78	2.65	2.27	1.48	1.78	1.64	1.47	0.71	1.19	0.56	0.91	0.76
2010	28.51	5.48	5.07	4.95	2.60	2.13	2.75	2.38	1.56	1.89	1.66	1.51	0.74	1.22	0.62	0.95	0.77
2011	28.76	5.69	5.16	4.99	3.01	2.41	2.85	2.46	1.64	1.99	1.70	1.58	0.79	1.25	0.78	0.97	0.78
2012	28.99	5.82	5.23	5.04	3.70	2.56	2.90	2.50	1.70	2.07	1.72	1.61	0.82	1.27	0.89	0.99	0.81
2013	29.30	5.92	5.33	5.11	4.16	2.57	2.97	2.54	1.79	2.12	1.77	1.65	0.86	1.29	0.93	1.01	0.85
2014	29.70	5.99	5.47	5.19	4.30	2.69	3.04	2.61	1.92	2.22	1.80	1.68	0.93	1.30	1.03	1.04	0.89
2015	30.00	6.06	5.56	5.29	4.47	2.92	3.14	2.73	2.04	2.30	1.85	1.71	1.02	1.31	1.07	1.06	0.93
2016	30.22	6.12	5.63	5.39	4.63	3.14	3.23	2.84	2.16	2.37	1.87	1.73	1.13	1.33	1.11	1.07	0.96
2017	30.51	6.22	5.73	5.49	4.86	3.58	3.34	3.04	2.32	2.42	1.90	1.76	1.24	1.34	1.14	1.10	1.00
2018	30.83	6.33	5.81	5.59	5.10	3.68	3.43	3.16	2.52	2.49	1.92	1.80	1.29	1.37	1.16	1.11	1.03
2019	31.11	6.41	5.85	5.69	5.20	3.88	3.49	3.30	2.68	2.54	1.94	1.82	1.32	1.40	1.18	1.13	1.06
2020	31.63	6.49	5.90	5.78	5.28	4.27	3.59	3.41	2.82	2.58	1.95	1.84	1.37	1.42	1.19	1.14	1.09
2021	32.84	6.61	6.05	5.91	5.43	4.75	3.79	3.54	2.91	2.66	2.00	1.90	1.52	1.44	1.24	1.15	1.13
2022	33.18	6.65	6.10	5.98	5.58	5.06	3.87	3.65	2.96	2.71	2.02	1.92	1.56	1.48	1.25	1.16	1.16

Year	IDN	ITA	ZAF	TUR	ISR	PAK	CHE	VNM	NOR	GRC	PHL	NLD	ESP	DNK	LKA	FIN	BEL
2000	0.82	-	0.30	0.19	0.38	0.24	0.21	0.23	0.30	0.22	0.19	0.19	0.38	0.29	0.05	0.00	0.20
2001	0.83	-	0.33	0.26	0.43	0.28	0.30	0.25	0.35	0.24	0.35	0.23	0.38	0.31	0.08	0.00	0.21
2002	0.83	-	0.35	0.26	0.44	0.32	0.31	0.25	0.36	0.24	0.37	0.27	0.38	0.31	0.09	0.05	0.21
2003	0.84	-	0.38	0.28	0.45	0.32	0.31	0.28	0.36	0.25	0.38	0.27	0.38	0.31	0.09	0.06	0.22
2004	0.85	-	0.39	0.28	0.46	0.34	0.32	0.32	0.37	0.25	0.39	0.27	0.38	0.31	0.10	0.06	0.22
2005	0.85	-	0.40	0.30	0.48	0.38	0.33	0.35	0.38	0.26	0.39	0.28	0.38	0.32	0.12	0.07	0.22
2006	0.87	0.00	0.41	0.35	0.50	0.42	0.34	0.36	0.40	0.27	0.40	0.31	0.38	0.33	0.14	0.08	0.23
2007	0.93	0.05	0.44	0.41	0.53	0.46	0.35	0.38	0.41	0.30	0.40	0.34	0.39	0.33	0.19	0.08	0.23
2008	0.94	0.16	0.46	0.43	0.54	0.48	0.35	0.40	0.42	0.30	0.41	0.34	0.39	0.34	0.21	0.14	0.24
2009	0.94	0.32	0.47	0.50	0.55	0.48	0.36	0.42	0.43	0.31	0.42	0.34	0.39	0.34	0.22	0.24	0.25
2010	0.95	0.43	0.50	0.53	0.55	0.50	0.38	0.43	0.44	0.32	0.42	0.35	0.39	0.34	0.24	0.26	0.27
2011	0.96	0.68	0.53	0.63	0.55	0.51	0.42	0.43	0.45	0.33	0.42	0.35	0.39	0.34	0.26	0.26	0.29
2012	0.98	0.71	0.55	0.67	0.57	0.51	0.46	0.52	0.45	0.33	0.42	0.35	0.39	0.35	0.27	0.28	0.32
2013	0.99	0.74	0.58	0.70	0.59	0.53	0.48	0.56	0.46	0.34	0.42	0.35	0.39	0.35	0.28	0.29	0.33
2014	1.01	0.76	0.61	0.71	0.61	0.55	0.50	0.57	0.47	0.36	0.42	0.36	0.39	0.35	0.28	0.31	0.33
2015	1.03	0.81	0.63	0.71	0.64	0.56	0.51	0.58	0.47	0.38	0.42	0.36	0.39	0.36	0.29	0.32	0.34
2016	1.04	0.84	0.64	0.72	0.66	0.57	0.51	0.58	0.47	0.40	0.42	0.36	0.39	0.37	0.29	0.33	0.34
2017	1.06	0.89	0.68	0.73	0.70	0.59	0.52	0.59	0.48	0.42	0.43	0.37	0.39	0.37	0.31	0.34	0.34
2018	1.07	0.93	0.74	0.75	0.73	0.62	0.53	0.59	0.49	0.44	0.43	0.38	0.39	0.37	0.31	0.35	0.35
2019	1.07	0.95	0.79	0.76	0.77	0.63	0.53	0.59	0.50	0.46	0.43	0.39	0.40	0.38	0.32	0.36	0.35
2020	1.08	0.97	0.84	0.79	0.80	0.68	0.54	0.59	0.50	0.47	0.43	0.40	0.40	0.38	0.35	0.37	0.35
2021	1.08	0.99	0.90	0.89	0.85	0.74	0.60	0.60	0.51	0.49	0.43	0.42	0.40	0.40	0.40	0.38	0.36
2022	1.08	1.01	0.95	0.90	0.87	0.76	0.63	0.60	0.51	0.50	0.44	0.43	0.41	0.40	0.40	0.39	0.37

Year	BGD	BRA	CHL	NZL	SAU	EGY	MEX	JOR	KWT	AUT	PER	PRT	OMN	ROU	ARE	HRV	MAR
2000	0.22	0.20	0.18	0.16	0.16	0.09	0.17	0.00	0.04	0.07	0.00	0.13	0.02	0.08	0.00	0.04	0.12
2001	0.22	0.22	0.19	0.18	0.18	0.10	0.18	0.05	0.04	0.08	0.05	0.13	0.02	0.10	0.01	0.05	0.12
2002	0.23	0.23	0.19	0.19	0.19	0.10	0.18	0.05	0.04	0.08	0.05	0.13	0.02	0.10	0.02	0.05	0.12
2003	0.24	0.23	0.21	0.19	0.19	0.11	0.19	0.05	0.05	0.09	0.08	0.14	0.03	0.10	0.02	0.06	0.12
2004	0.24	0.24	0.22	0.19	0.20	0.11	0.19	0.06	0.05	0.09	0.08	0.15	0.03	0.10	0.02	0.06	0.13
2005	0.25	0.25	0.23	0.20	0.20	0.13	0.20	0.07	0.06	0.10	0.09	0.15	0.04	0.11	0.04	0.12	0.13
2006	0.25	0.27	0.25	0.20	0.21	0.14	0.20	0.09	0.06	0.16	0.17	0.16	0.05	0.12	0.08	0.12	0.13
2007	0.27	0.28	0.26	0.21	0.22	0.16	0.21	0.12	0.06	0.20	0.19	0.17	0.15	0.14	0.10	0.12	0.13
2008	0.27	0.29	0.27	0.21	0.23	0.20	0.22	0.13	0.07	0.23	0.20	0.18	0.16	0.15	0.11	0.13	0.13
2009	0.27	0.30	0.27	0.21	0.23	0.21	0.22	0.14	0.14	0.24	0.21	0.18	0.16	0.15	0.11	0.13	0.13
2010	0.28	0.30	0.27	0.21	0.23	0.22	0.23	0.15	0.15	0.25	0.22	0.18	0.17	0.15	0.11	0.13	0.13
2011	0.29	0.30	0.28	0.21	0.24	0.22	0.25	0.15	0.17	0.25	0.23	0.18	0.17	0.17	0.11	0.13	0.14
2012	0.30	0.30	0.28	0.21	0.24	0.26	0.25	0.16	0.19	0.25	0.23	0.18	0.17	0.18	0.12	0.13	0.14
2013	0.30	0.31	0.29	0.22	0.25	0.26	0.26	0.17	0.26	0.26	0.23	0.18	0.18	0.18	0.12	0.13	0.14
2014	0.31	0.32	0.31	0.23	0.26	0.27	0.26	0.17	0.26	0.26	0.23	0.19	0.18	0.19	0.12	0.13	0.14
2015	0.32	0.33	0.31	0.24	0.27	0.27	0.26	0.18	0.26	0.26	0.23	0.19	0.18	0.19	0.13	0.13	0.14
2016	0.32	0.33	0.32	0.25	0.27	0.27	0.27	0.18	0.26	0.26	0.23	0.19	0.19	0.19	0.13	0.13	0.14
2017	0.32	0.33	0.32	0.26	0.28	0.28	0.27	0.20	0.27	0.26	0.23	0.19	0.19	0.20	0.13	0.14	0.15
2018	0.33	0.33	0.32	0.27	0.29	0.28	0.28	0.20	0.27	0.26	0.23	0.19	0.19	0.20	0.14	0.15	0.15
2019	0.33	0.34	0.33	0.27	0.29	0.28	0.28	0.21	0.27	0.26	0.23	0.20	0.20	0.20	0.14	0.15	0.15
2020	0.34	0.34	0.33	0.28	0.29	0.29	0.28	0.22	0.27	0.26	0.24	0.21	0.20	0.20	0.14	0.16	0.15
2021	0.34	0.35	0.34	0.31	0.30	0.29	0.28	0.24	0.27	0.26	0.24	0.21	0.20	0.20	0.15	0.16	0.15
2022	0.35	0.35	0.34	0.32	0.30	0.30	0.28	0.28	0.27	0.26	0.24	0.21	0.20	0.20	0.17	0.16	0.15

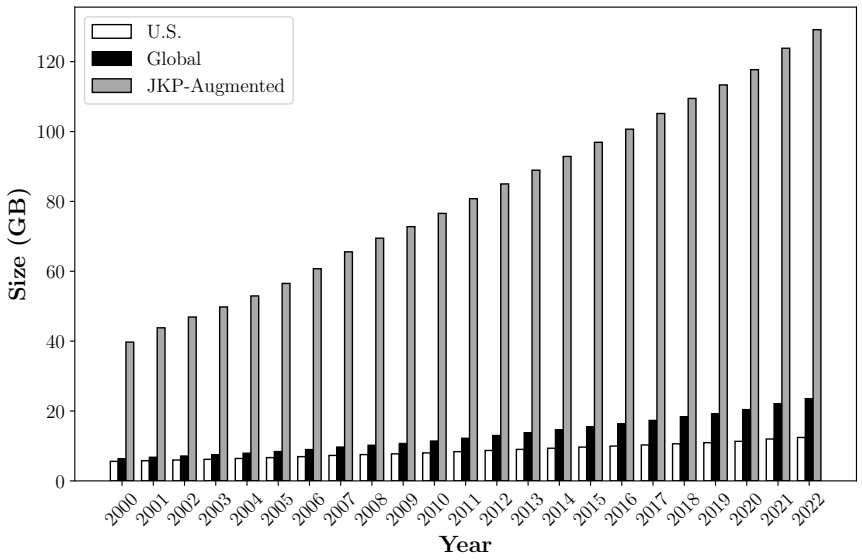
Note: This table presents the number of securities used to construct the JKP factors, as defined in Jensen et al. (2023), with values scaled in thousands. The numbers are presented cumulatively, showing the number of unique securities present up to and including each year. Although the dataset begins in 1925, coverage for many non-U.S. markets is limited in the earlier decades. Columns are ordered by their 2022 values. Blank cells denote missing data for the respective year.

Table A.9: Annual Breakdown of JKP Securities (continued)

Year	IRL	TUN	KEN	COL	MUS	HUN	QAT	LTU	SVN	CZE	LVA	CIV	EST	BHR	ISL	SRB	KAZ
2000	0.10	0.02	-	0.07	0.04	0.02	0.04	0.02	0.05	0.00	0.04	0.03	0.03	-	-	0.00	0.00
2001	0.11	0.05	-	0.07	0.04	0.03	0.04	0.04	0.05	0.03	0.04	0.03	0.03	-	0.00	0.00	0.00
2002	0.11	0.06	-	0.07	0.04	0.03	0.05	0.04	0.06	0.03	0.04	0.03	0.03	-	0.01	0.03	0.00
2003	0.11	0.06	-	0.07	0.04	0.03	0.05	0.04	0.06	0.04	0.04	0.04	0.03	-	0.02	0.03	0.00
2004	0.12	0.07	-	0.07	0.05	0.04	0.05	0.04	0.06	0.04	0.04	0.04	0.03	-	0.02	0.03	0.00
2005	0.12	0.08	-	0.07	0.05	0.04	0.06	0.04	0.06	0.05	0.05	0.04	0.03	-	0.03	0.03	0.01
2006	0.13	0.09	-	0.08	0.07	0.04	0.06	0.05	0.07	0.06	0.05	0.04	0.04	-	0.04	0.04	0.01
2007	0.13	0.09	-	0.08	0.08	0.04	0.07	0.05	0.07	0.06	0.07	0.04	0.04	-	0.04	0.04	0.01
2008	0.13	0.10	-	0.08	0.09	0.04	0.07	0.06	0.07	0.06	0.07	0.05	0.04	-	0.04	0.04	0.01
2009	0.13	0.11	0.03	0.09	0.09	0.06	0.08	0.06	0.07	0.06	0.07	0.05	0.04	0.02	0.05	0.04	0.02
2010	0.13	0.12	0.05	0.09	0.09	0.06	0.08	0.06	0.07	0.06	0.07	0.05	0.04	0.02	0.05	0.04	0.03
2011	0.13	0.12	0.06	0.09	0.09	0.06	0.09	0.06	0.08	0.06	0.07	0.05	0.04	0.03	0.05	0.04	0.03
2012	0.13	0.12	0.11	0.10	0.09	0.06	0.09	0.07	0.08	0.06	0.07	0.06	0.04	0.03	0.05	0.04	0.04
2013	0.14	0.12	0.11	0.10	0.09	0.06	0.09	0.08	0.08	0.06	0.07	0.06	0.04	0.03	0.05	0.04	0.04
2014	0.14	0.13	0.12	0.10	0.09	0.07	0.09	0.09	0.08	0.07	0.07	0.07	0.04	0.04	0.05	0.04	0.04
2015	0.14	0.13	0.12	0.10	0.10	0.08	0.09	0.09	0.08	0.07	0.07	0.07	0.04	0.04	0.05	0.04	0.04
2016	0.14	0.13	0.12	0.10	0.10	0.08	0.10	0.09	0.08	0.07	0.07	0.07	0.04	0.05	0.05	0.04	0.04
2017	0.14	0.13	0.12	0.11	0.10	0.09	0.10	0.09	0.08	0.07	0.07	0.07	0.04	0.05	0.05	0.05	0.04
2018	0.15	0.14	0.13	0.11	0.10	0.09	0.10	0.09	0.08	0.07	0.07	0.07	0.05	0.05	0.05	0.05	0.04
2019	0.15	0.14	0.13	0.11	0.10	0.09	0.10	0.09	0.09	0.08	0.07	0.07	0.05	0.05	0.05	0.05	0.05
2020	0.15	0.14	0.13	0.11	0.10	0.10	0.10	0.09	0.09	0.08	0.07	0.07	0.05	0.05	0.05	0.05	0.05
2021	0.15	0.14	0.13	0.12	0.10	0.10	0.10	0.09	0.09	0.08	0.07	0.07	0.05	0.05	0.05	0.05	0.05
2022	0.15	0.14	0.14	0.12	0.10	0.10	0.10	0.10	0.09	0.08	0.07	0.07	0.06	0.06	0.05	0.05	0.05

Note: This table presents the number of securities used to construct the JKP factors, as defined in Jensen et al. (2023), with values scaled in thousands. The numbers are presented cumulatively, showing the number of unique securities present up to and including each year. Although the dataset begins in 1925, coverage for many non-U.S. markets is limited in the earlier decades. Columns are ordered by their 2022 values. Blank cells denote missing data for the respective year.

Figure A.4: Comparative Data Volume across U.S., Global, and JKP-Augmented Datasets



Note: The figure reports the annual dataset sizes (in gigabytes) for the U.S., Global, and JKP-augmented samples over the period 2000–2022. The U.S. sample includes only the domestic market, the Global sample covers all international markets, and the JKP-augmented sample extends the global data by incorporating the JKP factors defined in Jensen et al. (2023).

B Appendix: Extended Results

Table B.1: Comparison of Daily Out-of-Sample Forecasting Performance of Benchmark Models Using Diebold-Mariano Test

		OLS+H			LASSO+H			RIDGE+H			Enet+H			PCR			XGBoost			CatBoost			LightGBM			NN-S			NN-L		
		21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
OLS+H	5	-0.30	-4.77*	-7.76*	-16.98*	-9.65*	-10.71*	-11.86*	-10.3*	-10.73*	-12.52*	-13.04*	-10.76*	-10.69*	-9.81*	-7.41*	-6.67*	-5.90*	-7.51*	-15.03*	0.77*	4.02*	3.32*	3.17*	0.26	3.22*	3.11*	3.66*	0.66	4.05*	3.33*	2.75*	-14.96*	-17.51*	-10.16*	-16.00*	-14.62*	-11.20*	-13.38*	-15.15*
	21		-5.96*	-8.71*	-17.01*	-9.62*	-10.70*	-11.84*	-10.33*	-10.82*	-12.63*	-13.11*	-10.76*	-10.71*	-9.82*	-7.38*	-6.62*	-5.28*	-7.48*	-14.97*	0.81	4.16*	3.43*	3.29*	0.28	3.34*	3.21*	3.79*	0.69	4.18*	3.44*	2.85*	-14.94*	-17.52*	-10.16*	-16.06*	-14.68*	-11.29*	-13.47*	-15.24*
	252			-8.67*	-17.07*	-9.09*	-10.28*	-11.47*	-9.93*	-10.42*	-12.59*	-13.11*	-10.33*	-10.28*	-9.33*	-6.70*	-6.03*	-5.28*	-6.88*	-14.45*	1.63	5.03*	4.52*	4.48*	1.02	4.18*	4.03*	4.76*	1.48	5.13*	4.52*	3.98*	-14.42*	-17.28*	-9.64*	-15.89*	-14.53*	-11.09*	-13.09*	-15.04*
	512				-16.93*	-8.33*	-9.62*	-10.85*	-9.11*	-9.64*	-12.07*	-12.74*	-9.53*	-9.50*	-8.48*	-5.70*	-5.32*	-4.57*	-6.17*	-13.85*	2.65*	5.99*	5.69*	5.77*	1.97*	5.13*	4.90*	5.79*	2.47*	6.19*	5.69*	5.25*	-13.76*	-16.84*	-9.02*	-15.37*	-14.21*	-10.70*	-12.37*	-14.46*
LASSO+H	5					17.61*	16.35*	15.37*	22.96*	22.12*	19.82*	18.36*	22.22*	21.64*	20.30*	18.71*	8.58*	8.76*	7.80*	1.41	16.71*	15.97*	16.41*	16.68*	16.05*	15.48*	14.43*	15.76*	16.52*	16.71*	16.62*	16.94*	0.56	-0.53	6.52*	4.71*	-0.98	4.97*	5.01*	1.52
	21						-12.02*	-17.51*	1.91	1.14	-3.61*	-4.91*	1.18	1.13	2.82*	12.05*	0.54	1.38	-0.80	-13.60*	8.35*	9.17*	9.17*	9.29*	7.54*	8.53*	8.00*	8.91*	8.05*	9.43*	9.18*	9.13*	-7.47*	-10.27*	-2.85*	-7.44*	-9.61*	-4.99*	-6.15*	-8.32*
	252							-7.82*	6.18*	5.03*	-0.08	-1.60	5.98*	5.98*	9.43*	15.47*	2.40*	3.07*	1.16	-10.76*	9.77*	10.25*	10.34*	10.47*	8.94*	8.94*	9.05*	10.03*	9.45*	10.56*	10.37*	10.37*	-6.14*	-8.68*	-1.26	-5.42*	-8.54*	-3.53*	-4.41*	-7.01*
	512								9.37*	7.68*	2.34*	0.54	9.55*	9.06*	11.71*	19.27*	3.59*	4.19*	2.36*	-9.71*	10.59*	10.98*	11.13*	11.30*	9.74*	10.35*	9.70*	10.74*	10.26*	11.35*	11.17*	11.20*	-5.30*	-7.78*	-0.32	-4.37*	-7.52*	-2.53*	-3.33*	-5.90*
RIDGE+H	5									-1.70	-9.41*	-10.33*	-2.39*	-2.11*	0.21	5.18*	-0.25	0.44	-1.29	-11.14*	9.28*	10.12*	10.34*	10.53*	8.32*	9.47*	8.72*	9.91*	8.96*	10.68*	10.50*	10.55*	-8.36*	-11.70*	-3.64*	-9.37*	-11.14*	-6.09*	-8.02*	-10.21*
	21										-8.64*	-9.61*	-0.46	-0.59	1.14	5.12*	-0.02	0.64	-1.04	-10.59*	9.63*	10.51*	10.69*	10.92*	8.67*	9.87*	9.00*	10.24*	9.32*	11.13*	10.86*	10.95*	-8.21*	-11.36*	-3.42*	-8.83*	-10.98*	-5.99*	-7.74*	-10.03*
	252											-3.73*	7.46*	7.06*	7.16*	8.83*	1.87	2.43*	0.91	-8.21*	11.34*	11.93*	12.26*	12.57*	10.42*	11.32*	10.45*	11.75*	11.07*	12.60*	12.44*	12.65*	-6.35*	-9.68*	-1.36	-6.11*	-8.71*	-3.91*	-4.64*	-7.19*
	512												8.68*	8.12*	7.71*	9.60*	2.50*	3.02*	1.55	-7.35*	11.62*	12.25*	12.56*	12.85*	10.76*	11.60*	10.75*	12.02*	11.38*	12.91*	12.74*	12.93*	-5.51*	-8.77*	-0.59	-5.07*	-7.79*	-2.98*	-3.69*	-6.14*
Enet+H	5													-0.29	2.18*	6.51*	0.06	0.78	-1.03	-11.78*	9.24*	10.06*	10.26*	10.45*	8.30*	9.43*	8.69*	9.81*	8.93*	10.62*	10.41*	10.44*	-8.05*	-11.46*	-3.40*	-8.76*	-10.52*	-5.82*	-7.30*	-9.57*
	21														2.72*	6.50*	0.10	0.80	-1.00	-11.63*	9.28*	10.07*	10.28*	10.49*	8.34*	9.46*	8.73*	9.85*	8.97*	10.62*	10.44*	10.48*	-8.04*	-11.35*	-3.36*	-8.74*	-10.62*	-5.89*	-7.14*	-9.56*
	252															6.69*	-0.33	0.45	-1.51	-12.36*	8.82*	9.66*	9.81*	9.96*	7.89*	9.04*	8.40*	9.47*	8.49*	10.10*	9.93*	9.91*	-8.38*	-11.49*	-3.62*	-8.92*	-11.01*	-6.13*	-7.61*	-10.14*
	512																-2.63*	-1.62	-4.08*	-16.17*	6.40*	7.70*	7.60*	7.67*	5.60*	7.05*	6.63*	7.48*	6.11*	7.90*	7.61*	7.47*	-9.57*	-12.77*	-5.32*	-10.65*	-11.65*	-7.31*	-9.09*	-11.02*
PCR	5																	1.71	-2.03*	-12.34*	5.91*	7.06*	6.79*	6.77*	5.36*	6.59*	6.42*	6.85*	5.70*	7.04*	6.75*	6.52*	-6.78*	-8.17*	-2.60*	-5.45*	-7.62*	-3.89*	-4.75*	-6.57*
	21																		-3.34*	-12.13*	5.27*	6.56*	6.24*	6.18*	4.77*	6.07*	5.95*	6.34*	5.09*	6.48*	6.17*	5.93*	-7.30*	-8.52*	-3.12*	-5.85*	-7.92*	-4.34*	-5.22*	-6.96*
	252																			-11.23*	6.73*	7.91*	7.66*	7.64*	6.15*	7.41*	7.25*	7.72*	6.51*	7.88*	7.60*	7.38*	-6.14*	-7.43*	-1.75	-4.58*	-7.04*	-3.19*	-4.00*	-5.88*
	512																				12.66*	13.18*	13.30*	13.44*	11.92*	12.61*	12.06*	12.89*	12.38*	13.47*	13.29*	13.23*	-0.38	-1.50	5.12*	2.65*	-1.63	2.70*	3.03*	0.19
XGBoost	5																					4.79*	3.28*	2.89*	-3.15*	3.69*	3.20*	4.27*	-0.88	4.87*	3.51*	2.64*	-15.02*	-18.24*	-9.47*	-16.33*	-17.35*	-13.22*	-15.01*	-18.32*
	21																						-3.68*	-3.80*	-5.36*	-2.43*	-1.10	-1.05	-4.88*	-1.12	-3.08*	-4.26*	-16.62*	-18.67*	-10.72*	-16.66*	-16.94*	-13.44*	-15.17*	-17.82*
	252																							-1.67	-3.84*	1.55	1.60	2.40*	-3.37*	3.36*	0.56	-2.42*	-16.33*	-18.93*	-10.48*	-17.13*	-16.79*	-13.32*	-15.06*	-18.21*
	512																								-3.40*	56.19*	2.04*	3.51*	-2.93*	3.64*	1.82	-1.10	-16.16*	-19.11*	-10.68*	-17.22*	-16.64*	-13.29*	-15.01*	-17.87*
CatBoost	5																									4.38*	3.76*	4.80*	2.99*	5.46*	4.08*	3.28*	-14.38*	-17.53*	-8.88*	-15.47*	-17.02*	-12.85*	-14.13*	-17.68*
	21																										0.22	0.45	-3.81*	1.24	-1.38	-2.64*	-15.93*	-17.93*	-10.20*	-15.89*	-16.61*	-13.14*	-14.59*	-17.49*
	252																											0.27	-3.30*	0.49	-1.36	-2.37*	-15.60*	-17.20*	-9.68*	-14.87*	-15.95*	-12.40*	-13.56*	-16.90*
	512																												-4.29*	0.41	-2.10*	-3.76*	-16.07*	-18.27*	-10.35*	-16.28*	-16.85*	-13.32*	-14.64*	-17.81*
LightGBM	5																													5.91*	4.21*	3.13*	-14.89*	-18.13*	-9.38*	-16.04*	-17.20*	-13.16*	-14.71*	-18.04*
	21																														-3.51*	-5.48*	-16.59*	-19.12*	-10.98*	-17.31*	-17.16*	-13.79*	-15.58*	-18.21*
	252																															-4.86*	-16.20*	-19.01*	-10.51*	-17.34*	-16.99*	-13.47*	-15.27*	-18.50*
	512																																-15.95*	-19.11*	-10.50*	-17.35*	-16.90*	-13.51*	-15.07*	-18.17*
NN-S	5																																	-0.97	5.42*	2.57*	-1.30	3.28*	3.13*	0.65
	21																																		7.33*	5.02*	-0.39	4.70*	4.79*	1.74
	252																																			-2.85*	-5.08*	-1.99*	-2.14*	-4.00*
	512																																				-4.71*	0.80	0.70	-2.62*
NN-L	5																																					5.14*	5.08*	2.69*
	21																																						-0.21	-2.93*
	252																																							-3.63*

Note: This table reports the modified Diebold–Mariano (DM) test statistics, following the approach of Gu et al. (2020), for out-of-sample stock-level forecast comparisons across benchmark models. A positive statistic indicates that the model in the column achieves superior predictive performance relative to the model in the corresponding row. An asterisk denotes statistical significance at the 5% level. The benchmark models encompass linear approaches (OLS, Lasso, Ridge, Elastic Net, and PCR), ensemble methods (XGBoost, CatBoost, and LightGBM), and neural networks (NN-S and NN-L). Results are reported for four rolling window sizes (5, 21, 252, and 512). 'H' indicates that the model is estimated using the Huber loss.

Table B.2: Benchmark Models - Spread Portfolio Performance

Model	**OLS+H**				**LASSO+H**				**RIDGE+H**				**Enet+H**				**PCR**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
Low	-32.35	-32.62	-33.13	-32.96	-32.02	-31.40	-31.47	-31.44	-32.14	-31.36	-32.61	-32.12	-31.52	-31.64	-31.64	-31.44	-21.66	-20.54	-20.55	-23.50
2	-8.97	-8.83	-8.82	-8.63	-8.15	-4.67	-4.94	-4.96	-7.83	-8.51	-7.95	-7.85	-7.23	-6.91	-6.37	-5.07	1.30	0.94	1.54	-1.06
3	-0.62	-0.84	-0.98	-1.19	0.02	2.29	2.21	2.17	0.25	1.16	0.48	0.78	0.87	1.00	1.53	2.13	6.83	6.59	6.00	5.07
4	4.45	3.94	4.49	4.80	4.40	6.83	6.45	6.34	4.56	4.11	4.83	4.63	5.16	5.13	5.82	6.56	7.75	8.79	9.04	7.90
5	7.86	7.99	8.09	8.03	8.46	9.02	8.96	8.93	8.05	8.45	8.52	9.09	8.79	8.64	8.49	8.89	10.84	10.58	11.19	9.94
6	12.51	12.04	11.99	12.20	12.14	11.80	12.04	11.79	12.53	12.10	12.05	11.78	12.41	12.31	12.05	11.95	12.79	12.29	12.54	13.07
7	16.41	16.92	16.60	16.69	16.35	14.70	14.62	14.93	16.03	15.85	16.28	16.39	15.23	15.21	14.97	14.77	14.31	14.23	15.37	15.39
8	22.25	22.39	22.33	22.82	21.50	19.24	19.48	19.44	21.71	21.65	21.96	22.81	20.66	20.68	20.49	19.41	17.97	17.77	17.77	17.82
9	31.43	31.50	32.04	31.81	30.95	27.54	27.74	27.96	30.40	30.22	30.74	29.29	29.83	29.90	28.99	27.88	22.49	22.18	21.78	23.39
High	58.29	58.80	58.68	57.72	57.64	55.94	56.20	56.12	57.72	57.61	56.97	56.48	57.07	56.97	56.96	56.22	38.67	38.46	36.62	43.27
H-L	45.32	45.71	45.91	45.34	44.83	43.67	43.84	43.78	44.93	44.48	44.79	44.30	44.29	44.31	44.30	43.83	30.17	29.50	28.59	33.39
Model	**XGBoost**				**CatBoost**				**LightGBM**				**NN-S**				**NN-L**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
Low	-34.62	-36.00	-35.12	-34.68	-31.91	-36.52	-35.00	-36.06	-33.75	-34.82	-33.92	-33.66	-26.53	-30.34	-29.40	-29.15	-31.34	-30.62	-29.45	-27.90
2	-3.48	-4.60	-6.16	-5.38	-0.94	-3.02	-3.94	-4.06	-3.15	-4.82	-6.46	-7.05	-7.09	-7.52	-5.35	-5.20	-8.64	-8.33	-5.89	-4.93
3	4.41	4.15	2.14	2.09	4.28	4.49	2.72	3.30	3.05	1.89	1.89	1.10	0.38	0.60	1.39	1.88	-8.64	-8.33	-5.89	-4.93
4	6.18	8.21	6.77	6.35	7.46	8.32	6.38	6.84	6.01	6.45	5.97	5.85	4.12	5.11	6.07	6.32	4.26	4.04	5.11	6.73
5	10.41	9.69	9.80	8.74	9.38	10.24	9.84	9.36	9.95	9.79	9.24	9.33	7.92	9.16	10.30	9.95	8.59	8.58	8.78	8.90
6	12.93	12.00	11.58	11.29	11.51	11.43	11.88	12.41	12.30	12.32	11.29	12.02	11.44	12.51	12.66	12.48	12.24	12.62	13.11	12.90
7	15.07	14.89	14.48	14.64	15.23	15.19	15.65	14.06	15.50	14.81	15.21	15.05	15.27	15.81	16.06	15.13	16.31	16.60	16.48	15.64
8	19.18	18.43	19.93	19.65	18.19	18.65	18.23	19.32	20.69	20.31	20.73	20.26	20.97	21.53	20.45	19.99	22.54	21.79	20.82	20.12
9	26.90	26.80	27.60	28.45	24.92	25.96	27.53	27.67	27.20	28.31	28.29	29.01	30.14	29.54	28.20	28.85	30.60	31.69	28.64	27.41
High	54.32	57.71	60.26	60.12	53.17	56.53	58.00	58.44	53.48	57.03	59.06	59.37	54.65	54.90	50.90	51.06	56.74	55.31	52.23	49.75
H-L	44.47	46.86	47.69	47.40	42.54	46.52	46.50	47.25	43.61	45.93	46.49	46.52	40.59	42.62	40.15	40.11	44.04	42.97	40.84	38.82

Note: This table presents the annualized average returns of decile spread portfolios formed using forecasts from various predictive models across different window sizes (5, 21, 252, and 512 trading days). For each model and window size, the returns of decile portfolios from the lowest (Low) to the highest (High) forecasted return are reported. The final row (H-L) represents the return spread between the highest and lowest deciles. Models evaluated include linear (OLS+H, LASSO+H, RIDGE+H, Elastic Net+H, and PCR), ensemble (XGBoost, CatBoost, and LightGBM), and neural network (NN-S and NN-L) models. 'H' indicates that the model is estimated using the Huber loss. Portfolios are constructed using equal-weighted decile sorting based on predicted returns.

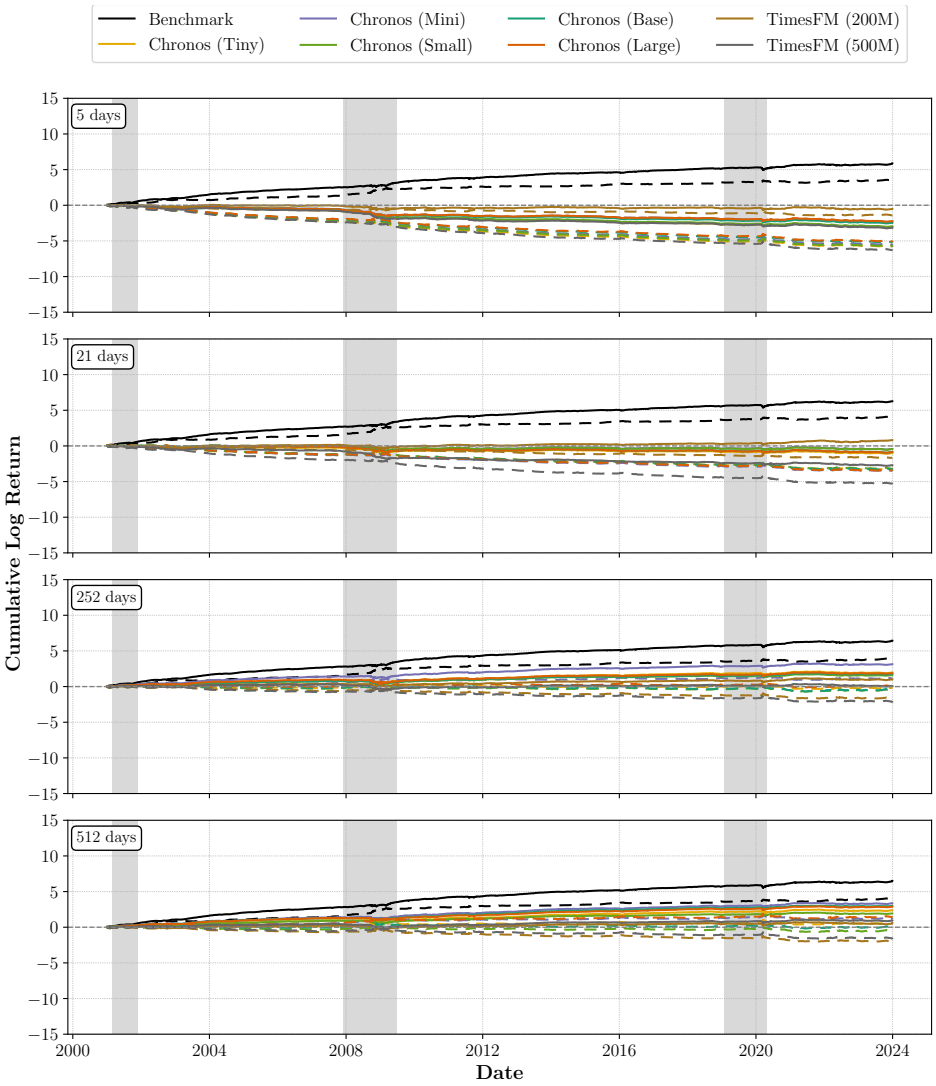
Table B.3: Zero-Shot TSFMs - Spread Portfolio Performance

Model	**Benchmark**				**Chronos (Tiny)**				**Chronos (Mini)**				**Chronos (Small)**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
Low	-31.91	-36.52	-35.00	-36.06	49.11	27.27	0.54	-6.10	47.24	28.75	-11.45	-11.38	49.05	26.20	2.19	2.69
2	-0.94	-3.02	-3.94	-4.06	25.05	17.80	5.40	4.73	24.98	19.14	4.34	3.80	25.48	19.25	8.55	7.73
3	4.28	4.49	2.72	3.30	18.50	14.73	8.36	9.44	17.84	15.43	8.18	8.65	19.28	15.83	10.24	8.67
4	7.46	8.32	6.38	6.84	14.27	13.41	11.19	10.17	13.14	13.44	10.22	9.47	14.75	14.38	11.22	10.71
5	9.38	10.24	9.84	9.36	13.13	13.08	12.20	12.07	11.26	11.95	11.25	10.44	12.34	12.16	11.70	10.76
6	11.51	11.43	11.88	12.41	9.56	12.18	11.84	13.62	9.13	8.98	13.17	11.94	9.03	10.01	13.46	12.42
7	15.23	15.19	15.65	14.06	7.78	9.34	14.21	13.78	6.37	7.39	14.17	13.97	7.32	8.06	12.76	12.91
8	18.19	18.65	18.23	19.32	4.42	8.14	14.36	15.56	3.54	6.16	15.74	15.69	3.29	5.25	13.96	14.63
9	24.92	25.96	27.53	27.67	-3.73	4.49	16.13	17.41	-2.00	4.09	18.60	19.73	-3.08	4.46	13.74	14.60
High	53.17	56.53	58.00	58.44	-25.03	-7.36	18.84	22.41	-18.45	-2.26	28.89	30.80	-24.40	-2.53	15.26	17.97
H-L	42.54	46.52	46.50	47.25	-37.07	-17.31	9.15	14.25	-32.85	-15.50	20.17	21.09	-36.73	-14.36	6.53	7.64

Model	**Chronos (Base)**				**Chronos (Large)**				**TimesFM 1 (200M)**				**TimesFM 2 (500M)**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
Low	43.81	27.40	1.68	-1.70	43.28	28.69	-2.15	-14.07	10.83	11.23	12.81	15.83	53.16	44.44	17.27	12.38
2	23.50	21.61	10.90	7.02	22.98	19.83	7.65	4.86	15.88	10.68	12.12	14.54	31.51	29.18	19.41	15.22
3	18.66	16.62	10.84	8.70	17.16	16.90	10.79	9.43	15.39	10.96	11.90	14.32	22.96	22.12	16.66	15.10
4	15.58	14.36	11.88	10.14	14.23	14.21	11.85	12.10	14.94	9.92	11.45	12.71	16.82	17.38	15.04	14.50
5	12.65	11.92	11.86	10.38	12.13	12.11	12.38	12.42	14.42	9.95	11.41	12.19	11.94	13.17	13.36	11.67
6	9.54	11.16	12.06	10.28	9.55	10.63	13.08	13.68	14.05	11.36	11.19	10.11	7.91	9.33	10.49	10.99
7	7.90	8.13	11.66	11.62	8.00	8.66	13.28	14.13	12.43	10.77	11.05	10.40	2.98	5.03	9.02	9.37
8	3.66	5.39	12.37	13.26	4.88	5.99	13.17	15.68	11.10	11.94	11.00	9.37	-0.71	0.77	6.10	7.70
9	-2.14	2.46	13.43	16.39	-1.24	2.15	15.42	18.63	6.91	12.95	10.14	7.47	-7.43	-6.40	3.68	6.70
High	-20.08	-5.98	16.42	27.00	-17.91	-6.11	17.63	26.26	-2.83	13.32	10.01	6.12	-26.09	-22.22	2.05	9.45
H-L	-31.94	-16.69	7.37	14.35	-30.59	-17.40	9.89	20.17	-6.83	-1.83	-1.40	-4.85	-39.62	-33.33	-7.61	-1.47

Note: This table presents the annualized average returns of decile spread portfolios formed using forecasts from various predictive models across different window sizes (5, 21, 252, and 512 trading days). The benchmark model is CatBoost, the best-performing model among the benchmarks. The time series foundation models (TSFMs) include Chronos (tiny, mini, small, base, and large), TimesFM (version 1 with 200 million and version 2 with 500 million parameters), and Uni2TS (Moirai-small and Moirai-base). Zero-shot inference is performed using the pre-trained models released by the respective authors. For each model and window size, the returns of decile portfolios from the lowest (Low) to the highest (High) forecasted return are reported. The final row (H-L) represents the return spread between the highest and lowest deciles. Portfolios are constructed using equal-weighted decile sorting based on predicted returns.

Figure B.1: Cumulative Log Returns of Zero-Shot TSFMs: Long and Short Portfolios



Note: This figure displays the cumulative log returns of long and short portfolios, separately constructed using various forecasting models over rolling windows of 5, 21, 252, and 512 trading days. The benchmark model is CatBoost, the best-performing model among the benchmarks. The time series foundation models (TSFMs) include Chronos (tiny, mini, small, base, and large) and TimesFM (version 1 with 200 million and version 2 with 500 million parameters). Zero-shot inference is performed using the pre-trained models released by the respective authors. Each subplot corresponds to a specific horizon, as indicated by the text labels in the upper-left corners. For each model, the solid line denotes the performance of the long portfolio, while the dashed line represents the corresponding short portfolio. The benchmark model (CatBoost) is highlighted in black with bold lines. Shaded areas indicate U.S. recession periods, as defined by the National Bureau of Economic Research (NBER). All portfolios are equally weighted.

Table B.4: Zero-Shot TSFMs - Forecasting Performance (Extended)

Model	Chronos-Bolt (Tiny)				Chronos-Bolt (Mini)				Chronos-Bolt (Small)				Chronos-Bolt (Base)				TimesFM 2.5				Moirai (Small)			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
R^2_{OOS}	-118.33	-21.90	-2.51	-1.42	-119.61	-21.79	-1.82	-0.99	-95.72	-23.76	-1.69	-0.75	-106.92	-20.23	-1.15	-0.65	-44.98	-15.75	-2.78	-1.75	-8.3e10	-1.7e10	-4.0e4	-18.40
	-108.72	-21.08	-2.56	-1.52	-110.67	-20.88	-1.68	-0.99	-88.80	-23.06	-1.61	-0.83	-92.78	-19.18	-1.07	-0.62	-44.33	-15.14	-2.96	-2.07	-1.5e10	-873.50	-12.18	-9.63
	-130.31	-22.74	-2.53	-1.40	-130.68	-22.39	-1.99	-1.05	-104.29	-24.47	-1.77	-0.68	-122.91	-20.99	-1.31	-0.72	-47.87	-17.17	-2.32	-1.13	-1.9e11	-2.6e9	-9.4e4	-30.94
Overall Acc.	48.46	48.91	49.46	49.68	48.62	48.65	49.77	50.00	48.25	48.51	49.97	50.26	48.34	48.71	50.25	50.38	48.42	48.68	50.07	50.34	48.23	48.45	48.63	48.65
	49.05	49.08	49.74	49.87	48.97	49.24	50.16	50.36	49.15	49.15	50.06	50.45	49.12	49.16	50.37	50.46	49.18	49.37	50.02	50.11	49.43	49.56	49.61	49.62
	46.95	48.26	49.17	49.60	47.48	47.26	49.38	49.73	46.21	46.91	50.12	50.36	46.51	47.50	50.54	50.75	46.72	47.16	50.80	51.49	45.80	46.33	46.74	46.78
Up Acc.	41.14	37.17	48.43	47.19	36.45	44.91	55.96	58.22	45.39	42.68	49.44	53.70	44.39	43.03	52.28	48.80	50.02	51.43	56.72	56.82	57.97	59.91	59.27	59.02
	43.46	40.82	55.86	56.19	38.91	48.57	63.27	67.22	47.71	45.70	57.20	62.85	46.81	45.87	61.94	61.25	52.19	54.75	63.16	64.28	59.97	61.69	61.13	61.04
	37.14	31.70	39.34	36.43	32.17	39.78	48.36	48.47	41.46	38.42	41.96	45.26	40.30	39.33	41.62	36.23	46.52	46.82	50.50	50.07	54.75	57.52	57.23	57.14
Down Acc.	55.52	60.20	50.12	51.73	60.41	52.12	43.41	41.71	50.94	54.04	50.10	46.49	52.09	54.07	47.80	51.51	46.76	45.81	43.24	43.72	38.61	37.14	38.09	38.38
	54.74	57.41	42.76	42.66	59.31	49.64	35.99	32.35	50.49	52.49	41.92	36.78	51.37	52.28	37.54	38.49	45.87	43.47	35.78	34.86	38.27	36.68	37.34	37.44
	55.59	62.87	57.71	61.10	61.00	53.80	50.11	50.70	50.34	54.36	57.24	54.76	51.95	54.68	58.25	63.50	46.81	47.35	50.90	52.62	37.79	36.32	37.35	37.54
F1	0.48	0.48	0.49	0.49	0.48	0.48	0.49	0.49	0.48	0.48	0.49	0.49	0.48	0.48	0.49	0.49	0.48	0.49	0.49	0.50	0.48	0.48	0.48	0.48
	0.49	0.49	0.49	0.49	0.48	0.49	0.48	0.48	0.49	0.49	0.49	0.48	0.49	0.49	0.48	0.48	0.49	0.49	0.48	0.48	0.49	0.48	0.49	0.49
	0.46	0.46	0.48	0.48	0.46	0.47	0.49	0.49	0.46	0.46	0.49	0.49	0.46	0.47	0.49	0.49	0.47	0.47	0.50	0.51	0.46	0.46	0.46	0.46

Model	Moirai (Base)				Moirai (Large)				Moirai 2				Kairos (10M)				Kairos (23M)				Kairos (50M)			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
R^2_{OOS}	-1.9e8	-367.36	-1133.48	-1.8e7	-2.0e7	-8.4e7	-1.0e5	-803.37	-39.65	-12.08	-2.28	-1.91	-302.09	-62.29	-2.51	-1.60	-181.76	-44.37	-2.73	-4.62	-185.76	-33.37	-1.96	-1.13
	-3181.88	-277.52	-68.40	-60.02	-3.5e7	-2.7e6	-3630.24	-1015.80	-38.84	-12.48	-2.48	-2.02	-282.02	-57.04	-2.52	-1.39	-168.83	-38.71	-3.01	-5.07	-166.09	-29.01	-1.83	-1.06
	-5.1e8	-470.75	-1074.89	-5.1e7	-3.2e7	-5.8e6	-2.3e4	-2068.74	-41.57	-12.00	-1.98	-1.61	-330.02	-68.44	-2.64	-1.98	-202.33	-50.23	-2.58	-4.28	-212.01	-37.57	-2.06	-1.19
Overall Acc.	48.30	48.50	48.95	49.10	48.61	48.75	49.04	49.18	48.40	49.11	50.18	50.39	49.10	49.04	49.89	50.12	48.66	48.58	50.27	50.15	48.85	48.81	50.15	50.21
	49.65	49.67	49.64	49.60	49.92	50.01	49.76	49.89	49.05	49.39	50.37	50.44	50.62	50.42	50.01	50.18	49.52	49.67	50.19	50.08	49.70	49.57	49.90	50.27
	45.74	46.24	47.65	48.10	46.22	46.50	47.76	48.02	46.78	48.50	50.61	51.14	46.80	46.80	50.04	50.52	46.97	46.48	50.80	50.52	47.19	47.37	50.82	50.65
Up Acc.	64.53	63.41	57.18	53.45	75.31	70.73	60.30	60.86	41.88	50.70	69.16	69.22	91.87	84.86	51.00	48.49	58.15	62.94	54.54	48.54	61.65	58.31	41.00	47.45
	66.29	65.15	59.58	56.07	76.40	72.69	63.16	64.50	44.55	53.67	74.63	75.56	92.07	85.03	59.20	59.64	59.70	64.28	62.74	55.36	62.78	60.56	49.06	58.00
	61.73	60.71	54.25	50.90	74.12	68.48	56.78	56.08	37.43	46.99	63.85	62.69	91.70	85.02	40.62	33.60	55.58	61.29	43.44	38.37	59.94	55.23	32.84	36.52
Down Acc.	32.39	33.85	40.81	44.73	22.46	27.18	37.86	37.55	54.65	47.37	31.36	31.76	7.34	14.02	48.46	51.34	39.34	34.47	45.93	51.48	36.27	39.39	58.74	52.48
	32.16	33.34	39.11	42.66	22.19	26.17	35.51	34.32	53.55	44.62	24.63	23.86	7.39	14.24	39.87	39.74	38.82	34.26	36.88	44.20	35.95	37.87	50.17	41.47
	31.48	33.31	41.71	45.54	21.37	26.89	39.62	40.72	55.01	49.73	38.66	40.70	6.85	12.78	58.22	65.34	39.24	33.25	57.19	61.17	35.80	40.31	66.65	63.01
F1	0.47	0.47	0.49	0.49	0.45	0.46	0.48	0.48	0.48	0.49	0.48	0.48	0.38	0.42	0.50	0.50	0.48	0.48	0.50	0.50	0.48	0.48	0.49	0.50
	0.48	0.48	0.49	0.49	0.46	0.47	0.48	0.48	0.49	0.49	0.46	0.46	0.39	0.43	0.49	0.49	0.49	0.48	0.49	0.50	0.49	0.49	0.49	0.49
	0.45	0.46	0.48	0.48	0.43	0.45	0.47	0.48	0.46	0.48	0.50	0.51	0.37	0.40	0.49	0.48	0.47	0.46	0.50	0.49	0.47	0.47	0.49	0.49

Note: This table presents each metric as a set of three values, ordered from top to bottom: full sample, top 25% of firms by market capitalization (large-cap), and bottom 25% (small-cap) for various predictive models across different window sizes (5, 21, 252, and 512 trading days). The time series foundation models (TSFMs) include Chronos-Bolt (tiny, mini, small, and base), TimesFM 2.5, Uni2TS (Moirai 1.1 small, base, and large, as well as Moirai 2), and Kairos (10M, 23M and 50M parameter models). Zero-shot inference is performed using the pre-trained models released by the respective authors. If the model yields a distribution as its output, we use the mean value. Otherwise, we adopt the point estimation approach specified by the respective authors. The hyperparameters are also set according to those proposed by the authors. Metrics are first computed separately for each calendar year using all stock-date observations within that year. The reported values represent the average of these yearly statistics. Metrics include out-of-sample R^2 (R^2_{OOS}), overall directional accuracy, upward and downward classification accuracy, and macro-averaged F1 score. 'Overall Acc.' denotes overall directional accuracy, 'Up Acc.' and 'Down Acc.' represent the model's accuracy in predicting upward and downward excess returns respectively, and 'F1' refers to the macro-averaged F1 score.

Table B.5: Zero-Shot TSFMs - Forecasting Performance (Extended) - Continued

Model	Moment (Small)				Moment (Base)				Moment (Large)				Lag-Llama				TiRex			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
R^2_{OOS}	-9.26	-18.23	-3.36	-5.14	-14.61	-21.55	-24.16	-23.04	-11.75	-19.14	-8.07	-5.06	-52.21	-39.68	-8.20	-9.48	-40.38	-25.88	-2.57	-1.25
	-8.82	-18.22	-3.26	-5.11	-15.78	-23.18	-26.54	-26.10	-10.66	-19.09	-6.88	-5.02	-49.37	-38.53	-9.18	-11.17	-39.48	-25.43	-2.95	-1.76
	-10.20	-18.97	-3.33	-5.21	-13.90	-19.93	-22.43	-21.04	-13.45	-19.51	-9.11	-5.17	-57.32	-42.50	-7.44	-8.42	-42.42	-26.77	-2.02	-0.50
Overall Acc.	48.75	49.71	49.87	49.35	49.91	50.41	50.61	50.61	48.78	49.83	49.84	50.58	48.51	48.63	51.39	51.31	48.56	48.56	50.30	50.64
	49.43	49.78	50.84	51.00	48.83	49.03	48.99	48.99	49.75	50.26	50.42	50.85	49.12	49.06	50.43	50.64	49.05	49.28	50.18	50.50
	47.17	49.46	48.74	46.90	51.38	52.38	52.92	52.93	46.85	49.21	49.03	50.56	47.07	47.46	53.41	52.89	47.31	46.98	51.26	51.83
Up Acc.	54.80	53.17	78.02	93.90	18.91	9.29	0.01	0.02	67.74	61.09	66.41	75.62	44.70	40.77	39.66	50.34	43.53	50.14	59.79	64.93
	56.63	54.48	81.13	95.85	20.90	10.43	0.01	0.01	68.89	61.90	67.10	76.27	47.11	43.63	41.56	52.30	46.07	53.21	65.72	72.13
	52.23	51.83	73.39	89.92	15.59	7.86	0.02	0.04	66.62	60.78	66.87	76.09	40.75	36.38	39.17	48.91	39.42	45.61	54.20	57.78
Down Acc.	42.71	46.20	22.27	5.84	80.05	90.46	99.99	99.98	30.17	38.74	33.51	26.07	52.12	56.17	62.74	52.22	53.35	46.86	40.73	36.36
	41.72	44.70	19.11	4.25	77.73	89.09	99.99	99.99	29.64	37.93	32.76	24.21	51.05	54.51	59.45	48.83	51.94	44.93	33.50	27.46
	42.58	47.26	26.72	8.56	83.13	91.94	99.98	99.97	29.20	38.87	33.10	27.82	52.57	57.21	66.04	56.40	54.24	48.11	48.49	46.36
F1	0.49	0.50	0.46	0.37	0.44	0.40	0.34	0.34	0.47	0.49	0.48	0.47	0.48	0.48	0.51	0.51	0.48	0.48	0.50	0.49
	0.49	0.49	0.45	0.37	0.45	0.40	0.33	0.33	0.47	0.49	0.48	0.47	0.49	0.49	0.50	0.50	0.49	0.49	0.48	0.47
	0.47	0.49	0.46	0.38	0.44	0.40	0.35	0.35	0.45	0.49	0.48	0.48	0.47	0.46	0.52	0.53	0.47	0.47	0.51	0.52

Model	FlowState (SF 1)				FlowState (SF 2)				TTM				Toto				Sundial			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
R^2_{OOS}	-115.42	-45.13	-21.14	-19.79	-127.97	-43.16	-5.82	-4.91	-64.61	-38.69	-8.20	-11.88	-784.15	-146.25	-235.58	-114.18	-29.48	-11.59	-2.91	-2.19
	-107.01	-42.18	-18.79	-17.51	-116.40	-40.26	-5.32	-4.23	-54.75	-31.04	-8.49	-10.74	-603.62	-76.87	-193.24	-118.51	-27.96	-10.78	-3.07	-2.74
	-127.63	-49.44	-24.15	-22.78	-143.19	-47.11	-6.57	-5.83	-75.67	-46.64	-8.40	-13.43	-900.83	-114.94	-205.77	-111.49	-32.10	-12.95	-2.59	-1.40
Overall Acc.	48.43	48.56	48.60	48.64	48.51	48.95	49.75	49.89	48.38	48.44	49.10	48.32	48.89	49.80	50.56	50.70	48.48	48.94	50.28	50.54
	49.26	49.02	49.25	49.29	49.42	49.40	49.79	49.98	49.03	49.07	48.96	49.70	49.36	49.99	50.20	50.21	49.19	49.54	50.25	50.13
	46.58	47.35	47.25	47.27	46.60	47.93	49.74	49.90	46.80	46.94	49.06	45.75	47.72	49.66	51.76	52.15	46.81	47.73	50.88	51.83
Up Acc.	49.93	42.08	46.63	46.72	52.69	44.11	39.46	42.17	41.84	43.28	32.18	62.88	52.12	59.80	55.38	55.59	49.35	54.85	59.89	58.20
	51.73	44.78	50.97	51.05	54.67	46.29	44.53	48.02	44.57	46.36	37.02	64.86	54.44	63.49	59.53	60.17	51.69	57.25	62.91	60.75
	46.76	37.76	40.51	40.44	49.88	40.70	33.20	35.28	37.34	38.66	25.71	60.54	49.22	56.26	51.62	51.35	45.60	51.73	56.81	56.25
Down Acc.	46.89	54.76	50.35	50.34	44.36	53.57	59.55	57.17	54.65	53.31	65.37	33.91	45.63	39.85	45.71	45.81	47.51	43.01	40.73	42.98
	46.57	53.25	47.18	47.16	43.81	52.47	54.89	51.62	53.49	51.63	61.01	33.57	43.87	35.64	40.22	39.64	46.39	41.24	36.78	38.89
	46.36	55.77	53.13	53.22	43.64	54.28	64.30	62.74	55.12	54.20	69.68	32.51	46.32	43.68	51.80	52.81	47.80	44.06	45.50	47.84
F1	0.48	0.48	0.48	0.48	0.48	0.49	0.49	0.49	0.48	0.48	0.47	0.47	0.49	0.49	0.50	0.50	0.48	0.49	0.50	0.50
	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.50	0.49	0.49	0.48	0.48	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49
	0.47	0.47	0.47	0.47	0.47	0.47	0.48	0.48	0.46	0.46	0.46	0.45	0.48	0.49	0.52	0.52	0.47	0.48	0.51	0.52

Note: This table presents each metric as a set of three values, ordered from top to bottom: full sample, top 25% of firms by market capitalization (large-cap), and bottom 25% (small-cap) for various predictive models across different window sizes (5, 21, 252, and 512 trading days). The time series foundation models (TSFMs) include Moment (small, base, and large), Lag-Llama, TiRex, FlowState (for SF 1, the scaled factor is set to 0.0656; for SF 2, it is set to 3.43 for daily data with weekly cycles), TTM, Toto, and Sundial. Zero-shot inference is performed using the pre-trained models released by the respective authors. If the model yields a distribution as its output, we use the mean value. Otherwise, we adopt the point estimation approach specified by the respective authors. The hyperparameters are also set according to those proposed by the authors. Metrics are first computed separately for each calendar year using all stock-date observations within that year. The reported values represent the average of these yearly statistics. Metrics include out-of-sample R^2 (R^2_{OOS}), overall directional accuracy, upward and downward classification accuracy, and macro-averaged F1 score. 'Overall Acc.' denotes overall directional accuracy, 'Up Acc.' and 'Down Acc.' represent the model's accuracy in predicting upward and downward excess returns respectively, and 'F1' refers to the macro-averaged F1 score.

Table B.6: Zero-Shot TSFMs - Portfolio Performance (Extended)

Model	Chronos-Bolt (Tiny)				Chronos-Bolt (Mini)				Chronos-Bolt (Small)				Chronos-Bolt (Base)				TimesFM 2.5				Moirai (Small)			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
Annualized Return	-40.76	-33.69	-19.04	-16.62	-38.28	-32.92	-13.28	-9.44	-41.13	-36.14	-9.97	-4.01	-41.12	-32.05	-5.87	-5.94	-37.37	-31.01	9.58	17.57	-38.55	-33.06	-26.35	-26.27
	-14.50	-9.28	-3.78	-2.33	-12.39	-10.31	-2.05	0.21	-14.28	-12.07	0.05	3.35	-13.59	-10.01	1.88	1.49	-12.29	-9.26	10.67	14.70	-11.17	-9.57	-7.15	-7.61
	-26.26	-24.41	-15.26	-14.28	-25.88	-22.61	-11.23	-9.65	-26.86	-24.07	-10.02	-7.36	-27.52	-22.04	-7.75	-7.43	-25.08	-21.75	-1.09	2.87	-27.39	-23.49	-19.20	-18.66
Standard Deviation	9.14	9.40	9.85	9.50	8.88	9.14	8.78	8.22	9.15	9.36	9.80	9.22	8.92	9.23	9.57	9.32	8.49	8.80	8.33	8.18	8.08	8.01	7.58	7.55
	11.38	10.93	10.94	11.03	11.15	11.25	11.74	12.00	11.57	11.15	10.86	10.83	11.48	11.10	11.11	10.74	11.88	11.59	10.87	10.79	11.85	11.55	11.86	11.75
	13.91	14.68	15.66	15.52	13.90	14.36	14.84	14.47	13.80	14.48	15.93	15.74	13.80	14.63	15.83	15.83	13.97	14.30	14.46	14.53	13.38	14.06	14.15	14.03
Sharpe Ratio	-4.46	-3.59	-1.93	-1.75	-4.31	-3.60	-1.51	-1.15	-4.50	-3.86	-1.02	-0.44	-4.61	-3.47	-0.61	-0.64	-4.40	-3.52	1.15	2.15	-4.77	-4.13	-3.47	-3.48
	-1.27	-0.85	-0.35	-0.21	-1.11	-0.92	-0.17	0.02	-1.23	-1.08	0.00	0.31	-1.18	-0.90	0.17	0.14	-1.03	-0.80	0.98	1.36	-0.94	-0.83	-0.60	-0.65
	-1.89	-1.66	-0.97	-0.92	-1.86	-1.58	-0.76	-0.67	-1.95	-1.66	-0.63	-0.47	-1.99	-1.51	-0.49	-0.47	-1.79	-1.52	-0.08	0.20	-2.05	-1.67	-1.36	-1.33
Daily Return (bps)	-16.18	-13.37	-7.56	-6.59	-15.19	-13.06	-5.27	-3.74	-16.32	-14.34	-3.95	-1.59	-16.32	-12.72	-2.33	-2.36	-14.83	-12.31	3.80	6.97	-15.30	-13.12	-10.46	-10.43
	-5.76	-3.68	-1.50	-0.93	-4.92	-4.09	-0.81	0.08	-5.67	-4.79	0.02	1.33	-5.39	-3.97	0.75	0.59	-4.88	-3.67	4.23	5.83	-4.43	-3.80	-2.84	-3.02
	-10.42	-9.69	-6.06	-5.67	-10.27	-8.97	-4.46	-3.83	-10.66	-9.55	-3.97	-2.92	-10.92	-8.75	-3.07	-2.95	-9.95	-8.63	-0.43	1.14	-10.87	-9.32	-7.62	-7.40
Max DD	99.99	99.96	98.83	97.94	99.99	99.95	95.73	89.88	99.99	99.98	91.30	69.28	99.99	99.94	77.95	78.33	99.98	99.92	35.14	28.43	99.99	99.95	99.77	99.77
	97.36	91.16	70.80	57.65	95.68	93.07	58.01	51.94	97.25	95.34	53.91	41.27	96.75	92.45	49.79	49.05	95.70	91.32	39.14	28.80	94.40	91.93	85.84	87.26
	99.80	99.70	97.81	97.23	99.79	99.55	94.73	92.60	99.83	99.68	93.50	88.46	99.85	99.49	89.49	88.96	99.74	99.45	55.52	52.04	99.84	99.63	99.02	98.89
Max DD (1-day)	11.79	7.73	13.07	12.24	6.95	8.07	13.02	11.88	11.79	8.13	13.02	12.24	6.98	7.90	12.17	11.88	6.57	6.79	12.24	12.24	6.04	11.93	12.01	12.40
	6.47	7.12	6.54	5.99	6.39	7.11	6.55	6.58	7.06	7.06	6.45	6.51	6.62	6.84	6.30	6.09	7.43	6.72	6.83	6.62	7.25	6.46	6.81	6.48
	11.65	8.59	11.65	11.65	7.73	8.42	11.65	11.65	11.65	8.39	11.65	11.65	8.04	8.73	11.65	11.65	8.65	8.92	11.65	11.65	8.09	11.65	11.65	11.65
Skew	-2.37	-1.51	-2.97	-2.75	-1.16	-1.53	-3.52	-3.16	-2.48	-1.50	-2.84	-2.67	-1.26	-1.55	-2.63	-2.50	-1.26	-1.60	-3.09	-3.03	-1.44	-3.33	-3.67	-3.86
	-0.51	-0.70	-0.72	-0.66	-0.50	-0.63	-0.65	-0.55	-0.54	-0.62	-0.81	-0.66	-0.49	-0.63	-0.66	-0.65	-0.50	-0.60	-0.67	-0.57	-0.53	-0.55	-0.55	-0.53
	-0.22	0.05	-0.39	-0.36	0.21	0.11	-0.35	-0.33	-0.22	0.08	-0.36	-0.32	0.16	0.04	-0.36	-0.30	0.08	-0.00	-0.31	-0.25	0.09	-0.35	-0.36	-0.30
Kurt	42.26	19.30	48.93	45.84	16.94	21.33	68.54	62.15	42.91	20.56	47.63	47.00	17.66	21.34	41.08	40.23	19.07	19.18	64.43	67.89	19.98	69.07	75.41	84.16
	6.22	6.37	5.47	5.70	6.24	5.68	5.83	5.93	6.38	6.09	5.65	6.06	6.00	6.35	4.71	5.25	6.09	5.92	6.84	6.50	6.43	5.61	6.27	5.99
	15.41	11.19	12.71	12.78	11.14	11.70	13.15	13.97	16.11	11.43	11.38	11.20	12.05	11.33	11.66	11.77	12.38	11.22	14.01	13.93	12.51	16.12	13.15	12.22

Model	Moirai (Base)				Moirai (Large)				Moirai 2				Kairos (10M)				Kairos (23M)				Kairos (50M)			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
Annualized Return	-35.21	-29.43	-21.64	-20.92	-24.67	-26.88	-24.92	-23.58	-39.75	-23.52	9.28	17.11	-20.72	-25.82	-6.83	-9.37	-26.25	-31.13	0.39	-3.52	-23.95	-26.84	-2.06	-4.94
	-8.84	-5.28	-5.91	-6.67	-2.14	-4.97	-6.49	-5.25	-12.59	-5.25	11.66	16.28	-4.47	-8.93	1.05	1.64	-6.89	-9.93	4.74	3.88	-4.71	-7.36	1.66	1.20
	-26.37	-24.16	-15.73	-14.26	-22.53	-21.92	-18.44	-18.32	-27.16	-18.27	-2.38	0.82	-16.25	-16.89	-7.88	-11.01	-19.36	-21.19	-4.35	-7.40	-19.24	-19.48	-3.73	-6.14
Standard Deviation	7.14	7.17	7.01	7.21	6.68	6.77	7.40	7.30	9.37	8.86	8.04	7.37	7.66	7.93	6.67	7.07	6.66	7.83	6.23	6.63	6.76	7.72	7.53	8.45
	11.47	11.53	11.95	11.98	11.62	12.00	11.72	11.63	11.24	11.50	11.51	11.38	13.48	13.23	11.29	11.29	12.26	12.06	12.24	11.55	12.41	12.12	10.74	10.82
	13.12	13.78	13.96	13.80	12.84	13.21	14.01	14.14	14.15	14.48	14.26	13.80	10.58	12.14	13.61	13.61	12.85	13.49	12.80	13.84	13.17	14.05	14.37	14.89
Sharpe Ratio	-4.93	-4.10	-3.09	-2.90	-3.69	-3.97	-3.37	-3.23	-4.24	-2.65	1.15	2.32	-2.71	-3.26	-1.02	-1.33	-3.94	-3.97	0.06	-0.53	-3.54	-3.48	-0.27	-0.58
	-0.77	-0.46	-0.49	-0.56	-0.18	-0.41	-0.55	-0.45	-1.12	-0.46	1.01	1.43	-0.33	-0.67	0.09	0.15	-0.56	-0.82	0.39	0.34	-0.38	-0.61	0.15	0.11
	-2.01	-1.75	-1.13	-1.03	-1.76	-1.66	-1.32	-1.30	-1.92	-1.26	-0.17	0.06	-1.54	-1.39	-0.58	-0.81	-1.51	-1.57	-0.34	-0.54	-1.46	-1.39	-0.26	-0.41
Daily Return (bps)	-13.97	-11.68	-8.59	-8.30	-9.79	-10.67	-9.89	-9.36	-15.77	-9.33	3.68	6.79	-8.22	-10.25	-2.71	-3.72	-10.42	-12.35	0.16	-1.40	-9.50	-10.65	-0.82	-1.96
	-3.51	-2.09	-2.35	-2.65	-0.85	-1.97	-2.57	-2.08	-5.00	-2.08	4.63	6.46	-1.77	-3.54	0.42	0.65	-2.73	-3.94	1.88	1.54	-1.87	-2.92	0.66	0.48
	-10.46	-9.59	-6.24	-5.66	-8.94	-8.70	-7.32	-7.27	-10.78	-7.25	-0.94	0.33	-6.45	-6.70	-3.13	-4.37	-7.68	-8.41	-1.72	-2.94	-7.63	-7.73	-1.48	-2.44
Max DD	99.97	99.89	99.32	99.20	99.67	99.80	99.68	99.57	99.99	99.57	37.58	27.25	99.24	99.76	79.52	88.86	99.77	99.93	17.69	65.03	99.62	99.80	52.37	72.70
	90.22	78.50	81.41	84.25	61.51	78.19	83.35	77.58	95.86	77.91	36.45	32.10	75.97	91.25	48.46	48.76	85.47	92.95	42.78	39.86	76.46	86.86	45.69	48.72
	99.80	99.68	97.86	96.99	99.52	99.45	98.81	98.78	99.84	98.78	68.85	52.15	97.88	98.24	88.35	93.86	99.02	99.36	76.03	88.70	99.01	99.06	75.88	84.70
Max DD (1-day)	5.16	11.64	11.96	12.01	5.70	5.76	5.81	6.01	8.41	12.39	6.62	5.97	4.37	12.85	6.29	11.96	4.57	7.33	12.17	11.96	4.67	12.18	7.32	12.12
	6.68	6.33	7.30	6.65	6.33	6.65	6.69	6.53	7.27	6.24	6.14	6.24	7.06	6.75	6.95	6.58	7.36	6.41	6.87	6.89	6.82	6.44	6.04	6.71
	8.06	11.65	11.65	11.65	8.42	8.39	8.75	8.82	8.69	11.65	9.03	8.88	8.27	11.65	7.34	11.65	8.09	8.28	11.65	11.65	8.18	11.65	8.34	11.65
Skew	-1.35	-3.95	-4.21	-3.92	-1.70	-1.54	-1.70	-1.88	-1.62	-2.93	-1.72	-1.77	-0.44	-3.55	-1.62	-5.10	-0.98	-1.63	-5.48	-5.44	-0.90	-3.98	-1.37	-3.19
	-0.53	-0.43	-0.53	-0.54	-0.47	-0.46	-0.56	-0.60	-0.61	-0.52	-0.64	-0.67	-0.29	-0.33	-0.56	-0.52	-0.45	-0.46	-0.50	-0.52	-0.35	-0.49	-0.54	-0.66
	0.09	-0.35	-0.31	-0.32	0.05	0.10	-0.03	-0.05	0.07	-0.22	-0.08	-0.10	0.12	-0.42	-0.11	-0.55	0.01	0.06	-0.45	-0.50	0.07	-0.33	-0.03	-0.43
Kurt	18.74	91.45	99.27	89.09	23.26	24.34	21.29	21.74	21.26	57.54	21.97	24.82	10.07	84.04	20.31	114.33	14.17	22.56	163.84	133.10	14.74	84.37	21.71	62.92
	6.69	6.02	6.86	6.42	5.94	6.28	6.14	5.82	6.51	5.77	5.46	6.80	6.52	5.33	6.15	5.41	6.41	5.20	5.61	5.85	6.32	5.42	6.34	6.42
	12.16	16.87	13.28	12.25	13.65	12.71	11.09	10.56	12.44	15.72	12.39	13.11	19.56	22.66	9.01	15.05	11.64	12.85	14.59	14.03	10.68	16.07	9.10	12.63

Note: This table reports average yearly portfolio performance metrics across different rolling window sizes (5, 21, 252, and 512 trading days) for each model. The time series foundation models (TSFMs) include Chronos-Bolt (tiny, mini, small, and base), TimesFM 2.5, Uni2TS (Moirai 1.1 small, base, and large, as well as Moirai 2), and Kairos (10M, 23M and 50M parameter models). Zero-shot inference is performed using the pre-trained models released by the respective authors. If the model yields a distribution as its output, we use the mean value. Otherwise, we adopt the point estimation approach specified by the respective authors. The hyperparameters are also set according to those proposed by the authors. Each cell displays three values from top to bottom: long–short portfolio, long-only leg, and short-only leg. Metrics include annualized return, standard deviation, Sharpe ratio, daily return (in basis points), maximum drawdown (Max DD), one-day maximum drawdown (Max DD (1-day)), skewness, and kurtosis of portfolio. Portfolios are formed using decile sorting based on model forecasts, with equal weighting across stocks.

Table B.7: Zero-Shot TSFMs - Portfolio Performance (Extended) - Continued

Model	Moment (Small)				Moment (Base)				Moment (Large)				Lag-Llama				TiRex			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
	-30.32	-10.00	-1.80	-15.00	-24.61	-8.05	1.92	10.43	-29.91	-7.98	-2.88	27.75	-39.93	-40.29	25.30	24.30	-37.21	-30.59	17.28	28.33
Annualized Return	-9.29	-1.27	4.43	-0.11	-4.49	0.09	4.68	10.41	-9.57	1.14	3.66	21.27	-13.48	-13.63	20.24	18.73	-10.90	-8.43	14.68	20.65
	-21.03	-8.72	-6.23	-14.89	-20.12	-8.15	-2.75	0.02	-20.34	-9.12	-6.53	6.48	-26.45	-26.65	5.06	5.58	-26.32	-22.16	2.60	7.68
	9.75	7.85	6.80	8.55	10.65	10.36	12.24	11.02	9.43	6.75	6.32	7.02	8.99	8.76	6.52	6.27	9.04	8.53	8.06	7.85
Standard Deviation	11.92	11.66	13.20	14.37	9.72	8.48	6.68	7.03	12.43	12.26	11.52	11.40	11.50	11.28	11.01	11.44	11.51	11.64	11.23	11.09
	14.20	13.19	11.09	9.05	15.25	15.52	16.21	15.39	13.59	13.35	13.18	12.85	13.98	13.90	13.30	12.71	14.25	14.03	14.37	14.21
	-3.11	-1.27	-0.27	-1.75	-2.31	-0.78	0.16	0.95	-3.17	-1.18	-0.45	3.95	-4.44	-4.60	3.88	3.87	-4.12	-3.59	2.14	3.61
Sharpe Ratio	-0.78	-0.11	0.34	-0.01	-0.46	0.01	0.70	1.48	-0.77	0.09	0.32	1.87	-1.17	-1.21	1.84	1.64	-0.95	-0.72	1.31	1.86
	-1.48	-0.66	-0.56	-1.64	-1.32	-0.52	-0.17	0.00	-1.50	-0.68	-0.50	0.50	-1.89	-1.92	0.38	0.44	-1.85	-1.58	0.18	0.54
	-12.03	-3.97	-0.72	-5.95	-9.77	-3.20	0.76	4.14	-11.87	-3.16	-1.14	11.01	-15.84	-15.99	10.04	9.64	-14.77	-12.14	6.86	11.24
Daily Return (bps)	-3.69	-0.51	1.76	-0.05	-1.78	0.04	1.86	4.13	-3.80	0.45	1.45	8.44	-5.35	-5.41	8.03	7.43	-4.32	-3.35	5.82	8.19
	-8.34	-3.46	-2.47	-5.91	-7.99	-3.23	-1.09	0.01	-8.07	-3.62	-2.59	2.57	-10.49	-10.58	2.01	2.21	-10.44	-8.79	1.03	3.05
	99.91	91.32	44.47	97.37	99.69	87.36	40.38	27.40	99.91	84.74	53.57	14.08	99.99	99.99	11.92	28.03	99.98	99.92	22.22	14.35
Max DD	91.40	52.67	44.11	62.15	71.57	47.98	28.14	20.00	92.11	43.71	37.90	22.61	96.69	96.75	28.78	35.41	93.97	89.59	24.62	22.54
	99.35	90.12	81.02	97.00	99.23	90.15	78.69	56.84	99.22	89.96	83.95	42.41	99.81	99.82	44.83	46.33	99.81	99.49	50.73	47.77
	8.09	6.89	4.20	3.55	8.33	6.78	6.72	5.71	7.75	5.21	6.17	12.67	12.08	6.09	4.34	4.17	7.74	6.93	12.24	12.24
Max DD (1-day)	7.92	6.60	8.44	8.09	7.38	4.17	4.84	5.64	7.87	6.29	5.87	5.88	6.45	6.73	6.66	6.71	7.37	6.50	6.41	6.62
	9.71	8.93	4.77	6.15	9.26	8.64	6.50	6.23	9.83	7.49	8.43	11.65	11.65	7.83	7.08	7.23	9.00	8.56	11.65	11.65
	-1.87	-1.40	0.05	0.31	-1.50	-0.89	-0.44	-0.39	-1.79	-0.61	-1.15	-4.73	-2.55	-1.30	-0.73	-0.96	-1.53	-1.44	-3.24	-3.34
Skew	-0.62	-0.51	-0.31	-0.19	-0.83	-0.49	-0.71	-0.80	-0.56	-0.30	-0.43	-0.29	-0.50	-0.62	-0.48	-0.51	-0.54	-0.50	-0.54	-0.53
	-0.14	-0.04	0.07	-0.05	-0.18	0.02	0.08	0.15	-0.13	0.15	0.03	-0.45	-0.29	0.09	0.12	0.10	0.08	0.07	-0.28	-0.26
	22.01	18.92	17.25	7.31	17.85	12.18	5.13	4.49	20.70	16.37	20.96	130.38	46.21	16.36	9.95	10.08	20.92	19.29	72.15	80.33
Kurt	6.80	6.00	7.97	7.36	8.22	6.47	18.07	18.21	6.62	6.16	6.42	6.15	6.23	6.76	7.23	6.15	5.86	5.74	5.56	6.03
	13.75	10.97	7.31	10.41	10.84	8.61	5.20	5.14	15.35	9.49	10.49	19.77	15.37	10.65	8.55	10.38	12.35	11.43	14.46	15.42

Model	FlowState (SF 1)				FlowState (SF 2)				TTM				Toto				Sundial			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
	-41.67	-40.08	-40.30	-39.52	-38.16	-30.37	-19.59	-17.70	-41.99	-41.95	-33.82	-40.07	-21.50	-10.02	20.91	25.72	-36.99	-26.08	13.73	25.48
Annualized Return	-14.73	-13.54	-12.94	-12.37	-13.23	-9.91	-4.92	-4.69	-13.65	-13.75	-6.38	-12.06	-4.47	-0.54	15.39	18.33	-11.97	-7.67	11.32	17.18
	-26.94	-26.54	-27.36	-27.14	-24.93	-20.45	-14.67	-13.00	-28.34	-28.20	-27.44	-28.01	-17.03	-9.48	5.52	7.39	-25.02	-18.41	2.41	8.31
	8.95	9.31	9.30	9.19	8.86	8.50	8.80	8.69	9.13	9.49	10.17	9.30	5.78	6.60	4.71	4.14	9.34	8.60	6.06	6.05
Standard Deviation	11.81	11.42	11.48	11.51	11.79	11.65	11.17	11.21	10.85	10.62	10.19	11.68	11.65	11.73	11.57	11.65	11.73	11.79	12.16	11.98
	13.57	14.22	14.11	14.07	13.47	14.01	15.05	14.95	13.55	14.17	15.26	14.03	13.08	13.67	12.95	12.57	14.31	14.32	13.53	13.55
	-4.65	-4.30	-4.34	-4.30	-4.31	-3.57	-2.23	-2.04	-4.60	-4.42	-3.33	-4.31	-3.72	-1.52	4.44	6.22	-3.96	-3.03	2.27	4.21
Sharpe Ratio	-1.25	-1.19	-1.13	-1.08	-1.12	-0.85	-0.44	-0.42	-1.26	-1.30	-0.63	-1.03	-0.38	-0.05	1.33	1.57	-1.02	-0.65	0.93	1.43
	-1.99	-1.87	-1.94	-1.93	-1.85	-1.46	-0.97	-0.87	-2.09	-1.99	-1.80	-2.00	-1.30	-0.69	0.43	0.59	-1.75	-1.29	0.18	0.61
	-16.54	-15.91	-15.99	-15.68	-15.14	-12.05	-7.77	-7.02	-16.66	-16.65	-13.42	-15.90	-8.53	-3.97	8.30	10.20	-14.68	-10.35	5.45	10.11
Daily Return (bps)	-5.85	-5.37	-5.13	-4.91	-5.25	-3.93	-1.95	-1.86	-5.42	-5.46	-2.53	-4.79	-1.77	-0.21	6.11	7.27	-4.75	-3.04	4.49	6.82
	-10.69	-10.53	-10.86	-10.77	-9.89	-8.12	-5.82	-5.16	-11.25	-11.19	-10.89	-11.12	-6.76	-3.76	2.19	2.93	-9.93	-7.31	0.96	3.30
	99.99	99.99	99.99	99.99	99.99	99.91	98.94	98.37	99.99	99.99	99.96	99.99	99.30	90.18	10.43	8.96	99.98	99.76	23.41	14.36
Max DD	97.56	96.72	96.17	95.69	96.56	92.60	77.29	75.99	96.69	96.71	81.94	95.39	73.66	50.88	26.07	27.33	95.30	87.54	33.07	24.51
	99.83	99.82	99.85	99.84	99.73	99.25	97.34	96.14	99.88	99.87	99.85	99.87	98.32	91.13	43.05	43.72	99.74	98.82	50.55	44.36
	11.79	12.09	13.07	12.08	11.79	11.71	13.18	13.18	7.81	12.01	12.07	12.78	5.15	5.43	7.42	6.10	8.21	12.01	7.77	12.41
Max DD (1-day)	6.60	6.19	6.36	6.50	6.40	6.27	6.67	6.66	7.16	5.90	5.99	6.56	7.16	6.29	6.24	5.79	7.83	5.81	6.70	7.02
	11.65	11.65	11.65	11.65	11.65	11.65	11.65	11.65	8.08	11.65	11.65	11.65	7.92	8.44	7.79	7.82	9.27	11.65	8.12	11.65
	-2.45	-2.49	-2.82	-2.53	-2.62	-2.68	-3.37	-3.35	-1.64	-2.63	-2.34	-3.04	-1.33	-1.62	-4.05	-2.88	-1.74	-3.13	-2.73	-7.30
Skew	-0.47	-0.55	-0.57	-0.55	-0.44	-0.35	-0.58	-0.64	-0.73	-0.74	-0.69	-0.59	-0.56	-0.52	-0.42	-0.33	-0.62	-0.58	-0.42	-0.38
	-0.28	-0.28	-0.27	-0.28	-0.33	-0.24	-0.31	-0.32	0.10	-0.24	-0.29	-0.32	0.10	0.02	0.07	0.13	-0.04	-0.27	0.05	-0.31
	43.30	43.11	53.75	44.63	45.77	50.87	67.59	70.45	20.59	43.39	36.63	52.83	18.49	22.40	90.50	82.53	22.72	58.78	61.19	225.81
Kurt	6.08	5.79	5.71	5.85	5.96	6.00	6.11	5.99	6.94	5.68	5.16	5.88	7.86	6.24	6.39	6.17	6.42	4.98	5.50	5.65
	15.86	15.21	15.35	15.33	15.86	15.08	12.67	12.78	12.10	15.42	13.16	16.43	10.30	11.27	10.85	11.92	13.16	16.43	12.26	16.94

Note: This table reports average yearly portfolio performance metrics across different rolling window sizes (5, 21, 252, and 512 trading days) for each model. The time series foundation models (TSFMs) include Moment (small, base, and large), Lag-Llama, TiRex, FlowState (for SF 1, the scaled factor is set to 0.0656; for SF 2, it is set to 3.43 for daily data with weekly cycles), TTM, Toto, and Sundial. Zero-shot inference is performed using the pre-trained models released by the respective authors. If the model yields a distribution as its output, we use the mean value. Otherwise, we adopt the point estimation approach specified by the respective authors. The hyperparameters are also set according to those proposed by the authors. Each cell displays three values from top to bottom: long–short portfolio, long-only leg, and short-only leg. Metrics include annualized return, standard deviation, Sharpe ratio, daily return (in basis points), maximum drawdown (Max DD), one-day maximum drawdown (Max DD (1-day)), skewness, and kurtosis of portfolio returns. Portfolios are formed using decile sorting based on model forecasts, with equal weighting across stocks.

Table B.8: Zero-Shot TSFMs - Spread Portfolio Performance (Extended)

Model	Chronos-Bolt (Tiny)				Chronos-Bolt (Mini)				Chronos-Bolt (Small)				Chronos-Bolt (Base)				TimesFM 2.5				Moirai (Small)			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
Low	52.52	48.82	30.52	28.57	51.77	45.22	22.46	19.30	53.71	48.14	20.03	14.72	55.05	44.09	15.49	14.86	50.16	43.51	2.17	-5.74	54.77	46.98	38.40	37.32
2	28.51	28.72	24.00	22.27	30.90	29.88	18.76	16.38	29.86	29.87	16.62	13.60	29.61	28.59	16.12	16.18	28.55	29.32	10.50	8.78	29.20	28.00	27.17	27.05
3	21.42	22.36	19.06	17.91	20.80	23.62	16.83	15.91	22.43	24.73	16.15	12.42	21.58	23.94	14.41	14.88	21.65	22.31	11.70	10.33	20.27	21.15	21.11	21.74
4	15.84	15.72	15.19	14.76	16.55	18.13	14.67	12.63	17.64	17.89	13.82	12.78	16.21	17.68	12.67	13.27	15.89	16.70	11.45	10.30	15.68	16.12	17.06	16.88
5	12.08	11.15	12.20	11.75	10.97	12.56	12.82	12.62	12.08	12.82	12.34	11.42	11.88	12.54	12.12	12.48	12.27	12.17	10.51	10.71	11.24	11.65	12.48	13.04
6	9.06	7.02	8.94	9.12	8.20	7.80	11.13	11.29	8.17	8.12	10.16	10.39	7.32	8.08	11.02	11.56	8.47	8.18	9.83	10.69	6.63	8.16	8.65	8.44
7	5.86	2.97	6.27	6.99	3.71	3.66	8.74	8.84	4.42	4.58	9.56	10.43	4.20	3.59	9.78	10.07	5.18	4.29	10.84	10.84	3.51	4.06	4.80	4.66
8	1.00	-1.08	3.45	3.94	-0.25	-1.70	7.01	8.07	-0.38	-2.13	7.59	9.58	-0.12	-0.04	8.90	8.28	0.20	-0.47	11.38	12.31	-0.72	0.91	0.92	1.13
9	-6.02	-5.89	-0.82	0.60	-6.63	-7.32	2.93	5.79	-8.14	-8.66	4.88	9.22	-7.29	-7.21	6.99	6.70	-6.56	-6.25	11.56	13.64	-7.01	-6.63	-5.04	-3.77
High	-29.01	-18.56	-7.56	-4.67	-24.79	-20.62	-4.10	0.43	-28.56	-24.14	0.10	6.69	-27.19	-20.02	3.76	2.98	-24.57	-18.52	21.33	29.40	-22.34	-19.15	-14.30	-15.23
H-L	-40.76	-33.69	-19.04	-16.62	-38.28	-32.92	-13.28	-9.44	-41.13	-36.14	-9.97	-4.01	-41.12	-32.05	-5.87	-5.94	-37.37	-31.01	9.58	17.57	-38.55	-33.06	-26.35	-26.27

Model	Moirai (Base)				Moirai (Large)				Moirai 2				Kairos (10M)				Kairos (23M)				Kairos (50M)			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
Low	52.73	48.32	31.46	28.52	45.07	43.83	36.87	36.65	54.31	36.55	4.76	-1.64	32.51	33.77	15.76	22.02	38.72	42.39	8.69	14.81	38.48	38.96	7.45	12.28
2	27.57	26.93	24.55	22.33	24.35	25.41	23.87	22.78	31.79	23.06	7.06	5.49	19.72	22.60	15.43	17.67	26.33	28.36	9.61	13.97	22.57	26.15	14.70	15.90
3	19.04	19.58	20.02	19.97	18.21	17.91	18.91	17.83	24.24	19.57	7.72	8.11	16.10	18.38	16.38	15.20	18.96	21.66	10.70	12.66	16.54	19.81	14.26	15.50
4	14.00	14.72	16.72	16.61	13.26	14.62	15.21	14.04	16.32	15.25	9.47	9.04	15.14	16.76	13.83	12.58	15.06	16.00	10.88	11.31	13.92	15.51	14.63	14.31
5	10.74	10.42	12.88	12.92	9.83	10.44	12.36	11.94	11.74	11.53	9.51	8.92	12.02	14.07	11.62	10.45	11.46	12.24	11.52	11.40	10.81	11.76	13.31	13.52
6	7.15	6.05	9.95	10.54	7.29	7.86	9.24	9.17	6.89	8.37	10.33	10.26	10.87	11.39	11.27	9.48	8.64	9.61	11.87	10.20	8.55	8.20	13.21	11.11
7	5.24	3.36	7.09	8.55	2.54	4.86	5.94	6.61	2.67	6.64	11.42	10.89	7.12	8.00	9.65	8.22	5.13	4.47	12.88	10.40	6.84	4.99	11.72	10.67
8	-0.95	-0.83	3.58	4.99	-0.63	0.96	3.60	3.36	-2.87	2.19	12.73	11.78	5.44	4.21	8.47	6.78	3.06	1.45	12.48	9.27	3.49	2.33	10.00	8.84
9	-6.63	-6.76	-3.19	0.15	-4.41	-4.75	-1.77	-0.62	-8.66	-1.41	14.93	15.83	1.27	-0.05	6.75	5.57	-2.33	-5.04	13.16	9.47	-0.54	-1.73	8.67	6.74
High	-17.68	-10.55	-11.82	-13.33	-4.28	-9.93	-12.98	-10.51	-25.18	-10.49	23.32	32.57	-8.93	-17.86	2.10	3.28	-13.78	-19.87	9.47	7.77	-9.43	-14.72	3.33	2.40
H-L	-35.21	-29.43	-21.64	-20.92	-24.67	-26.88	-24.92	-23.58	-39.75	-23.52	9.28	17.11	-20.72	-25.82	-6.83	-9.37	-26.25	-31.13	0.39	-3.52	-23.95	-26.84	-2.06	-4.94

Note: This table presents the annualized average returns of decile spread portfolios formed using forecasts from various predictive models across different window sizes (5, 21, 252, and 512 trading days). The time series foundation models (TSFMs) include Chronos-Bolt (tiny, mini, small, and base), TimesFM 2.5, Uni2TS (Moirai 1.1 small, base, and large, as well as Moirai 2), and Kairos (10M, 23M and 50M parameter models). Zero-shot inference is performed using the pre-trained models released by the respective authors. If the model yields a distribution as its output, we use the mean value. Otherwise, we adopt the point estimation approach specified by the respective authors. The hyperparameters are also set according to those proposed by the authors. For each model and window size, the returns of decile portfolios from the lowest (Low) to the highest (High) forecasted return are reported. The final row (H-L) represents the return spread between the highest and lowest deciles. Portfolios are constructed using equal-weighted decile sorting based on predicted returns.

Table B.9: Zero-Shot TSFMs - Spread Portfolio Performance (Extended) - Continued

Model	Moment (Small)				Moment (Base)				Moment (Large)				Lag-Llama				TiRex			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
Low	42.05	17.45	12.46	29.78	40.25	16.29	5.51	-0.03	40.68	18.23	13.07	-12.97	52.89	53.30	-10.12	-11.15	52.64	44.32	-5.21	-15.36
2	29.03	15.76	11.96	16.46	25.98	17.76	10.67	4.80	26.49	14.54	12.21	1.94	28.04	27.29	-2.02	-1.67	29.53	28.22	7.24	4.08
3	20.78	15.63	11.80	13.61	18.44	15.22	12.34	8.72	20.81	12.66	11.11	5.97	19.68	20.69	3.66	4.00	20.88	21.23	10.61	8.16
4	16.32	14.97	11.24	11.82	13.92	14.55	11.65	9.59	20.81	12.66	11.11	5.97	15.74	15.42	6.85	7.42	15.06	16.68	10.43	10.48
5	12.37	13.22	10.98	10.53	10.63	12.28	13.64	11.44	13.82	11.83	10.97	9.34	11.76	11.53	9.88	10.00	11.27	11.54	10.95	10.44
6	8.26	11.58	11.54	8.71	7.58	10.77	13.23	11.80	8.81	11.11	10.67	10.06	8.11	8.27	11.77	11.97	6.65	7.01	10.61	10.48
7	5.16	10.56	11.30	8.47	4.83	9.51	12.69	13.45	6.05	10.39	11.90	12.65	5.62	5.11	13.42	13.67	3.45	4.46	11.19	12.01
8	-0.14	9.07	10.02	6.66	1.14	8.57	11.24	14.65	1.39	10.01	11.63	15.01	1.62	1.36	15.66	17.03	-0.51	-0.01	12.17	13.24
9	-3.99	5.58	11.10	5.45	-2.57	6.12	10.94	16.03	-3.78	7.79	10.75	18.73	-5.26	-4.47	21.68	22.56	-5.96	-5.34	13.90	16.44
High	-18.58	-2.55	8.86	-0.23	-8.97	0.19	9.35	20.82	-19.14	2.28	7.32	42.54	-26.96	-27.27	40.48	37.45	-21.79	-16.86	29.36	41.30
H-L	-30.32	-10.00	-1.80	-15.00	-24.61	-8.05	1.92	10.43	-29.91	-7.98	-2.88	27.75	-39.93	-40.29	25.30	24.30	-37.21	-30.59	17.28	28.33

Model	FlowState (SF 1)				FlowState (SF 2)				TTM				Toto				Sundial			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
Low	53.88	53.07	54.73	54.28	49.86	40.91	29.33	26.01	56.69	56.39	54.87	56.02	34.07	18.96	-11.04	-14.78	50.05	36.83	-4.82	-16.62
2	25.97	27.11	27.44	27.62	24.91	23.76	19.08	18.74	32.18	31.52	29.16	30.15	25.28	16.34	4.55	4.62	30.25	24.22	7.94	4.98
3	18.90	20.40	19.95	18.53	18.23	18.91	16.56	15.54	23.22	22.90	21.15	21.88	20.10	13.94	9.62	8.26	23.31	19.53	10.16	8.66
4	14.95	15.62	14.96	14.57	14.16	15.36	13.72	13.97	17.37	17.28	14.50	15.89	15.15	12.78	9.93	10.00	16.44	15.67	10.03	10.64
5	11.68	12.11	11.76	11.01	12.12	12.42	12.43	12.30	11.44	11.86	8.83	10.83	11.16	12.42	10.98	10.38	11.07	13.10	12.19	11.00
6	9.88	8.64	8.14	7.74	9.86	9.79	10.59	11.10	6.48	7.61	4.02	6.92	8.71	11.78	12.37	11.52	8.11	9.98	12.77	11.45
7	7.06	5.21	5.17	5.51	7.87	7.12	8.77	9.63	2.51	2.47	-0.01	2.67	5.19	9.60	13.00	12.47	3.26	6.41	12.24	13.45
8	3.07	1.62	0.46	1.13	12.74	4.19	3.98	6.74	7.78	-2.31	-2.77	-1.11	1.99	8.55	14.24	14.45	-0.19	2.63	13.11	16.05
9	-4.70	-5.45	-5.49	-4.40	-3.50	-1.18	3.88	5.57	-8.96	-8.98	-5.77	-7.90	-1.46	7.96	16.84	17.68	-7.13	-1.77	15.01	17.31
High	-29.46	-27.09	-25.88	-24.75	-26.46	-19.83	-9.84	-9.38	-27.30	-27.50	-12.77	-24.12	-8.94	-1.07	30.78	36.66	-23.94	-15.34	22.64	34.35
H-L	-41.67	-40.08	-40.30	-39.52	-38.16	-30.37	-19.59	-17.70	-41.99	-41.95	-33.82	-40.07	-21.50	-10.02	20.91	25.72	-36.99	-26.08	13.73	25.48

Note: This table presents the annualized average returns of decile spread portfolios formed using forecasts from various predictive models across different window sizes (5, 21, 252, and 512 trading days). The time series foundation models (TSFMs) include Moment (small, base, and large), Lag-Llama, TiRex, FlowState (for SF 1, the scaled factor is set to 0.0656; for SF 2, it is set to 3.43 for daily data with weekly cycles), TTM, Toto, and Sundial. Zero-shot inference is performed using the pre-trained models released by the respective authors. If the model yields a distribution as its output, we use the mean value. Otherwise, we adopt the point estimation approach specified by the respective authors. The hyperparameters are also set according to those proposed by the authors. For each model and window size, the returns of decile portfolios from the lowest (Low) to the highest (High) forecasted return are reported. The final row (H-L) represents the return spread between the highest and lowest deciles. Portfolios are constructed using equal-weighted decile sorting based on predicted returns.

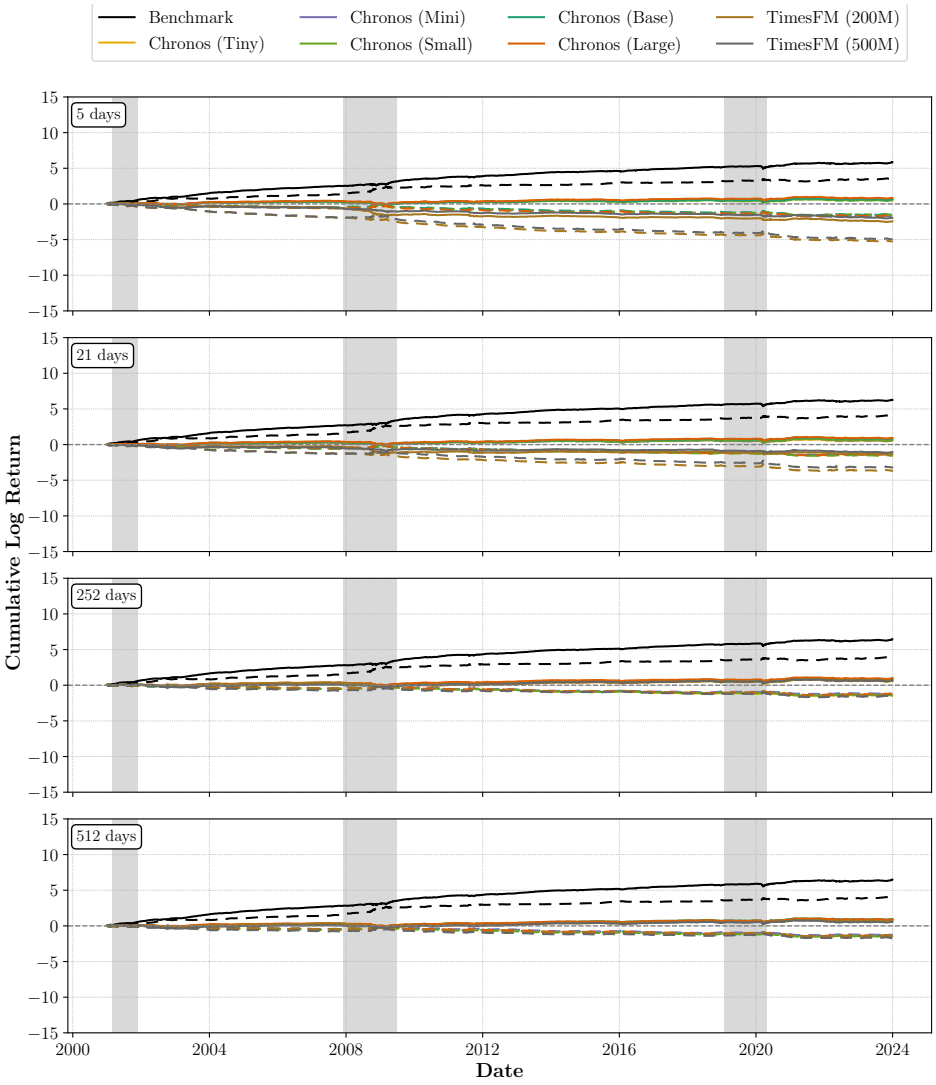
Table B.10: Fine-Tuned TSFMs - Spread Portfolio Performance

Model	Benchmark				Chronos (Tiny)				Chronos (Mini)				Chronos (Small)			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
Low	-31.91	-36.52	-35.00	-36.06	14.03	12.01	12.29	11.16	14.14	11.55	9.43	10.76	12.00	13.31	11.94	12.67
2	-0.94	-3.02	-3.94	-4.06	12.67	12.36	11.74	12.20	14.32	11.58	11.83	11.58	14.99	12.14	10.64	11.34
3	4.28	4.49	2.72	3.30	11.27	11.63	10.77	11.26	11.99	12.31	12.30	11.79	13.39	11.33	12.79	11.69
4	7.46	8.32	6.38	6.84	12.31	11.89	12.13	12.24	11.45	12.22	12.20	12.39	11.71	11.46	12.34	11.70
5	9.38	10.24	9.84	9.36	11.01	12.25	12.45	11.44	10.36	11.61	11.80	12.14	11.14	11.44	11.93	10.98
6	11.51	11.43	11.88	12.41	10.90	11.54	12.19	11.61	11.35	11.97	12.25	11.86	11.07	11.95	11.56	12.65
7	15.23	15.19	15.65	14.06	10.88	12.65	11.87	11.98	11.12	11.33	11.22	10.88	11.18	11.45	10.58	11.08
8	18.19	18.65	18.23	19.32	11.21	11.83	11.79	11.22	11.17	11.15	11.60	11.48	10.47	11.58	11.94	10.86
9	24.92	25.96	27.53	27.67	11.63	10.09	10.48	10.83	9.72	11.47	11.68	11.18	10.87	10.54	10.16	10.78
High	53.17	56.53	58.00	58.44	7.17	6.84	7.38	9.14	7.47	7.88	8.78	9.04	6.26	7.88	9.22	9.34
H-L	42.54	46.52	46.50	47.25	-3.43	-2.59	-2.46	-1.01	-3.34	-1.84	-0.32	-0.86	-2.87	-2.72	-1.36	-1.67

Model	Chronos (Base)				Chronos (Large)				TimesFM 1 (200M)				TimesFM 2 (500M)			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
Low	11.58	10.39	11.78	11.60	13.14	10.94	9.74	11.26	44.28	30.81	10.82	12.20	41.56	26.82	12.13	14.30
2	12.17	12.26	10.90	10.56	14.58	11.36	11.79	11.92	29.05	22.37	11.09	11.04	27.85	21.47	12.02	12.79
3	14.38	11.87	11.19	11.73	11.71	10.92	11.18	12.00	21.30	18.69	11.58	12.10	19.69	17.37	12.57	13.35
4	13.35	11.72	12.01	11.93	12.03	11.30	11.88	11.03	16.15	16.16	13.30	12.31	15.77	15.21	13.04	12.60
5	11.29	10.58	12.23	12.63	11.09	11.60	11.57	11.80	12.17	12.46	12.10	12.03	12.20	12.60	11.67	11.24
6	12.01	12.37	12.78	11.07	10.66	11.72	11.74	11.56	8.29	10.39	11.28	11.10	8.25	9.66	11.30	10.48
7	10.54	12.04	11.67	12.38	9.92	12.76	11.35	11.92	4.39	6.80	11.06	10.71	4.28	8.20	10.60	10.93
8	10.66	11.12	10.84	10.97	10.70	10.75	12.45	11.49	0.46	4.04	11.83	11.78	1.63	5.88	11.19	10.08
9	9.52	11.56	10.78	10.65	10.31	12.18	11.19	10.73	-4.98	-0.36	10.85	10.73	-4.13	2.02	9.50	9.25
High	7.58	9.17	8.90	9.55	8.93	9.55	10.20	9.38	-19.88	-10.10	7.36	7.26	-15.87	-7.98	7.25	6.25
H-L	-2.00	-0.61	-1.44	-1.03	-2.11	-0.69	0.23	-0.94	-32.08	-20.45	-1.73	-2.47	-28.72	-17.40	-2.44	-4.02

Note: This table presents the annualized average returns of decile spread portfolios formed using forecasts from various predictive models across different window sizes (5, 21, 252, and 512 trading days). The benchmark model is CatBoost, the best-performing model among the benchmarks. The time series foundation models (TSFMs) include Chronos (tiny, mini, small, base, and large) and TimesFM (version 1 with 200 million and version 2 with 500 million parameters). The models released by the respective authors are fine-tuned on an annual basis. For each model and window size, the returns of decile portfolios from the lowest (Low) to the highest (High) forecasted return are reported. The final row (H-L) represents the return spread between the highest and lowest deciles. Portfolios are constructed using equal-weighted decile sorting based on predicted returns.

Figure B.2: Cumulative Log Returns of Fine-Tuned TSFMs: Long and Short Portfolios



Note: This figure displays the cumulative log returns of long and short portfolios, separately constructed using various forecasting models over rolling windows of 5, 21, 252, and 512 trading days. The benchmark model is CatBoost, the best-performing model among the benchmarks. The time series foundation models (TSFMs) include Chronos (tiny, mini, small, base, and large) and TimesFM (version 1 with 200 million and version 2 with 500 million parameters). The models released by the respective authors are fine-tuned on an annual basis. Each subplot corresponds to a specific horizon, as indicated by the text labels in the upper-left corners. For each model, the solid line denotes the performance of the long portfolio, while the dashed line represents the corresponding short portfolio. The benchmark model (CatBoost) is highlighted in black with bold lines. Shaded areas indicate U.S. recession periods, as defined by the National Bureau of Economic Research (NBER). All portfolios are equally weighted.

Table B.11: Comparison of Daily Out-of-Sample Forecasting Performance of Pre-Trained TSFMs Using Diebold-Mariano Test

		Chronos (Tiny)				Chronos (Mini)				Chronos (Small)				TimesFM (8M)				TimesFM (20M)			
		5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
Benchmark	5	-8.97*	-4.55*	-2.97*	-3.41*	-12.60*	-6.02*	-2.15*	-2.51*	-10.87*	-5.56*	-2.69*	-2.66*	-28.48*	-20.24*	-32.31*	-27.52*	-29.19*	-22.93*	-26.47*	-25.38*
	21	-9.85*	-6.51*	-5.10*	-5.55*	-13.66*	-8.49*	-3.46*	-3.79*	-12.53*	-7.06*	-4.12*	-4.02*	-28.89*	-20.58*	-32.64*	-27.72*	-29.64*	-23.24*	-26.68*	-25.54*
	252	-9.87*	-6.67*	-5.53*	-6.10*	-13.75*	-8.84*	-3.63*	-3.97*	-12.74*	-7.22*	-4.27*	-4.19*	-28.86*	-20.40*	-32.30*	-27.60*	-29.64*	-23.10*	-26.38*	-25.34*
	512	-9.96*	-6.70*	-5.29*	-5.85*	-13.80*	-8.75*	-3.52*	-3.86*	-12.73*	-7.22*	-4.17*	-4.08*	-28.91*	-20.58*	-32.58*	-27.72*	-29.67*	-23.26*	-26.76*	-25.62*
Chronos (Tiny)	5					-1.61	8.00*	7.22*	6.77*	2.49*	6.82*	7.80*	7.35*	-31.37*	-19.11*	-23.37*	-21.26*	-32.43*	-19.62*	-12.76*	-13.21*
	21					-11.94*	-2.36*	1.89	1.40	-16.05*	-2.15*	1.95	1.68	-30.78*	-21.57*	-30.89*	-26.00*	-31.77*	-23.73*	-22.13*	-21.28*
	252					-12.23*	-5.54*	0.22	-0.46	-11.88*	-4.02*	-0.40	-0.49	-29.17*	-19.50*	-30.80*	-26.37*	-30.25*	-22.40*	-22.88*	-22.17*
	512					-11.72*	-3.93*	0.42	-0.26	-10.08*	-3.43*	-0.11	-0.19	-28.60*	-18.92*	-30.78*	-26.52*	-29.50*	-21.57*	-22.69*	-22.24*
Chronos (Mini)	5									4.40*	9.52*	11.15*	10.83*	-27.17*	-15.32*	-22.90*	-21.00*	-28.02*	-16.67*	-12.24*	-12.84*
	21									-12.17*	-0.94	3.51*	3.07*	-29.72*	-20.45*	-30.91*	-25.96*	-30.63*	-22.71*	-21.90*	-21.13*
	252									-9.18*	-3.25*	-0.80	-1.42	-28.11*	-18.35*	-30.11*	-26.11*	-29.29*	-21.18*	-21.53*	-21.12*
	512									-8.20*	-2.71*	0.11	0.05	-27.60*	-17.84*	-29.86*	-26.10*	-28.70*	-20.58*	-21.09*	-20.89*
Chronos (Small)	5													-30.50*	-18.81*	-25.94*	-22.76*	-31.80*	-20.85*	-15.72*	-15.73*
	21													-29.03*	-19.45*	-29.30*	-25.34*	-30.03*	-21.78*	-20.40*	-20.03*
	252													-28.86*	-19.35*	-30.61*	-26.12*	-30.20*	-22.36*	-21.91*	-21.22*
	512													-28.40	-18.67*	-30.30	-26.06*	-29.61	-21.56*	-21.41	-20.94*
TimesFM (8M)	5																	-1.71	15.90*	17.96*	16.57*
	21																	-18.46*	-0.55	5.97*	4.36*
	252																	-10.45*	4.35*	14.83*	11.33*
	512																	-6.08*	7.33*	14.73*	17.21*

Note: This table reports the modified Diebold–Mariano (DM) test statistics, following the approach of Gu et al. (2020), for out-of-sample stock-level forecast comparisons across pre-trained time series foundation models (TSFMs). A positive statistic indicates that the model in the column achieves superior predictive performance relative to the model in the corresponding row. An asterisk denotes statistical significance at the 5% level. The benchmark model is CatBoost, the best-performing model among the benchmarks. TSFMs include Chronos (tiny, mini, and small) and TimesFM (with 8 million and 20 million parameters). The models are pre-trained from scratch using U.S. excess return data on an annual basis.

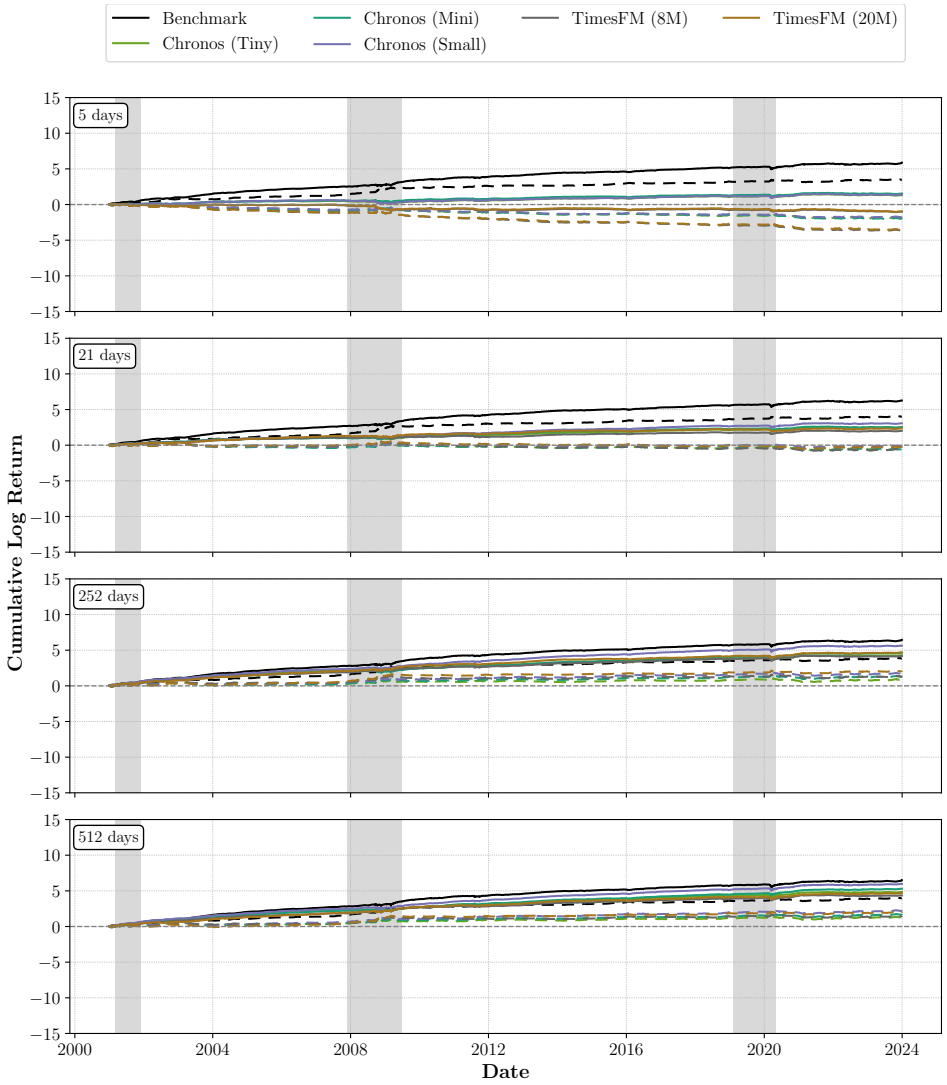
Table B.12: Pre-Trained TSFMs - Spread Portfolio Performance

Model	**Benchmark**				**Chronos (Tiny)**				**Chronos (Mini)**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512
Low	-31.91	-36.52	-35.00	-36.06	14.95	4.33	-8.55	-12.36	16.03	3.85	-13.08	-15.67
2	-0.94	-3.02	-3.94	-4.06	12.96	7.77	1.47	-1.21	12.77	8.05	0.29	-1.19
3	4.28	4.49	2.72	3.30	11.67	9.12	5.44	2.70	11.05	8.64	5.22	3.30
4	7.46	8.32	6.38	6.84	11.05	8.97	6.47	7.22	11.61	9.65	6.34	6.72
5	9.38	10.24	9.84	9.36	11.05	10.01	8.58	9.20	9.85	9.56	8.98	9.19
6	11.51	11.43	11.88	12.41	9.93	10.32	10.78	11.44	10.02	10.89	10.08	11.41
7	15.23	15.19	15.65	14.06	9.00	10.97	13.51	13.93	8.55	11.81	13.92	12.50
8	18.19	18.65	18.23	19.32	10.70	12.89	16.24	16.90	9.29	12.30	16.64	16.91
9	24.92	25.96	27.53	27.67	9.09	15.43	20.32	21.64	9.27	14.81	22.67	22.32
High	53.17	56.53	58.00	58.44	12.66	23.27	38.82	43.63	14.62	23.53	42.03	47.59
H-L	42.54	46.52	46.50	47.25	-1.14	9.47	23.69	28.00	-0.70	9.84	27.55	31.63

Model	**Chronos (Small)**				**TimesFM (8M)**				**TimesFM (20M)**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512
Low	14.08	0.17	-16.68	-20.16	30.12	3.68	-12.26	-13.02	29.43	1.21	-18.43	-17.91
2	12.07	6.62	-0.34	-2.30	20.77	8.39	0.17	-1.19	19.18	5.51	-2.65	-1.55
3	11.48	8.47	3.57	4.06	16.72	8.13	3.73	3.72	16.34	5.82	2.72	2.52
4	10.40	8.30	5.36	6.13	14.28	9.32	6.54	6.92	13.66	7.80	6.60	6.01
5	10.22	9.99	8.37	8.59	11.91	10.77	8.62	9.22	11.73	9.49	9.59	8.63
6	10.13	10.74	10.79	11.11	9.98	11.47	11.74	11.68	10.73	12.06	12.18	12.27
7	10.16	11.66	13.00	12.68	8.20	12.38	14.00	14.12	8.67	13.45	14.45	14.72
8	10.70	13.89	16.21	16.81	4.73	12.67	18.05	17.82	5.96	15.90	19.04	19.00
9	10.40	15.09	22.26	22.65	1.96	15.70	22.66	22.40	2.56	17.84	25.21	24.76
High	13.44	28.15	50.54	53.52	-7.42	18.76	38.02	39.60	-7.01	22.21	42.57	42.82
H-L	-0.32	13.99	33.61	36.84	-18.77	7.54	25.14	26.31	-18.22	10.50	30.50	30.36

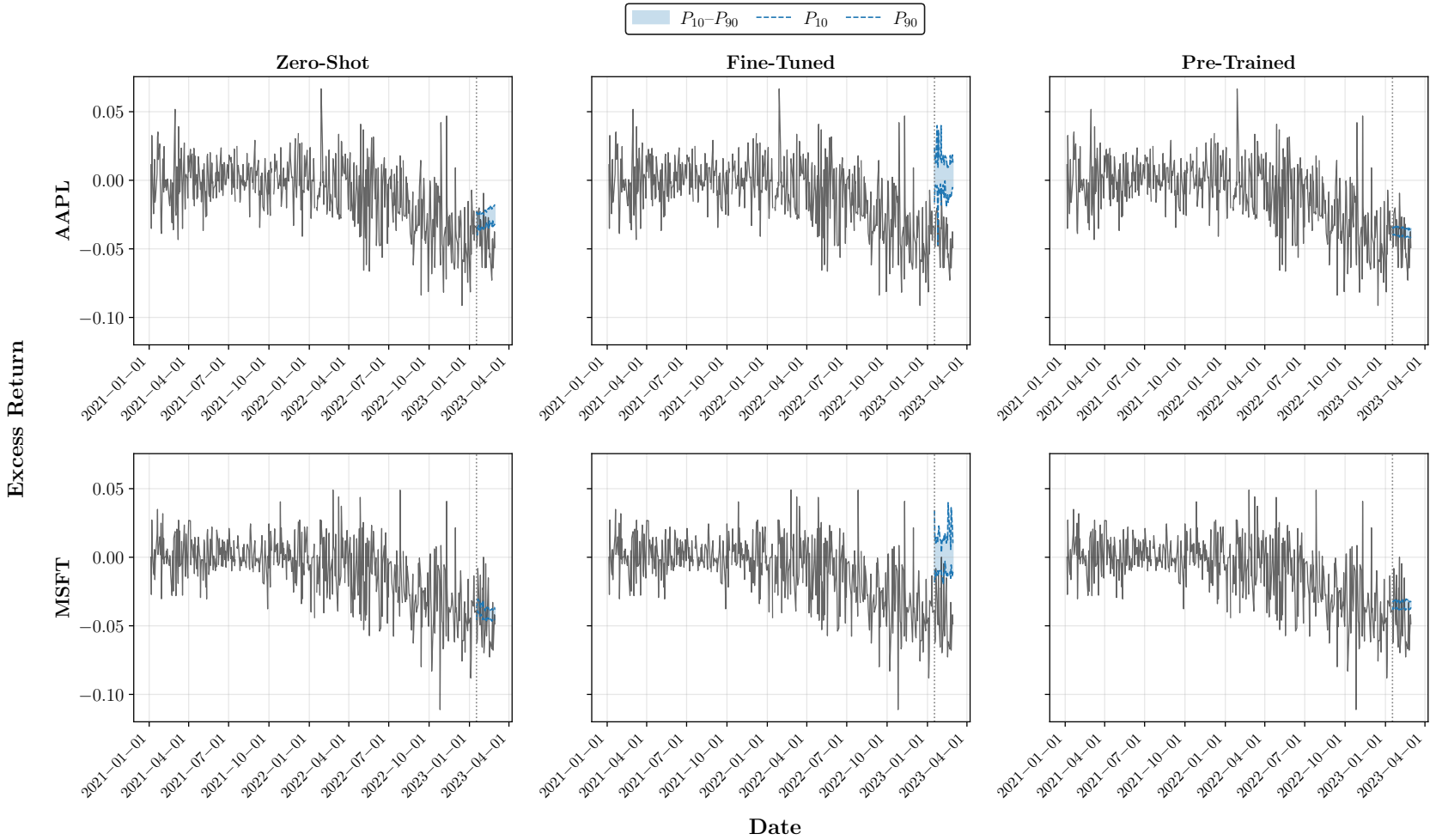
Note: This table presents the annualized average returns of decile spread portfolios formed using forecasts from various predictive models across different window sizes (5, 21, 252, and 512 trading days). The benchmark model is CatBoost, the best-performing model among the benchmarks. The time series foundation models (TSFMs) include Chronos (tiny, mini, and small) and TimesFM (with 8 million and 20 million parameters). Zero-shot inference is performed using the pre-trained models. For each model and window size, the returns of decile portfolios from the lowest (Low) to the highest (High) forecasted return are reported. The final row (H-L) represents the return spread between the highest and lowest deciles. Portfolios are constructed using equal-weighted decile sorting based on predicted returns.

Figure B.3: Cumulative Log Returns of Pre-Trained TSFMs: Long and Short Portfolios



Note: This figure displays the cumulative log returns of long and short portfolios, separately constructed using various forecasting models over rolling windows of 5, 21, 252, and 512 trading days. The benchmark model is CatBoost, the best-performing model among the benchmarks. The time series foundation models (TSFMs) include Chronos (tiny, mini, and small) and TimesFM (with 8 million and 20 million parameters). Zero-shot inference is performed using the pre-trained models. Each subplot corresponds to a specific horizon, as indicated by the text labels in the upper-left corners. For each model, the solid line denotes the performance of the long portfolio, while the dashed line represents the corresponding short portfolio. The benchmark model (CatBoost) is highlighted in black with bold lines. Shaded areas indicate U.S. recession periods, as defined by the National Bureau of Economic Research (NBER). All portfolios are equally weighted.

Figure B.4: Chronos TSFM 30-Day-Ahead Forecasts of Selected U.S. Equity Excess Returns



Note: This figure presents 30-day-ahead interval forecasts (P_{10}–P_{90}) of daily excess returns for selected U.S. stocks (AAPL and MSFT), generated using Chronos TSFMs. The left column displays zero-shot forecasts from the publicly available Chronos model, the middle column shows forecasts based on a fine-tuned, publicly available Chronos model, and the right column reports forecasts from a Chronos model pre-trained on global excess returns. All models use the small configuration. The black lines represent realized excess returns, the shaded blue regions denote 80% predictive intervals, and the dashed blue lines indicate the 10^{th} and 90^{th} percentile boundaries. The vertical dotted line marks the forecast start date. The forecasting horizon spans 30 trading days, from September 19, 2025, to October 31, 2025. The estimation window comprises 512 trading days, covering the period from September 11, 2023, to September 18, 2025. Consistent with Section 5.1.2, the number of generated samples is fixed at 20.

Table B.13: Forecasting Performance of Neural Networks with Varying Number of Units

	U.S.															
Model	**NN (8 Units, NN-S)**				**NN (32 Units, NN-L)**				**NN (128 Units)**				**NN (512 Units)**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
R^2_{OOS}	-3.10	-3.35	-1.91	-2.35	-3.36	-2.31	-2.29	-2.58	-4.82	-3.02	-2.36	-2.12	-3.95	-3.39	-2.28	-2.17
	-4.48	-5.12	-2.54	-2.79	-5.13	-3.05	-2.82	-3.40	-8.59	-3.85	-2.94	-2.52	-6.14	-4.76	-2.66	-2.43
	-1.95	-2.08	-0.97	-1.38	-2.04	-1.39	-1.36	-1.60	-2.92	-1.89	-1.47	-1.24	-2.51	-2.19	-1.38	-1.32
Overall Acc.	51.23	51.23	51.27	51.34	50.94	51.12	51.08	51.16	50.57	50.83	51.00	51.10	51.01	50.65	51.10	51.24
	50.90	51.01	51.15	51.22	50.70	50.74	50.80	50.95	50.12	50.94	50.83	50.35	50.16	50.44	50.97	50.70
	52.52	52.44	52.31	52.46	51.96	52.48	52.33	52.39	51.45	51.17	51.78	52.54	52.55	51.19	51.80	52.61
Up Acc.	68.63	69.44	70.53	69.42	68.96	67.82	66.87	69.14	55.37	70.71	59.22	47.83	43.27	57.08	60.50	53.35
	71.26	73.96	76.27	74.88	72.05	72.97	71.06	74.39	55.15	71.39	59.18	46.27	42.27	56.81	60.66	52.86
	67.39	66.84	67.79	66.86	68.00	65.11	64.65	66.07	57.16	71.36	60.31	50.21	44.93	58.58	61.48	54.87
Down Acc.	33.97	33.15	32.25	33.36	33.14	34.60	35.59	33.37	46.48	31.27	42.85	54.34	58.73	44.58	41.59	48.96
	29.26	26.65	24.64	26.09	28.11	27.25	29.59	26.24	45.71	29.61	41.87	54.82	58.73	44.15	40.25	48.12
	38.91	39.27	38.28	39.22	37.42	41.01	41.24	39.87	46.91	32.74	44.03	54.56	59.24	44.88	42.91	50.29
F1	0.49	0.49	0.49	0.49	0.49	0.49	0.50	0.49	0.44	0.44	0.46	0.48	0.43	0.43	0.46	0.49
	0.47	0.46	0.46	0.46	0.47	0.47	0.48	0.47	0.43	0.43	0.44	0.46	0.41	0.41	0.45	0.47
	0.51	0.51	0.51	0.51	0.51	0.52	0.52	0.51	0.45	0.46	0.47	0.50	0.45	0.44	0.48	0.50

	Global															
Model	**NN (8 Units, NN-S)**				**NN (32 Units, NN-L)**				**NN (128 Units)**				**NN (512 Units)**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
R^2_{OOS}	-0.56	-0.66	-0.72	-1.10	-0.60	-0.71	-0.59	-0.64	-0.69	-0.67	-0.69	-0.74	-0.98	-0.69	-0.73	-0.65
	-0.62	-0.78	-1.02	-1.95	-0.65	-0.99	-0.64	-0.72	-0.78	-0.80	-0.73	-0.91	-1.24	-0.71	-0.82	-0.68
	-0.28	-0.34	-0.25	-0.61	-0.29	-0.28	-0.29	-0.33	-0.43	-0.32	-0.36	-0.40	-0.66	-0.40	-0.48	-0.33
Overall Acc.	50.47	50.51	50.02	50.50	50.59	50.26	50.32	50.26	50.49	49.99	50.11	49.61	50.83	50.41	49.92	49.99
	50.23	50.19	50.04	50.43	50.58	50.13	50.15	50.11	50.06	49.37	49.83	50.26	50.33	50.01	50.60	50.77
	50.85	51.05	50.07	50.86	50.85	50.62	50.52	50.58	51.20	50.72	50.52	48.70	51.70	51.09	49.02	49.00
Up Acc.	39.96	48.12	56.16	52.03	59.24	48.18	52.44	45.29	45.24	46.70	47.68	72.50	41.17	45.47	69.44	71.39
	39.42	46.71	56.59	51.78	58.28	47.18	51.54	44.69	44.64	45.65	46.89	72.56	39.85	44.26	68.74	71.11
	40.52	49.50	56.37	52.51	60.36	49.44	53.58	46.07	45.70	47.78	48.58	72.97	42.19	46.33	70.26	72.15
Down Acc.	60.63	52.75	44.62	48.51	41.41	52.54	48.16	55.17	55.12	53.86	52.85	27.75	59.57	55.20	30.97	29.03
	60.93	53.58	43.77	48.36	41.99	53.05	48.60	55.41	55.48	54.49	53.30	27.54	60.51	55.91	31.44	29.05
	60.59	52.48	45.20	48.77	41.10	52.23	47.82	55.06	55.11	53.60	52.63	27.51	59.35	55.18	30.62	28.73
F1	0.36	0.38	0.38	0.37	0.37	0.37	0.37	0.37	0.35	0.36	0.36	0.35	0.37	0.37	0.36	0.36
	0.36	0.37	0.37	0.37	0.36	0.36	0.36	0.36	0.35	0.35	0.35	0.35	0.36	0.36	0.35	0.36
	0.37	0.39	0.38	0.38	0.38	0.38	0.38	0.37	0.36	0.37	0.37	0.35	0.38	0.38	0.35	0.36

Note: This table presents each metric as a set of three values, ordered from top to bottom: full sample, top 25% of firms by market capitalization (large-cap), and bottom 25% (small-cap) for various predictive models across different window sizes (5, 21, 252, and 512 trading days). Metrics are first computed separately for each calendar year using all stock-date observations within that year. The reported values represent the average of these yearly statistics. Metrics include out-of-sample R^2 (R^2_{OOS}), overall directional accuracy, upward and downward classification accuracy, and macro-averaged F1 score. The top panel reports the results for models trained exclusively on U.S. data, while the bottom panel presents the results when global data are used for training. All neural network models employ the same architecture, differing only in the number of units, which are set to 8, 32, 128, and 512, respectively. The results for the neural network with 8 units are denoted as NN-S, while those for the network with 32 units are denoted as NN-L, as also reported in Section 5 and Section 5.2. 'Overall Acc.' denotes overall directional accuracy, 'Up Acc.' and 'Down Acc.' represent the model's accuracy in predicting upward and downward excess returns respectively, and 'F1' refers to the macro-averaged F1 score.

Table B.14: Portfolio Performance of Neural Networks with Varying Number of Units

	U.S.															
Model	**NN (8 Units, NN-S)**				**NN (32 Units, NN-L)**				**NN (128 Units)**				**NN (512 Units)**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
Annualized Return	40.59	42.62	40.15	40.11	44.04	42.97	40.84	38.82	44.39	43.72	39.35	39.28	43.90	43.44	40.05	40.12
	27.33	27.45	25.45	25.53	28.37	27.66	26.12	24.87	28.62	28.16	25.40	24.96	28.43	27.95	25.76	25.33
	13.26	15.17	14.70	14.58	15.67	15.31	14.73	13.95	15.78	15.56	13.95	14.32	15.47	15.49	14.28	14.79
Standard Deviation	9.29	9.19	8.97	8.93	9.22	9.29	8.84	8.87	9.33	9.18	9.01	8.96	9.32	9.14	8.88	8.73
	14.00	13.85	13.62	13.71	13.84	13.83	13.70	13.57	13.93	13.79	13.69	13.65	13.92	13.85	13.67	13.59
	11.76	11.96	11.91	11.84	11.91	11.89	11.84	11.90	11.92	11.91	11.90	11.89	11.89	11.87	11.94	11.87
Sharpe Ratio	4.37	4.64	4.48	4.49	4.78	4.63	4.62	4.38	4.76	4.76	4.37	4.38	4.71	4.75	4.51	4.60
	1.95	1.98	1.87	1.86	2.05	2.00	1.91	1.83	2.05	2.04	1.86	1.83	2.04	2.02	1.88	1.86
	1.13	1.27	1.23	1.23	1.32	1.29	1.24	1.17	1.32	1.31	1.17	1.20	1.30	1.30	1.20	1.25
Daily Return (bps)	16.11	16.91	15.93	15.92	17.48	17.05	16.21	15.41	17.62	17.35	15.61	15.59	17.42	17.24	15.89	15.92
	10.84	10.89	10.10	10.13	11.26	10.97	10.36	9.87	11.36	11.17	10.08	9.90	11.28	11.09	10.22	10.05
	5.26	6.02	5.83	5.78	6.22	6.08	5.84	5.54	6.26	6.17	5.53	5.68	6.14	6.15	5.67	5.87
Max DD	18.91	18.50	16.58	16.98	17.01	18.53	14.48	17.29	18.03	17.55	18.87	14.94	17.55	18.38	16.32	14.20
	33.36	33.38	32.27	32.70	32.39	33.23	31.99	31.66	33.45	32.70	32.47	31.83	32.76	33.02	32.11	31.87
	30.49	30.76	31.78	31.25	30.53	29.63	30.79	29.22	28.81	28.71	32.23	28.75	30.35	29.89	31.47	29.26
Max DD (1-day)	5.62	5.47	4.80	5.15	5.51	5.61	5.49	5.29	5.65	5.43	5.33	5.22	5.75	5.78	5.53	4.88
	8.81	8.54	8.21	8.37	8.67	8.80	8.67	8.32	8.87	8.56	8.57	8.44	8.87	8.79	8.60	8.26
	6.31	5.53	4.98	5.28	5.58	5.63	5.18	5.19	5.88	5.71	5.24	5.44	4.97	5.27	5.23	5.40
Skew	1.31	1.31	1.20	1.67	1.24	1.37	1.27	1.36	1.31	1.35	1.45	1.40	1.32	1.28	1.23	1.43
	-0.12	-0.15	-0.09	-0.00	-0.11	-0.11	-0.14	-0.13	-0.15	-0.11	-0.05	-0.07	-0.12	-0.09	-0.16	-0.10
	0.48	0.51	0.51	0.52	0.52	0.53	0.45	0.49	0.52	0.53	0.56	0.46	0.59	0.52	0.50	0.53
Kurt	17.34	16.69	15.43	20.70	16.54	17.36	17.32	17.86	17.53	16.92	18.78	17.59	17.08	16.98	16.95	17.62
	11.96	11.61	11.69	12.29	11.45	11.68	11.98	11.78	12.02	11.35	11.97	11.66	11.77	11.72	11.84	11.55
	7.07	6.62	6.16	6.23	6.76	6.69	6.04	6.51	6.96	6.57	7.04	6.30	6.99	6.37	6.33	6.51

	Global															
Model	**NN (8 Units, NN-S)**				**NN (32 Units, NN-L)**				**NN (128 Units)**				**NN (512 Units)**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
Annualized Return	21.57	16.94	25.37	15.88	20.14	23.72	17.04	17.10	21.70	19.66	17.02	12.68	16.59	16.56	15.96	17.69
	19.07	16.05	19.20	15.15	18.41	20.10	16.46	15.76	17.99	17.57	16.29	13.29	17.46	17.28	15.68	16.84
	2.50	0.90	6.17	0.73	1.73	3.61	0.59	1.34	3.71	2.09	0.73	-0.60	-0.87	-0.72	0.29	0.85
Standard Deviation	7.59	7.43	7.44	6.88	8.12	8.21	7.05	6.95	8.47	7.82	7.23	6.39	8.16	7.72	6.94	6.77
	13.31	12.99	13.08	12.97	13.63	13.75	13.22	13.40	13.68	13.50	13.72	12.93	13.64	13.48	13.49	13.34
	11.25	12.01	11.94	11.69	11.48	11.58	11.78	11.48	12.06	12.03	11.25	11.59	12.21	11.68	11.62	11.58
Sharpe Ratio	2.84	2.28	3.41	2.31	2.48	2.89	2.42	2.46	2.56	2.51	2.35	1.98	2.03	2.14	2.30	2.61
	1.43	1.23	1.47	1.17	1.35	1.46	1.24	1.18	1.31	1.30	1.19	1.03	1.28	1.28	1.16	1.26
	0.22	0.07	0.52	0.06	0.15	0.31	0.05	0.12	0.31	0.17	0.06	-0.05	-0.07	-0.06	0.02	0.07
Daily Return (bps)	8.56	6.72	10.07	6.30	7.99	9.41	6.76	6.79	8.61	7.80	6.75	5.03	6.58	6.57	6.33	7.02
	7.57	6.37	7.62	6.01	7.31	7.98	6.53	6.25	7.14	6.97	6.46	5.27	6.93	6.86	6.22	6.68
	0.99	0.36	2.45	0.29	0.69	1.43	0.23	0.53	1.47	0.83	0.29	-0.24	-0.35	-0.28	0.11	0.34
Max DD	27.42	17.97	14.73	26.02	22.42	29.94	14.40	19.76	23.68	25.25	8.75	16.93	24.75	28.36	10.80	10.13
	29.80	36.48	28.29	32.36	33.74	32.88	29.05	30.31	36.97	29.90	28.89	26.89	29.68	30.67	29.89	33.29
	42.94	55.93	42.65	71.51	58.88	48.76	68.33	60.67	34.75	63.73	60.92	69.59	73.80	65.45	64.83	64.19
Max DD (1-day)	3.68	4.48	3.16	6.12	4.74	4.97	2.97	3.70	5.06	3.93	3.27	2.46	3.53	5.03	6.51	3.57
	7.97	8.26	6.83	7.38	8.51	8.36	7.09	8.19	8.70	8.55	7.86	6.37	8.40	8.46	7.62	8.04
	5.01	5.82	4.33	5.62	6.33	5.58	4.80	6.09	6.44	6.19	5.36	5.12	6.88	5.11	6.22	5.06
Skew	1.45	1.03	1.95	1.00	1.73	1.37	2.04	1.95	0.97	1.55	1.95	1.60	1.20	1.25	0.85	1.88
	0.00	-0.20	0.09	-0.05	-0.01	0.03	0.12	-0.04	-0.16	0.02	0.01	0.08	0.01	-0.07	-0.02	0.10
	0.36	0.25	0.63	0.18	0.48	0.35	0.59	0.40	0.03	0.51	0.43	0.41	0.16	0.22	0.20	0.40
Kurt	17.75	21.33	19.89	25.22	21.19	17.71	21.38	24.62	16.64	18.19	22.05	16.84	14.53	20.44	23.62	20.74
	11.91	10.84	7.82	9.58	11.62	10.56	10.24	11.24	11.81	11.57	9.91	8.77	10.60	11.12	10.64	11.47
	6.39	7.81	6.60	6.94	7.06	6.00	7.95	7.28	7.27	8.36	6.78	6.81	9.58	6.06	6.52	6.27

Note: This table reports average yearly portfolio performance metrics across different rolling window sizes (5, 21, 252, and 512 trading days) for each model. Each cell displays three values from top to bottom: long–short portfolio, long-only leg, and short-only leg. Metrics include annualized return, standard deviation, Sharpe ratio, daily return (in basis points), maximum drawdown (Max DD), one-day maximum drawdown (Max DD (1-day)), skewness, and kurtosis of portfolio returns. The top panel reports the results for models trained exclusively on U.S. data, while the bottom panel presents the results when global data are used for training. All neural network models employ the same architecture, differing only in the number of units, which are set to 8, 32, 128, and 512, respectively. The results for the neural network with 8 units are denoted as NN-S, while those for the network with 32 units are denoted as NN-L, as also reported in Section 5 and Section 5.2. Portfolios are formed using decile sorting based on model forecasts, with equal weighting across stocks.

Table B.15: Spread Portfolio Performance of Neural Networks with Varying Number of Units

	U.S.															
Model	**NN (8 Units, NN-S)**				**NN (32 Units, NN-L)**				**NN (128 Units)**				**NN (512 Units)**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
Low	-26.53	-30.34	-29.40	-29.15	-31.34	-30.62	-29.45	-27.90	-31.55	-31.12	-27.89	-28.64	-30.94	-30.97	-28.57	-29.58
2	-7.09	-7.52	-5.35	-5.20	-8.64	-8.33	-5.89	-4.93	-8.72	-7.73	-5.83	-4.86	-8.70	-8.25	-5.13	-4.90
3	0.38	0.60	1.39	1.88	-8.64	-8.33	-5.89	-4.93	-0.68	-0.44	1.50	1.45	-1.23	0.18	1.25	1.48
4	4.12	5.11	6.07	6.32	4.26	4.04	5.11	6.73	4.30	4.85	6.50	5.72	4.52	4.83	5.75	6.32
5	7.92	9.16	10.30	9.95	8.59	8.58	8.78	8.90	8.56	7.86	9.48	9.06	8.42	8.46	8.86	9.71
6	11.44	12.51	12.66	12.48	12.24	12.62	13.11	12.90	12.24	12.14	12.01	13.83	12.04	11.75	12.59	12.75
7	15.27	15.81	16.06	15.13	16.31	16.60	16.48	15.64	16.38	16.65	16.20	16.10	16.65	16.71	16.62	16.64
8	20.97	21.53	20.45	19.99	22.54	21.79	20.82	20.12	21.46	21.71	20.76	21.32	22.77	22.16	20.34	19.97
9	30.14	29.54	28.20	28.85	30.60	31.69	28.64	27.41	32.06	31.05	27.74	27.38	30.90	30.52	28.04	28.25
High	54.65	54.90	50.90	51.06	56.74	55.31	52.23	49.75	57.24	56.32	50.80	49.92	56.87	55.90	51.53	50.65
H-L	40.59	42.62	40.15	40.11	44.04	42.97	40.84	38.82	44.39	43.72	39.35	39.28	43.90	43.44	40.05	40.12
	Global															
Model	**NN (8 Units, NN-S)**				**NN (32 Units, NN-L)**				**NN (128 Units)**				**NN (512 Units)**			
Window Size	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
Low	-4.99	-1.80	-12.34	-1.45	-3.46	-7.23	-1.18	-2.68	-7.43	-4.19	-1.46	1.21	1.75	1.44	-0.57	-1.70
2	4.92	5.83	0.08	5.15	4.78	2.86	4.69	4.82	4.01	4.70	4.98	7.51	6.01	6.02	5.45	4.04
3	6.84	7.72	5.29	7.57	6.20	5.90	6.70	7.13	6.62	7.03	7.97	8.06	7.70	8.59	7.00	6.76
4	7.80	8.17	8.19	8.47	7.28	7.14	8.49	8.42	7.71	7.98	8.93	9.68	7.77	8.79	8.01	8.07
5	7.67	9.41	10.36	9.79	7.89	8.75	8.78	9.21	9.14	8.81	9.38	11.41	8.37	8.77	9.04	10.08
6	8.74	9.62	12.36	11.38	10.01	9.60	10.65	10.35	11.64	9.70	10.23	11.05	9.82	8.71	10.48	10.59
7	12.11	10.43	13.09	12.00	10.18	12.00	11.68	11.57	12.63	11.34	11.65	11.19	10.08	10.18	11.86	10.91
8	12.93	12.37	16.82	12.67	14.09	14.32	13.14	13.77	13.48	13.07	11.78	11.80	11.68	10.59	12.63	12.84
9	17.10	17.42	19.01	15.39	17.46	17.71	15.38	17.15	17.50	17.68	15.23	12.77	13.14	13.60	16.01	15.99
High	38.14	32.09	38.40	30.31	36.83	40.20	32.91	31.52	35.98	35.14	32.57	26.57	34.92	34.55	31.35	33.68
H-L	21.57	16.94	25.37	15.88	20.14	23.72	17.04	17.10	21.70	19.66	17.02	12.68	16.59	16.56	15.96	17.69

Note: This table presents the annualized average returns of decile spread portfolios formed using forecasts from various predictive models across different window sizes (5, 21, 252, and 512 trading days). For each model and window size, the returns of decile portfolios from the lowest (Low) to the highest (High) forecasted return are reported. The final row (H-L) represents the return spread between the highest and lowest deciles. The top panel reports the results for models trained exclusively on U.S. data, while the bottom panel presents the results when global data are used for training. All neural network models employ the same architecture, differing only in the number of units, which are set to 8, 32, 128, and 512, respectively. The results for the neural network with 8 units are denoted as NN-S, while those for the network with 32 units are denoted as NN-L, as also reported in Section 5 and Section 5.2. Portfolios are formed using decile sorting based on model forecasts, with equal weighting across stocks. Portfolios are constructed using equal-weighted decile sorting based on predicted returns.

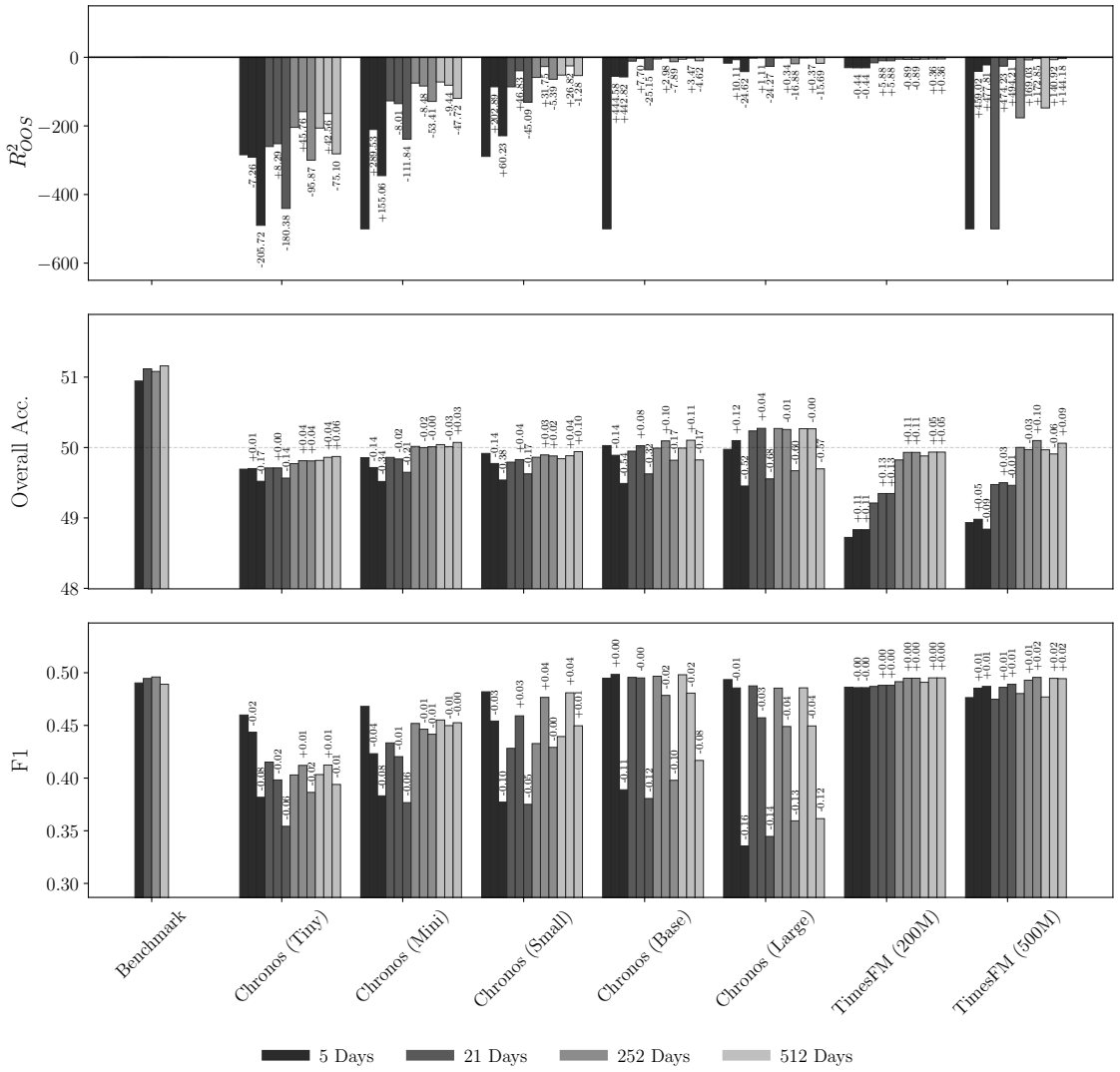
Table B.16: Comparison of Daily Out-of-Sample Forecasting Performance of Augmented Pre-Trained TSFMs Using Diebold-Mariano Test

		Chronos (Tiny)				Chronos (Mini)				Chronos (Small)				TimesFM (8M)				TimesFM (20M)			
		5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
Benchmark	5	-17.06*	-5.83*	-4.14*	-4.15*	-17.78*	-12.99*	-4.14*	-4.50*	-17.23*	-14.03*	-4.58*	-4.64*	-32.56*	-22.34*	-11.23*	-9.38*	-33.27*	-21.44*	-9.59*	-7.79*
	21	-18.20*	-6.05*	-6.17*	-6.10*	-19.20*	-13.89*	-5.99*	-6.35*	-17.99*	-15.04*	-6.36*	-6.29*	-32.94*	-22.51*	-11.73*	-9.61*	-34.06*	-22.13*	-13.18*	-10.81*
	252	-18.51*	-6.06*	-6.49*	-6.50*	-19.55*	-14.04*	-6.30*	-6.73*	-18.16*	-15.18*	-6.71*	-6.68*	-32.90*	-22.17*	-11.69*	-9.61*	-34.23*	-21.85*	-13.47*	-11.08*
	512	-18.33*	-6.06*	-6.24*	-6.20*	-19.25*	-13.92*	-6.03*	-6.43*	-18.06*	-15.08*	-6.43*	-6.41*	-32.98*	-22.44*	-11.73*	-9.62*	-34.12*	-22.16*	-13.05*	-10.78*
Chronos (Tiny)	5					-0.01	2.52*	16.42*	16.18*	-7.85*	2.40*	15.94*	15.80*	-29.84*	-8.18*	2.48*	-2.06*	-33.28*	-3.78*	15.06*	15.26*
	21					-0.28	0.66	5.20*	5.14*	-3.54*	0.55	5.10*	5.08*	-15.34*	-4.43*	1.03	-1.62	-14.55*	-1.90*	4.80*	5.00*
	252					-19.04*	-12.82*	0.50	-0.33	-17.45*	-13.81*	-1.05	-0.89	-32.77*	-19.64*	-9.34*	-8.50*	-34.93*	-18.65*	-2.96*	-1.18
	512					-18.49*	-12.50*	0.34	-0.77	-17.27*	-13.41*	-1.42	-1.65	-32.33*	-19.06*	-9.40*	-8.55*	-34.49*	-17.66*	-3.19*	-1.38
Chronos (Mini)	5									-7.06*	2.51*	17.40*	17.12*	-30.54*	-8.27*	2.60*	-2.06*	-34.38*	-3.88*	16.15*	16.14*
	21									-8.07*	-0.47	12.17*	11.85*	-27.31*	-9.70*	0.91	-2.90*	-29.75*	-5.68*	10.86*	11.15*
	252									-16.92*	-13.25*	-2.78*	-2.01*	-31.79*	-19.13*	-9.40*	-8.55*	-33.86*	-17.63*	-3.20*	-1.44
	512									-16.82*	-12.85*	-0.63	-1.20	-31.61*	-18.59*	-9.27*	-8.50*	-33.71*	-16.98*	-2.80*	-1.09*
Chronos (Small)	5													-23.59*	-2.97*	6.07*	0.72	-24.00*	1.61	15.78*	15.91*
	21													-27.80*	-9.70*	1.07	-2.84*	-30.83*	-5.77*	11.76*	12.03*
	252													-31.47*	-18.85*	-56.82*	-8.41*	-33.52*	-17.20*	-2.20*	-0.60
	512													-31.59*	-18.50*	-9.08*	-8.42*	-33.47*	-16.61*	-2.19*	-0.59
TimesFM (8M)	5																	4.21*	23.41*	31.51*	31.25*
	21																	-15.70*	7.49*	19.10*	19.27*
	252																	-22.12*	-5.47*	8.61*	9.11*
	512																	-12.68*	0.23	8.09*	8.37*
TimesFM (20M)	5																				
	21																				
	252																				
	512																				
Chronos (Tiny)*	5																				
	21																				
	252																				
	512																				
Chronos (Mini)*	5																				
	21																				
	252																				
	512																				
Chronos (Small)*	5																				
	21																				
	252																				
	512																				
TimesFM (8M)*	5																				
	21																				
	252																				
	512																				

		Chronos (Tiny)*				Chronos (Mini)*				Chronos (Small)*				TimesFM (8M)*				TimesFM (20M)*			
		5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
Benchmark	5	-35.73*	-15.87*	-3.67*	-3.67*	-30.93*	-16.65*	-3.02*	-3.23*	-38.59*	-19.40*	-3.63*	-3.51*	-28.10*	-17.48*	-12.89*	-12.43*	-29.08*	-20.66*	-56.31*	-9.72*
	21	-36.18*	-18.79*	-5.37*	-4.93*	-31.14*	-19.28*	-4.92*	-5.03*	-38.91*	-22.23*	-5.33*	-5.11*	-28.39*	-17.90*	-15.87*	-14.50*	-29.72*	-57.64*	-14.68*	-11.91*
	252	-36.29*	-19.25*	-5.63*	-5.13*	-31.20*	-19.73*	-5.23*	-5.40*	-39.08*	-22.72*	-5.66*	-5.45*	-28.44*	-17.86*	-15.65*	-14.21*	-29.79*	-21.91*	-14.72*	-11.93*
	512	-36.19*	-18.80*	-5.44*	-4.96*	-31.13*	-19.37*	-4.97*	-5.12*	-38.96*	-22.29*	-5.39*	-5.20*	-28.41*	-17.98*	-15.45*	-14.19*	-29.75*	-22.06*	-14.51*	-11.83*
Chronos (Tiny)	5	-30.90*	-31.71*	17.24*	15.65*	-27.12*	8.97*	17.63*	17.01*	-36.96*	6.15*	17.01*	16.68*	-24.72*	-5.75*	12.87*	12.74*	-27.72*	-0.09	13.19*	13.50*
	21	-15.72*	2.42*	5.29*	5.11*	-17.33*	2.10*	5.43*	5.37*	-19.70*	1.40	5.27*	5.26*	-18.02*	-3.24*	4.42*	4.49*	-13.56*	-0.26	4.56*	4.73*
	252	-36.00*	-17.70*	0.76	-0.84	-30.81*	-18.09*	4.61*	2.02*	54.10*	-21.30*	0.22	0.53	-28.26*	-16.00*	-5.11*	-3.93*	-31.33*	-18.51*	-3.81*	-2.34*
	512	-35.43*	-15.44*	0.43	-1.05	-30.65*	-15.93*	3.52*	3.36*	-38.64*	-19.01*	-53.21*	0.56*	-28.22*	-15.37*	-5.46*	-4.37*	-30.59*	-17.82*	-4.21*	-2.69*
Chronos (Mini)	5	-32.64*	11.35*	19.34*	17.45*	54.89*	10.14*	19.65*	18.67*	-37.04*	7.00*	19.19*	18.54*	-24.78*	-5.77*	13.78*	13.42*	-28.84*	-0.03	14.01*	14.18*
	21	-30.13*	5.71*	13.09*	12.00*	-26.45*	4.77*	13.41*	12.87*	-34.47*	2.46*	13.20*	12.65*	-24.77*	-7.53*	9.55*	9.48*	-25.22*	-2.28*	9.77*	10.05*
	252	-34.96*	-14.81*	0.31	-1.27	-30.45*	-15.30*	2.95*	2.44*	-38.27*	-18.33*	-0.15	0.30	-28.05*	-15.42*	-5.64*	-4.53*	-29.97*	-17.27*	-4.32*	-2.78*
	512	-34.65*	-13.75*	0.77	-0.54	-30.40*	-14.29*	3.04*	4.16*	-38.26*	-17.30*	0.53	1.33	-28.00*	-15.04*	-5.31*	-4.36*	-29.59*	-16.83*	-4.12*	-2.59*
Chronos (Small)	5	-24.53*	12.42*	17.66*	16.87*	-22.37*	11.77*	17.88*	17.52*	-29.64*	10.17*	17.55*	17.38*	-21.40*	-0.97	14.49*	14.34*	-21.81*	5.58*	14.73*	14.88*
	21	-29.79*	6.43*	13.86*	12.47*	-53.39*	5.68*	14.34*	13.68*	-34.68*	3.02*	13.84*	13.30*	-24.94*	-7.64*	10.30*	10.22*	-25.71*	-2.18*	10.52*	10.79*
	252	-34.66*	-55.41*	1.65	0.05	-30.30*	-14.44*	4.69*	4.20*	-37.99*	-17.45*	1.58	2.39*	-27.87*	-15.16*	-4.53*	-3.57*	-29.84*	-16.61*	-3.37*	-1.96*
	512	-34.56*	-12.89*	1.33	0.10	-30.25*	-13.51*	3.58*	4.98*	-38.12*	-16.42*	1.23	2.71*	-27.98*	-14.76*	-4.43*	-3.62*	-29.54*	-16.49*	-3.48*	-2.07*
TimesFM (8M)	5	0.05	31.48*	32.94*	31.56*	-5.66*	31.53*	32.61*	32.03*	-8.66*	30.49*	32.39*	32.09*	-9.06*	19.87*	29.99*	29.62*	5.84*	28.83*	30.08*	30.03*
	21	-19.88*	15.39*	20.23*	18.47*	-18.73*	14.70*	19.65*	19.06*	-23.32*	13.23*	19.39*	18.84*	-19.03*	2.75*	17.91*	17.91*	-14.02*	10.94*	18.16*	18.38*
	252	-23.86*	2.70*	9.33*	8.80*	-23.14*	2.03*	9.76*	9.69*	-27.52*	0.56	9.22*	9.34*	-22.42*	-6.79*	7.86*	8.07*	-19.70*	-2.66*	8.13*	8.62*
	512	-14.13*	5.13*	8.54*	8.34*	-15.75*	4.77*	8.73*	8.71*	-17.73*	3.99*	8.50*	8.56*	-16.57*	-1.31	7.69*	7.83*	-11.51*	2.10*	7.86*	8.13*
TimesFM (20M)	5	-3.67*	34.29*	34.88*	33.52*	-8.57*	34.72*	34.93*	34.34*	-13.17*	33.11*	34.92*	34.38*	-10.19*	19.21*	30.92*	30.45*	1.61	29.59*	31.10*	30.92*
	21	-26.33*	13.26*	18.84*	16.23*	-23.27*	12.39*	18.61*	17.66*	-29.22*	10.09*	18.09*	17.15*	-21.35*	-4.39*	16.64*	16.19*	-19.83*	6.33*	16.70*	16.61*
	252	-34.50*	-12.33*	3.03*	1.81	-29.91*	-13.04*	4.50*	4.39*	-37.56*	-16.01*	2.84*	3.20*	-27.52*	-14.97*	-3.53*	-2.33*	-28.69*	-17.69*	-2.21*	-0.12
	512	-34.07*	-12.34*	1.40	0.51	-29.92*	-13.02*	2.54*	2.52*	-37.41*	-15.79*	1.23	1.55	-27.52*	-15.00*	-5.36*	-4.78*	-28.23*	-17.60*	-4.51*	-2.68*
Chronos (Tiny)*	5	-6.53*	35.41*	36.20*	35.45*	-11.11*	34.10*	36.67*	35.66*	-7.85*	23.85*	32.95*	32.34*	4.55*	30.83*	33.01*	32.75*				
	21	-29.13*	-4.11*	17.76*	15.25*	-38.02*	-11.69*	16.57*	14.52*	-26.48*	-12.42*	9.20*	8.79*	-29.57*	-11.66*	9.48*	9.69*				
	252	-31.12*	-17.83*	2.05*	1.11	-39.53*	-21.21*	-0.59	0.01	-28.23*	-16.59*	-4.94*	-3.88*	-31.45*	-18.55*	-3.80*	-2.46*				
	512	-30.46*	-13.08*	2.93*	2.67*	-38.48*	-16.02*	1.37	1.99*	-27.89*	-14.85*	-3.64*	-2.93*	-30.29*	-15.71*	-2.75*	-1.64				
Chronos (Mini)*	5					-1.85	28.62*	30.89*	30.63*	-3.23*	20.96*	29.20*	28.93*	-29.60*	-10.11*	10.34*	10.44*				
	21					-38.78*	-9.78*	16.81*	14.91*	-26.29*	-11.91*	10.04*	9.58*	-29.60*	-10.11*	10.34*	10.44*				
	252					-39.01*	-21.23*	-3.09*	-2.10*	-28.36*	-16.20*	-6.50*	-5.18*	-31.04*	-18.56*	-5.20*	-3.55*				
	512					-38.71*	-18.65*	-1.95	-2.61*	-28.24*	-15.53*	-6.51*	-5.43*	-30.13*	-17.77*	-5.41*	-3.77*				
Chronos (Small)*	5									-2.87*	26.26*	36.19*	35.78*	12.62*	34.08*	34.08*	36.38*				
	21									-25.59*	-10.32*	12.83*	12.17*	-28.60*	-7.06*	-7.06*	13.05*				
	252									-28.25*	-16.09*	-4.83*	-3.82*	-30.84*	-17.89*	-17.89*	-2.37*				
	512									-28.21*	-15.28*	-5.21*	-4.30*	-30.25*	-17.24*	-17.24*	-2.81*				
TimesFM (8M)*	5													10.62*	24.08*	26.88*	26.95*				
	21													-16.34*	7.28*	14.33*	14.26*				
	252													-27.31*	-15.09*	1.77	3.68*				
	512													-26.67*	-14.70*	0.56	3.37*				

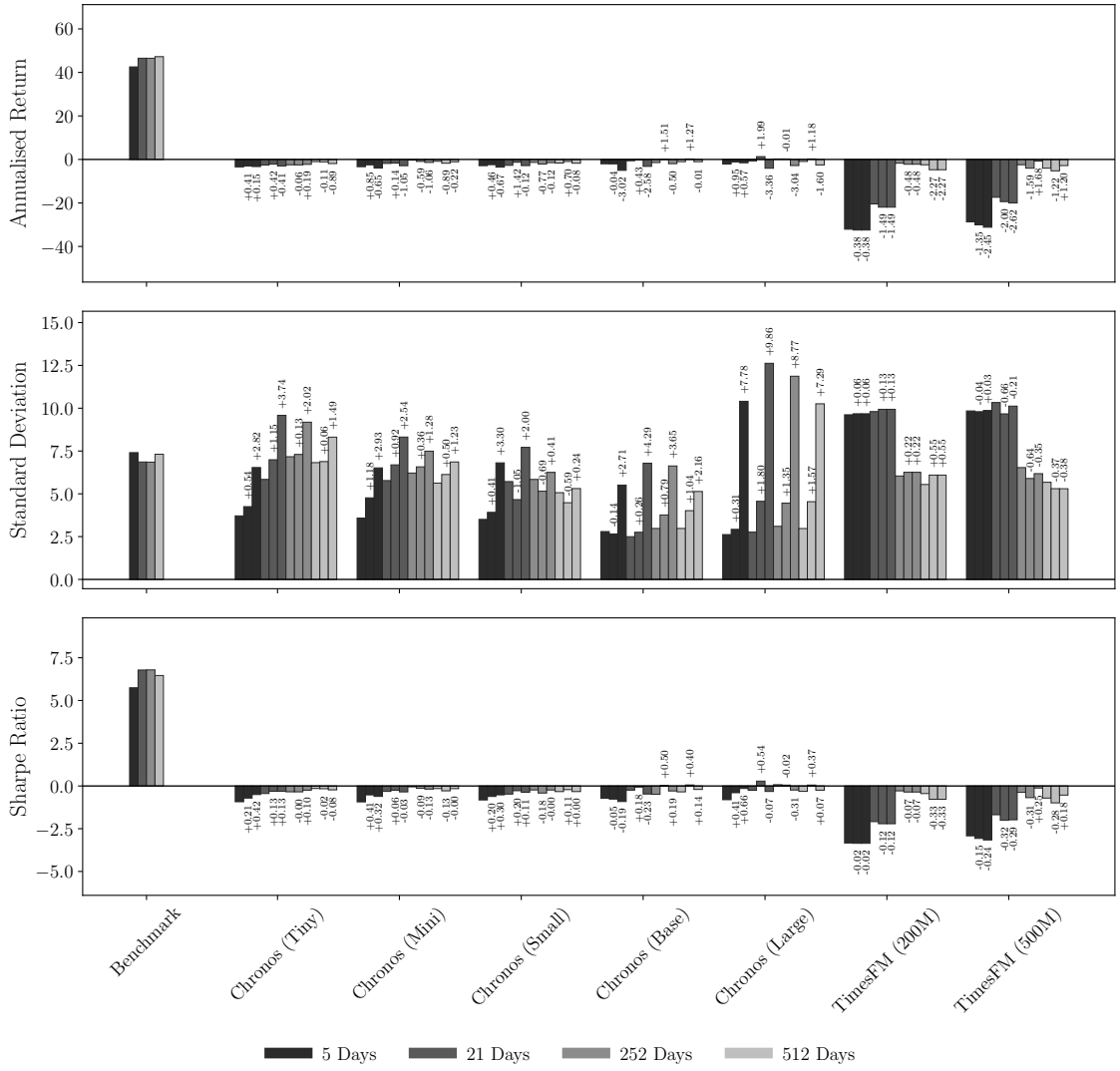
Note: This table reports the modified Diebold–Mariano (DM) test statistics, following the approach of Gu et al. (2020), for out-of-sample stock-level forecast comparisons across pre-trained time series foundation models (TSFMs). A positive statistic indicates that the model in the column achieves superior predictive performance relative to the model in the corresponding row. An asterisk denotes statistical significance at the 5% level. The benchmark model is CatBoost, the best-performing model among the benchmarks. TSFMs include Chronos (tiny, mini, and small) and TimesFM (with 8 million and 20 million parameters). The models are pre-trained from scratch and evaluated under two variants: (i) models pre-trained on global data augmented with JKP factors, and (ii) models pre-trained on global data augmented with synthetic data of equivalent size to the JKP factors, with the latter group indicated by an asterisk. JKP factors used in the augmented data are defined in Jensen et al. (2023), and the synthetic data is generated following Ansari et al. (2024). Tests for identical models across different window sizes have been omitted from this table for clarity.

Figure B.5: Impact of Global and Global-Augmented Fine-Tuning on Forecasting Performance



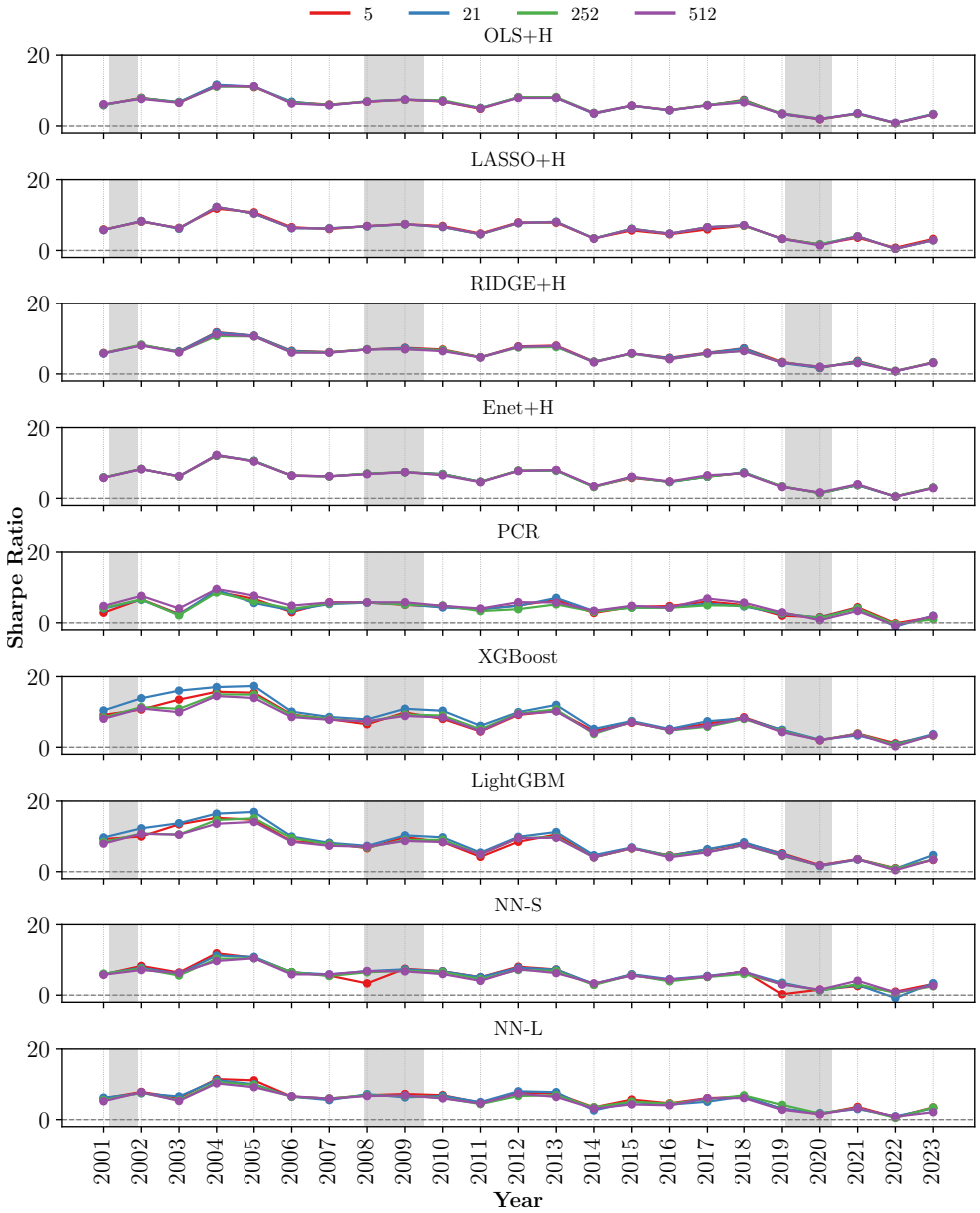
Note: This figure compares forecasting performance metrics (R^2_{OOS}, overall accuracy, and F1) for seven models under three fine-tuning regimes: U.S.-only (left bar in each triplet), global (middle bar), and global-augmented (right bar). Bars of the same color correspond to the same rolling training window size (5, 21, 252, and 512 trading days). Each triplet of adjacent bars illustrates the performance change attributable to expanding the fine-tuning data from U.S.-only to Global and further to Global-Augmented. Models include Chronos (tiny, mini, small, base, and large) and TimesFM (version 1 with 200 million and version 2 with 500 million parameters). JKP factors used in the augmented data are defined in Jensen et al. (2023). The benchmark group presents the results of the CatBoost model, trained on U.S. data, evaluated over trading window sizes of 5, 21, 252, and 512 days, respectively. 'Overall Acc.' denotes overall directional accuracy, and 'F1' refers to the macro-averaged F1 score. In the middle panel, the horizontal line indicates the 50% overall accuracy. To enhance interpretability of the plots, R^2_{OOS} values were truncated at –500 in order to mitigate the influence of extreme negative outliers on the visual scale.

Figure B.6: Impact of Global and Global-Augmented Fine-Tuning on Portfolio Performance



Note: This figure compares portfolio performance metrics (annualized return, standard deviation, and Sharpe ratio) for seven models under three fine-training regimes: U.S.-only (left bar in each triplet), global (middle bar), and global-augmented (right bar). Bars of the same color correspond to the same rolling training window size (5, 21, 252, and 512 trading days). Each triplet of adjacent bars illustrates the performance change attributable to expanding the fine-tuning data from U.S.-only to Global and further to Global-Augmented. Models include Chronos (tiny, mini, small, base, and large) and TimesFM (version 1 with 200 million and version 2 with 500 million parameters). JKP factors used in the augmented data are defined in Jensen et al. (2023). The benchmark group presents the results of the CatBoost model, trained on U.S. data, evaluated over trading window sizes of 5, 21, 252, and 512 days, respectively.

Figure B.7: Yearly Sharpe Ratios of Long–Short Portfolios (Benchmarks)



Note: The figure reports annual Sharpe ratios from 2001 to 2023 for long–short equal-weighted portfolios. Results are presented by model across multiple window sizes (5, 21, 252, and 512 trading days). The results are presented for benchmark models comprising linear methods (OLS, Lasso, Ridge, Elastic Net, and PCR), ensemble models (XGBoost and LightGBM), and neural networks (NN-S and NN-L); CatBoost is excluded for clarity. All models are trained on U.S. excess returns. Shaded regions correspond to U.S. recession periods as identified by the National Bureau of Economic Research (NBER).

Table B.17: Long–Short Portfolio Sharpe Ratios with Transaction Costs (Benchmarks)

Model	0 bps				20 bps				40 bps				Mixed			
	5	21	252	512	5	21	252	512	5	21	252	512	5	21	252	512
U.S.																
OLS+H	4.86	4.89	4.92	4.87	-3.14	-3.06	-2.97	-2.97	-11.14	-11.01	-10.87	-10.81	-2.95	-2.87	-2.79	-2.78
LASSO+H	4.84	4.86	4.87	4.83	-3.35	-4.61	-4.52	-4.45	-11.55	-14.07	-13.91	-13.71	-3.16	-4.40	-4.32	-4.24
RIDGE+H	4.88	4.81	4.79	4.74	-3.49	-3.50	-3.30	-3.34	-11.87	-11.81	-11.39	-11.43	-3.30	-3.30	-3.11	-3.15
Enet+H	4.83	4.83	4.85	4.86	-3.73	-3.76	-4.00	-4.52	-12.30	-12.36	-12.85	-13.90	-3.54	-3.56	-3.80	-4.31
PCR	3.61	3.58	3.52	3.94	-7.00	-7.23	-7.45	-6.03	-17.60	-18.02	-18.39	-15.99	-6.79	-7.02	-7.25	-5.83
XGBoost	5.79	6.53	5.86	5.71	-3.44	-3.23	-2.58	-2.51	-12.61	-12.96	-10.99	-10.70	-3.24	-3.05	-2.43	-2.36
CatBoost	5.74	6.78	6.79	6.46	-3.94	-3.52	-3.90	-3.13	-13.53	-13.74	-14.46	-12.64	-3.73	-3.33	-3.69	-2.96
LightGBM	5.65	6.12	5.66	5.54	-3.13	-2.77	-2.33	-2.21	-11.88	-11.65	-10.31	-9.95	-2.95	-2.61	-2.20	-2.08
NN-S	4.37	4.64	4.48	4.49	-3.06	-3.53	-4.31	-4.45	-10.46	-11.67	-13.09	-13.39	-2.88	-3.34	-4.13	-4.27
NN-L	4.78	4.63	4.62	4.38	-3.14	-3.17	-4.30	-4.62	-11.06	-11.00	-13.22	-13.58	-2.95	-2.99	-4.12	-4.44
Global																
OLS+H	4.86	4.80	4.73	4.63	-3.65	-3.58	-3.37	-3.39	-12.14	-11.95	-11.48	-11.42	-3.45	-3.39	-3.19	-3.21
LASSO+H	4.83	4.83	4.83	4.83	-4.74	-4.74	-4.74	-4.74	-14.31	-14.31	-14.31	-14.31	-4.53	-4.53	-4.53	-4.53
RIDGE+H	4.86	4.81	4.67	4.63	-3.85	-3.81	-3.61	-3.66	-12.56	-12.43	-11.91	-11.98	-3.65	-3.61	-3.43	-3.48
Enet+H	4.87	4.83	4.83	4.83	-4.43	-4.39	-4.74	-4.74	-13.71	-13.59	-14.31	-14.31	-4.22	-4.18	-4.53	-4.53
PCR	4.28	3.82	4.13	4.04	-4.17	-6.28	-5.26	-5.84	-12.60	-16.34	-14.65	-15.69	-3.99	-6.09	-5.07	-5.65
XGBoost	3.67	4.43	4.49	4.34	-6.59	-6.17	-6.01	-5.83	-16.89	-16.87	-16.58	-16.05	-6.25	-5.82	-5.68	-5.52
CatBoost	3.58	4.64	5.92	4.71	-6.79	-5.66	-5.30	-5.91	-17.19	-16.02	-16.57	-16.59	-6.45	-5.33	-4.95	-5.55
LightGBM	3.68	4.17	4.12	4.22	-6.39	-6.02	-5.79	-5.75	-16.46	-16.30	-15.75	-15.77	-6.05	-5.69	-5.47	-5.44
NN-S	2.84	2.28	3.41	2.31	-6.75	-8.16	-7.82	-10.12	-15.75	-18.19	-19.01	-22.44	-6.51	-7.92	-7.61	-9.89
NN-L	2.48	2.89	2.42	2.46	-7.22	-7.26	-9.59	-9.80	-16.67	-17.38	-21.56	-22.02	-6.97	-7.03	-9.37	-9.57

Note: This table reports annualized Sharpe ratios of equal-weighted long–short decile portfolios formed using daily forecasts from different benchmarks models across rolling window sizes of 5, 21, 252, and 512 trading days. Performance is evaluated net of transaction costs under four scenarios: no costs (0 bps), fixed costs of 20 bps and 40 bps, and a mixed-cost structure where small- and large-cap firms face different estimated trading costs following Frazzini et al. (2012). In the mixed specification, the estimated costs correspond roughly to 21.3 bps for small-cap stocks and 11.2 bps for large-cap stocks. Models evaluated include linear (OLS+H, LASSO+H, RIDGE+H, Elastic Net+H, and PCR), ensemble (XGBoost, CatBoost, and LightGBM), and neural network (NN-S and NN-L) models. 'H' indicates that the model is estimated using the Huber loss. The top panel presents results based on training with U.S. excess return data, while the bottom panel presents results based on training with global excess return data.

Table B.18: Summary of Annual Out-of-Sample Observations and Securities (International)

Year	HKG		TWN		KOR		DEU		GBR		IND		AUS	
	Obs.	Sec.	Obs.	Sec.	Obs.	Sec.	Obs.	Sec.	Obs.	Sec.	Obs.	Sec.	Obs.	Sec.
2001	1,238,341	4,786	995,662	3,837	1,073,007	4,125	1,824,311	7,048	2,465,255	9,668	1,842,126	7,091	1,207,650	4,665
2002	1,236,290	4,792	994,886	3,837	1,065,897	4,121	1,935,883	6,960	2,424,283	9,491	1,817,678	7,057	1,200,798	4,656
2003	1,213,823	4,733	966,107	3,774	1,029,813	4,082	1,700,445	6,710	2,348,867	9,327	1,816,530	6,983	1,251,863	4,489
2004	1,215,428	4,688	937,583	3,654	985,351	3,806	1,715,175	6,638	2,351,844	9,095	1,820,045	6,959	1,333,090	4,396
2005	1,192,231	4,634	911,653	3,551	1,150,095	3,788	1,880,044	6,611	2,327,546	9,066	1,802,587	6,917	1,119,183	4,235
2006	1,573,919	4,431	1,150,169	3,504	1,318,278	3,778	2,314,676	6,585	3,213,960	9,024	2,261,908	6,925	1,490,811	4,191
2007	1,547,799	4,349	1,113,478	3,342	1,308,759	3,694	2,198,717	6,554	3,143,655	8,937	2,125,273	6,825	1,434,553	4,107
2008	1,518,318	4,306	1,143,572	3,295	1,266,153	3,637	2,193,099	6,540	3,060,502	8,770	2,395,684	6,816	1,388,031	3,965
2009	1,471,357	4,195	1,031,838	2,969	1,222,300	3,515	2,244,036	6,453	2,957,055	8,536	2,358,222	6,698	1,328,250	3,862
2010	1,452,087	4,113	1,029,330	2,922	1,165,581	3,429	2,185,223	6,299	2,922,933	8,305	2,281,897	6,517	1,310,182	3,739
2011	1,425,928	4,067	995,423	2,897	1,126,497	3,256	2,173,449	6,256	2,901,070	8,336	2,192,005	6,294	1,283,970	3,698
2012	1,431,792	4,064	994,235	2,837	1,126,360	3,226	2,193,525	6,208	2,941,836	8,347	2,108,958	6,021	1,294,363	3,678
2013	1,307,518	3,991	904,954	2,788	1,024,836	3,168	2,035,204	6,220	2,707,805	8,287	1,868,990	5,758	1,182,591	3,630
2014	1,185,983	3,890	823,794	2,742	944,451	3,113	1,850,396	6,118	2,456,971	8,124	1,648,124	5,474	1,072,071	3,543
2015	1,176,935	3,888	829,623	2,759	939,926	3,128	1,860,585	6,156	2,457,552	8,121	1,558,442	5,404	1,073,547	3,558
2016	1,174,714	3,867	839,801	2,763	959,750	3,139	1,874,817	6,154	2,466,864	8,080	1,466,026	4,855	1,067,928	3,544
2017	1,138,404	3,754	815,834	2,710	943,948	3,109	1,834,716	6,073	2,392,129	7,927	1,362,392	4,699	1,040,577	3,447
2018	1,098,403	3,678	801,371	2,717	945,217	3,120	1,814,150	6,022	2,372,240	7,904	1,328,470	4,438	1,013,397	3,437
2019	1,026,929	3,484	769,380	2,605	905,210	3,062	1,764,806	5,941	2,272,442	7,735	1,262,477	4,326	968,710	3,291
2020	1,004,155	3,393	769,270	2,579	890,938	3,014	1,771,886	5,919	2,257,760	7,566	1,240,385	4,207	966,120	3,249
2021	983,320	3,372	771,854	2,587	877,874	2,961	1,757,357	5,943	2,233,372	7,511	1,223,579	4,191	948,289	3,224
2022	931,348	3,304	763,020	2,605	854,233	2,913	1,732,755	5,922	2,205,613	7,526	1,181,921	4,104	920,774	3,167
2023	869,487	5,359	745,000	4,222	831,595	4,842	1,644,674	7,848	2,151,807	11,331	1,136,178	7,556	894,515	5,047
Overall	**28,414,509**	**5,398**	**21,097,837**	**4,246**	**23,956,069**	**4,882**	**44,499,929**	**7,896**	**59,033,361**	**11,394**	**40,099,897**	**7,628**	**26,791,263**	**5,083**

Note: This table reports the annual number of out-of-sample excess return observations (Obs.) and the count of unique securities (Sec.) for eight other markets from 2001 to 2023. Excess returns are reported as raw observation counts, while the number of securities represents distinct securities included each year. The overall number of securities in the last row represents the count of unique securities across all years combined. The set of countries is determined by three criteria: the relative size of the equity market (measured by total market capitalization), the allowance of short selling, and the availability of reliable data. The ordering of countries is arbitrary.

C Appendix: Benchmark Models

C.1 Linear Model and Regularized Variants

C.1.1 Linear Regression and Huber Loss

In the context of time series forecasting, linear models provide a transparent and interpretable baseline for predicting future observations from historical data. A linear forecasting model assumes that the future excess return r_{t+1}^{ex} can be expressed as a weighted combination of C lagged returns:

$$r_{t+1}^{ex} = \theta r_{t-C+1:t}^{ex} + \varepsilon, \tag{C.1}$$

where θ is the coefficient vector and ε is an error term with zero mean. The model parameters are estimated by minimizing a loss function over all observed samples:

$$\min_{\theta} \ \mathcal{L}(\theta) = \mathbb{E}_{\mathcal{D}_{\text{fin}}} \left[\ell\big(r_{t+1}^{ex}, \theta r_{t-C+1:t}^{ex}\big) \right], \tag{C.2}$$

where the expectation is taken over the dataset. When the loss is the mean squared error, i.e., $\ell(x, \hat{x}) = (x - \hat{x})^2$, the model reduces to ordinary least squares (OLS) regression. However, the squared loss is highly sensitive to outliers and heavy-tailed noise distributions, which are often present in financial or economic time series. To improve robustness, the Huber loss is widely adopted, defined as:

$$\ell_\delta(r_{t+1}^{ex}, \hat{r}_{t+1}^{ex}) = \begin{cases} \frac{1}{2}\left(r_{t+1}^{ex} - \hat{r}_{t+1}^{ex}\right)^2, & \text{if } |r_{t+1}^{ex} - \hat{r}_{t+1}^{ex}| \le \delta, \\ \delta\big(|r_{t+1}^{ex} - \hat{r}_{t+1}^{ex}| - \frac{\delta}{2}\big), & \text{otherwise,} \end{cases} \tag{C.3}$$

where $\delta > 0$ is a threshold that determines the transition between the quadratic and linear regimes. The resulting Huber regression for time series forecasting can thus be written as:

$$\min_{\theta} \ \mathbb{E}_{\mathcal{D}_{\text{fin}}} \left[\ell_\delta\big(r_{t+1}^{ex}, \hat{r}_{t+1}^{ex}\big) \right]. \tag{C.4}$$

Equation (C.4) yields a convex and differentiable optimization objective that is robust to noise and scalable to large time series datasets.

Such linear baselines not only capture essential autoregressive dependencies but also provide a transparent benchmark for evaluating more complex nonlinear models or TSFMs. We tune the regularization strength for OLS+H. In line with Leippold et al. (2022), who follow the recommendation of Huber (2011), we fix the robustness parameter of the Huber loss function at 1.35. See Table C.1

for the search grids and fixed settings.

C.1.2 Ridge Regression

Regularization techniques play a crucial role in improving the generalization ability and numerical stability of linear forecasting models, particularly when the number of predictors is large or the data exhibit multicollinearity. In time series forecasting, regularized linear models, such as Ridge regression (Hoerl and Kennard, 1970) and Lasso regression (Tibshirani, 1996), help control overfitting while enhancing interpretability through coefficient shrinkage.

Ridge regression introduces an ℓ_2-norm penalty on the coefficient vector, encouraging small but non-zero weights across correlated lagged predictors. The optimization objective can be written as:

$$\min_{\theta} \ \mathbb{E}_{\mathcal{D}_{\text{fin}}} \left[\ell_{\delta}\big(r^{ex}_{t+1}, \hat{r}^{ex}_{t+1}\big) + \lambda \|\theta\|_2^2 \right], \tag{C.5}$$

where $\lambda > 0$ is a regularization hyperparameter controlling the degree of shrinkage. The ℓ_2 penalty effectively reduces variance without eliminating predictors entirely, making Ridge regression particularly suitable for multicollinear time series where several lagged features convey overlapping information. This stabilization improves forecasting robustness and numerical conditioning. We tune the regularization strength, while the Huber threshold is fixed at 1.35. Exact ranges are shown in Table C.1.

C.1.3 Lasso Regression

In contrast, the Lasso regression imposes an ℓ_1-norm penalty on the coefficients:

$$\min_{\theta} \ \mathbb{E}_{\mathcal{D}_{\text{fin}}} \left[\ell_{\delta}\big(r^{ex}_{t+1}, \hat{r}^{ex}_{t+1}\big) + \lambda \|\theta\|_1 \right], \tag{C.6}$$

where the ℓ_1 penalty encourages sparsity in θ, automatically selecting a subset of the most relevant lagged predictors and setting others to zero. This property is especially beneficial in long-horizon forecasting, where many potential lags or exogenous variables may contain redundant information. By performing implicit feature selection, the Lasso reduces model complexity while preserving predictive accuracy. We tune the regularization strength, while the Huber threshold is fixed at 1.35. See Table C.1 for details.

C.1.4 Elastic Net

Both Ridge and Lasso regularizations enhance the predictive performance of linear models, but in complementary ways. Ridge regression yields smoother, dense solutions with small coefficients, which

is ideal when all predictors contribute modestly. Lasso, on the other hand, produces sparse models that are easier to interpret and computationally efficient, particularly valuable in high-dimensional or noisy time series environments.

These regularized estimators can also be combined through the Elastic Net (Zou and Hastie, 2005), which linearly interpolates between ℓ_1 and ℓ_2 penalties to balance sparsity and stability. Formally, the Elastic Net estimator for time series forecasting minimizes the following objective:

$$\min_{\theta} \ \mathbb{E}_{\mathcal{D}_{\text{fin}}} \left[\ell_\delta\big(r^{ex}_{t+1}, \hat{r}^{ex}_{t+1}\big) + \lambda_1 \|\theta\|_1 + \lambda_2 \|\theta\|_2^2 \right], \tag{C.7}$$

where $\lambda_1, \lambda_2 > 0$ are hyperparameters controlling the balance between sparsity and smooth shrinkage. The first term penalizes large prediction errors, while the subsequent regularization terms jointly constrain the magnitude and number of active coefficients.

The ℓ_1 term promotes sparsity by driving uninformative lag coefficients to zero, effectively performing automatic feature selection. The ℓ_2 term, in turn, distributes weight across correlated predictors, improving numerical stability and preventing over-penalization of groups of correlated lags. This hybrid structure allows the Elastic Net to retain both interpretability and robustness, which are essential in high-dimensional or multi-frequency time series where predictors may exhibit strong collinearity.

In practical forecasting applications, Elastic Net regression is often preferred over pure Lasso or Ridge models, as it adaptively balances bias reduction and variance control. By tuning (λ_1, λ_2), one can smoothly interpolate between a fully sparse model (Lasso) and a dense but stable model (Ridge), thus obtaining a flexible yet parsimonious representation of temporal dependencies. We tune the overall regularization and the mixing parameter, while the Huber threshold is fixed at 1.35. Search grids are provided in Table C.1.

C.2 Principal Component Regression

Linear forecasting models can suffer from multicollinearity and redundant predictors when the lagged input vectors $\mathbf{x}_{1:C}$ contain highly correlated components. This issue is especially pronounced in high-dimensional or macro-financial time series, where many variables exhibit strong co-movement. Principal component regression (PCR) (Massy, 1965) mitigates this problem by projecting the original predictors onto a lower-dimensional subspace spanned by the principal components that capture the dominant variance in the data.

Formally, let $R \in \mathbb{R}^{H \times C}$ denote the matrix of lagged predictors, where each row corresponds to a

time series of C lagged excess returns $r^{ex}_{t-C+1:t}$. PCR first performs a principal component analysis (PCA) on the standardized predictor matrix,

$$R = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \tag{C.8}$$

where $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_C]$ contains the eigenvectors of the covariance matrix $R^\top R$. The first k principal components correspond to the top k columns of $\mathbf{V}$, denoted as $\mathbf{V}_k$. Each observation is then represented in the reduced feature space as:

$$\mathbf{z} = \mathbf{V}^\top R. \tag{C.9}$$

Regression is then performed on these transformed features rather than the original predictors, by minimizing the following objective over weight θ:

$$\min_{\theta} \ \mathbb{E}_{\mathcal{D}_{\text{fin}}} \left[\ell_\delta \big(r^{ex}_{t+1} - \theta \mathbf{z} \big) \right], \tag{C.10}$$

where the loss $\ell_\delta(\cdot)$ can again be chosen as the Huber loss to improve robustness against outliers. By constraining the regression to depend only on the first k principal components, PCR effectively filters out noise and collinear structure in the original feature space.

In practice, PCR provides a computationally efficient way to stabilize linear forecasts when the predictor dimension is large relative to the sample size or when input correlations distort coefficient estimation. It inherits the interpretability of linear regression while integrating dimensionality reduction, thus offering a balance between model simplicity and statistical efficiency. We tune the number of principal components, while the PCA variant and downstream regressor are fixed. See Table C.1 for details.

C.3 Tree-Based Models

Tree-based ensemble models constitute a class of non-parametric forecasting methods that approximate nonlinear mappings between predictors and targets without assuming a specific functional form. Unlike linear models, which rely on additive relationships between input features, tree-based models recursively partition the predictor space into disjoint regions and fit simple local approximations, typically constants, within each region. This hierarchical partitioning allows the model to capture complex, nonlinear interactions and threshold effects that are prevalent in financial time series.

In ensemble tree models, $f(\cdot)$ is represented as a sum of M regression trees:

$$f_\theta(\mathbf{x}) = \sum_{m=1}^{M} \eta_m T_m(\mathbf{x}), \tag{C.11}$$

where $T_m(\cdot)$ denotes the m-th decision tree, and η_m is its corresponding weight. Here, θ denotes the parameters of the ensemble model. The trees are built sequentially to minimize a differentiable loss function (e.g., squared, Huber, or quantile loss), enabling efficient gradient-based boosting.

Such models, including Gradient Boosted Trees, XGBoost, LightGBM, and CatBoost, are well suited for structured tabular data and perform competitively on time series forecasting tasks after appropriate lag feature construction.

C.3.1 XGBoost

Extreme gradient boosting (XGBoost) is a scalable and highly optimized variant of gradient boosting that improves both predictive accuracy and computational efficiency through refined objective formulation and system-level optimization, proposed by Chen and Guestrin (2016). For a time series with lagged returns $r^{ex}_{t-C+1:t}$, the model approximates a nonlinear mapping:

$$\hat{r}^{ex}_{t+1} = f_\theta(r^{ex}_{t-C+1:t}) = \sum_{m=1}^{M} T_m(r^{ex}_{t-C+1:t}), \tag{C.12}$$

where each T_m is a regression tree added sequentially to minimize a regularized objective:

$$\mathcal{L}^{(j)} = \mathbb{E}_{\mathcal{D}_{\text{fin}}}\left[\ell_\delta\big(r^{ex}_{t-C+1:t}, \hat{r}^{(j-1)}_{t+1} + T_j(r^{ex}_{t-C+1:t})\big) + \lambda_1\|T_j\|_1 + \lambda_2\|T_j\|_2^2\right]. \tag{C.13}$$

In Equation (C.13), $\mathcal{L}^{(j)}$ denotes the loss after adding the j-th tree, and $\hat{r}^{(j-1)}_{t+1}$ is the prediction with the first $(j-1)$ trees. XGBoost differentiates itself from classical gradient boosting through several key innovations that enhance efficiency, regularization, and interpretability:

1. Second-order Taylor expansion of the objective. Instead of using only first-order gradients, XGBoost employs a second-order approximation of the loss function.

2. Explicit regularization and split gain criterion. XGBoost introduces both ℓ_1 and ℓ_2 regularization terms on leaf weights to control model complexity. Tree splits are selected by maximizing a gain function that ensures each split provides sufficient reduction in loss to justify model complexity.

3. Sparsity-aware and parallel optimization. XGBoost handles missing or sparse features efficiently by automatically learning the optimal default direction for missing values during tree construc-

tion. Additionally, histogram-based binning and parallel split search enable scalable training on large datasets.

We tune the tree depth and learning rate, while the booster is fixed to gradient-boosted trees. Exact grids are given in Table C.1.

C.3.2 CatBoost

Categorical boosting tree (CatBoost) is proposed by Prokhorenkova et al. (2018), and extends the standard gradient boosting framework to improve robustness and prevent overfitting, particularly in the presence of categorical or time-dependent features. CatBoost minimizes the following objective:

$$\min_{\theta} \mathbb{E}_{\mathcal{D}_{\text{fin}}} \left[\ell_\delta \big(r^{ex}_{t+1} - f_\theta(r^{ex}_{t-C+1:t}) \big) + \lambda \, \Omega(\theta) \right], \tag{C.14}$$

where $\ell_\delta(\cdot)$ is typically the Huber loss, and $\Omega(\theta)$ penalizes tree complexity.

CatBoost introduces several methodological innovations that distinguish it from conventional boosting models. First, it uses ordered boosting, a permutation-driven scheme that ensures each sample's gradient is estimated only from earlier observations, thereby preventing target leakage and prediction shift. Second, it employs a symmetric tree structure where all nodes at a given depth share the same splitting rule, producing balanced and efficient trees that reduce overfitting and accelerate inference. Third, categorical variables are handled through ordered target encoding, where sequential target statistics are computed from preceding samples to preserve temporal causality.

Through these mechanisms, CatBoost achieves high predictive accuracy and interpretability while maintaining unbiased gradient estimation and efficient computation. Its design makes it particularly suitable for structured and temporal datasets, offering a reliable nonlinear baseline for comparison. We tune the tree depth and learning rate, while the loss is fixed to RMSE. See Table C.1 for ranges.

C.3.3 LightGBM

Light gradient boosting machine (LightGBM) is an efficient implementation of gradient boosting designed for high-dimensional and large-scale datasets, proposed by Ke et al. (2017). It retains the basic ensemble structure of additive regression trees used in XGBoost but introduces several algorithmic innovations that significantly improve training speed, memory efficiency, and scalability without sacrificing accuracy. LightGBM adopts the same additive tree-based ensemble structure as other gradient boosting models but introduces several computational and algorithmic improvements, such as leaf-wise growth, gradient-based sampling, and histogram-based split finding. Given lagged excess returns

$r^{ex}_{t-C+1:t}$ as input features, LightGBM minimizes the following regularized objective:

$$\min_{\theta} \mathbb{E}_{\mathcal{D}_{\text{fin}}} \left[\ell_\delta\big(r^{ex}_{t+1} - f(r^{ex}_{t-C+1:t})\big) + \lambda\, \Omega(\theta) \right], \tag{C.15}$$

where $\ell_\delta(\cdot)$ is typically the Huber loss used for robust regression, and $\Omega(\theta)$ penalizes the complexity of the ensemble of trees. In LightGBM, tree construction follows a leaf-wise growth strategy, and the optimization leverages gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB) to improve computational efficiency while maintaining predictive accuracy.

Through the aforementioned mechanisms, LightGBM achieves substantial gains in speed and scalability compared to traditional boosting frameworks. We tune the number of leaves and learning rate, while the maximum depth and boosting type are fixed. Grids are reported in Table C.1.

C.4 Neural Networks

Neural networks extend traditional regression models by learning nonlinear mappings between input features and target variables through a hierarchy of linear transformations and nonlinear activations. Among them, the feedforward neural network (FNN), also known as the multilayer perceptron (MLP), is one of the simplest yet most versatile architectures for modeling complex relationships in time series data.

The FNN consists of L hidden layers, each performing an affine transformation followed by a nonlinear activation:

$$\mathbf{h}^{(l)} = \sigma^{(l)}\Big(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\Big), \quad l = 1, \ldots, L, \tag{C.16}$$

where $\mathbf{h}^{(0)} = r^{ex}_{t-C+1:t} \in \mathbb{R}^C$ denotes the input, $\mathbf{W}^{(l)} \in \mathbb{R}^{n_l \times n_{l-1}}$ and $\mathbf{b}^{(l)} \in \mathbb{R}^{n_l}$ denote the weights and biases of layer l, and $\sigma^{(l)}(\cdot)$ is a nonlinear activation function, such as rectified linear unit. The final output layer produces a scalar forecast $\hat{r}^{ex}_{t+1} \in \mathbb{R}$:

$$\hat{r}^{ex}_{t+1} = \mathbf{w}_{\text{out}}^{\top}\mathbf{h}^{(L)} + b_{\text{out}}. \tag{C.17}$$

Thus, the overall mapping from lagged excess returns to future values can be compactly expressed as:

$$\hat{r}^{ex}_{t+1} = f_\theta(r^{ex}_{t-C+1:t}), \tag{C.18}$$

where f_θ represents the neural network parameterized by weights and biases $\theta = \{\mathbf{W}^{(l)}, \mathbf{b}^{(l)}\}_{l=1}^{L}$. The network parameters are estimated by minimizing a differentiable loss function, such as the mean

squared loss, over the training samples:

$$\min_{\theta} \mathbb{E}_{\mathcal{D}_{\text{fin}}} \left[\ell\big(r^{ex}_{t+1} - f_\theta(r^{ex}_{t-C+1:t})\big) + \lambda \left\|\theta\right\|_2^2 \right], \tag{C.19}$$

where the regularization term controls weight magnitudes and prevents overfitting. Parameter updates are performed via stochastic gradient descent (SGD) or its adaptive variants (Adam, RMSProp), allowing efficient optimization even in high-dimensional feature spaces.

By adjusting the input window length C and network depth L, the FNN can approximate both short-term and long-term temporal dependencies. Although it does not explicitly model sequence dynamics as recurrent or attention-based architectures do, it remains a strong baseline due to its simplicity, parallelizability, and universal function approximation capability.

In forecasting practice, FNNs offer a balance between model interpretability and flexibility: they can approximate nonlinear dependencies while remaining computationally efficient and straightforward to train, thus serving as a bridge between classical econometric models and more complex ML architectures. The neural network architecture and training settings are fixed as described in Table C.1.

Table C.1: Hyperparameter Settings for Benchmark Models

Model	Hyperparameters
OLS+H	**Tuned:** – Regularization strength (α): searched over values from 10^{-6} to 10^{1} on a logarithmic scale (8 values) **Fixed:** – Robustness parameter (ϵ): 1.35
LASSO+H	**Tuned:** – Regularization strength (α): searched over values from 10^{-6} to 10^{1} on a logarithmic scale (8 values) **Fixed:** – Robustness parameter (ϵ): 1.35 – Penalty: L1 (Lasso)
Ridge+H	**Tuned:** – Regularization strength (α): searched over values from 10^{-6} to 10^{1} on a logarithmic scale (8 values) **Fixed:** – Robustness parameter (ϵ): 1.35 – Penalty: L2 (Ridge)
Enet+H	**Tuned:** – Regularization strength (α): searched over values from 10^{-6} to 10^{1} on a logarithmic scale (8 values) – Elastic Net mixing parameter ($l1_ratio$): $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ **Fixed:** – Robustness parameter (ϵ): 1.35 – Penalty: Elastic Net
PCR	**Tuned:** – Number of principal components: candidate sets depending on the feature dimension are as follows: $p = 5$: $\{2, 3\}$; $p = 21$: $\{2, 4, 6, 8, 10, 12\}$; $p = 252$: $\{16, 32, 48, 64, 96, 126\}$; $p = 512$: $\{32, 64, 96, 128, 192, 256\}$ **Fixed:** – PCA algorithm: Incremental PCA – Regression: Linear Regression (OLS)
XGBoost	**Tuned:** – Maximum tree depth: $\{3, 5, 7, 9, 11\}$ – Learning rate (η): $\{0.005, 0.01, 0.05, 0.1, 0.2\}$ **Fixed:** – Booster: gradient boosted trees
CatBoost	**Tuned:** – Maximum tree depth: $\{2, 3, 4, 5, 6, 8, 10, 12, 14, 16\}$ – Learning rate (η): $\{0.01, 0.05, 0.1\}$ **Fixed:** – Loss function: RMSE (default for regression)
LightGBM	**Tuned:** – Number of leaves: $\{32, 64, 128, 256, 512\}$ – Learning rate (η): $\{0.01, 0.05, 0.1\}$ **Fixed:** – Maximum depth: No limit – Boosting type: Gradient Boosted Decision Trees (GBDT)
NN	**Fixed:** – Hidden layer size: Single hidden layer with different numbers of units – Activation: ReLU – Optimizer: Adam ($\eta = 0.001$) – Loss function: MSE with L1 regularization ($\lambda = 10^{-4}$) – Learning rate scheduler: Reduce Learning Rate on Plateau – Dropout: 0.2 – BatchNorm on hidden layer – Epochs: 30 – Early stopping patience: 5

Note: This table presents the hyperparameter settings for the benchmark models. The set of models is drawn from Gu et al. (2020) and Leippold et al. (2022), subject to the additional requirement that they can be trained efficiently on the large-scale dataset employed in this study. Specifically, we include OLS+H, LASSO+H, Ridge+H, Enet+H, PCR, and neural network (NN). We also add XGBoost, CatBoost, and LightGBM, which have been consistently recognized as among the best-performing models across a wide range of applications. Hyperparameter tuning is performed only for the first year (2000), and the selected hyperparameters are subsequently applied to the remaining years. This procedure ensures computational feasibility while enabling a fair comparison across different classes of models. The key hyperparameters that were optimized are reported under 'Tuned', while those held constant throughout the analysis are reported under 'Fixed'. 'H' indicates that the model is estimated using the Huber loss.

D Summary of Time Series Foundation Models

In this section, we first provide a brief introduction to the other TSFM variants used in our numerical experiments and conclude with summary tables (Table D.1 and Table D.2) to facilitate the comparison between models.

Moirai (Uni2TS) (Woo et al., 2024) introduces frequency-specialized input and output projections, patch-based tokenization, and a masked-encoder pre-training objective that reconstructs masked patches rather than relying purely on autoregression. Building on this TSFM, Moirai 2 adopts a decoder-only architecture and is pre-trained on a broad mixture of datasets. It replaces the previous distributional loss with a quantile-loss formulation, transitions from single-token to multi-token prediction for improved efficiency and stability, filters out low-quality or non-forecastable time series during pre-training, enriches patch embeddings with missing-value indicators to better handle missing data, and introduces a patch-level random masking mechanism to enhance robustness during inference.

Kairos (Feng et al., 2025) is a TSFM with an encoder–decoder Transformer and two core modules: Mixture-of-Size Dynamic Patching (MoS-DP) and Instance-Adaptive Rotary Position Embedding (IARoPE). MoS-DP adaptively tokenizes by routing each coarsest patch to experts tied to different patch sizes (including null sizes) and fusing ancestor features, yielding fine granularity where signals change quickly and coarser granularity in stable regions. IARoPE tailors positional encodings per instance by modulating rotary positional encodings (RoPE) with low-frequency Fast Fourier Transform (FFT)-based features. For forecasting, Kairos uses multi-patch prediction to decode multiple future patches in parallel, mitigating autoregressive error accumulation.

Moment (Goswami et al., 2024) is an encoder-only TSFM pre-trained with masked time-series modeling. It applies reversible instance normalization and tokenizes inputs into fixed-length patches, then learns to reconstruct masked patches with a lightweight prediction head; the Transformer uses relative positional encodings augmented by absolute sinusoidal embeddings. Moment operates channel-wise for multivariate inputs. With simple linear probes or zero-shot use, the learned patch representations transfer across forecasting, anomaly detection, classification, and imputation under limited supervision.

Lag-Llama (Rasul et al., 2023) is a decoder-only TSFM for univariate probabilistic time series forecasting. Each timestep is tokenized as a vector of lagged values and date-time features; tokens pass through a shared linear projection and a causally masked Transformer using root mean square layer normalization (RMSNorm) and RoPE. A distribution head outputs a Student-t distribution for the next value, enabling autoregressive multi-step simulation and uncertainty intervals. The lag-as-

covariates, frequency-agnostic design and decoder-only architecture make rollout straightforward.

TiRex (Auer et al., 2025) is a decoder-only TSFM built on xLSTM (Beck et al., 2025) blocks that preserve recurrence for state tracking. The model normalizes each series with z-score instance normalization, tokenizes with non-overlapping input and output windows, and maps patches to and from the hidden space via lightweight two-layer residual blocks with RMSNorm and residual connections inside each xLSTM block. It predicts nine equidistant quantiles under a quantile-loss objective, enabling calibrated probabilistic forecasts. To stabilize long-horizon rollout, TiRex introduces contiguous patch masking (CPM) during pre-training, where consecutive future patches are masked and treated as missing, so multi-patch horizons at inference are handled without autoregressive re-feeding and with coherent uncertainty propagation.

FlowState (Graf et al., 2025) pairs a state space model (SSM)–based encoder with a functional-basis decoder (FBD) that produces a continuous forecast which is then sampled at arbitrary horizons; by modulating the discretization step with a scale factor, the same model adapts on the fly to different sampling rates without input patching or quantization. It enforces strictly causal normalization via running mean–variance to prevent leakage, and is pre-trained with parallel forecasts from progressively longer contexts to improve robustness.

TTM (Ekambaram et al., 2024), short for 'tiny time mixers,' is a compact TSFM that replaces quadratic self-attention with MLP-Mixer blocks augmented by lightweight gated attention, arranged in a multi-level backbone with a slim decoder. Inputs undergo per-instance z-score normalization and non-overlapping patch tokenization; the backbone operates channel-independently, while the decoder can mix channels and fuse known exogenous signals. Pre-training adopts direct forecasting with MSE, freezing the backbone and fine-tuning a small TTM head (slim decoder and forecast linear head); an exogenous mixer leverages known future covariates via a stride-1 patched TSMixer block. To improve scale and resolution robustness, TTM introduces adaptive patching across backbone levels, diverse resolution sampling, and a learnable resolution prefix token.

Toto (Cohen et al., 2024), short for 'time series optimized Transformer for observability,' is a decoder-only TSFM for multivariate observability data that uses non-overlapping patch embeddings with per-variate causal patch-based instance normalization (with a clipping mechanism) to handle extreme nonstationarity; proportional factorized attention that prioritizes time-wise over variate-wise mixing for scalable axis interactions; and a Student-t mixture model (SMM) head trained with a composite objective combining negative log-likelihood and a robust point loss.

Sundial (Liu et al., 2025) is a decoder-only TSFM that keeps inputs continuous via patch embed-

ding (length-P) with per-patch binary masks and applies stationarization (instance normalization) as re-normalization, avoiding discrete tokenization. Its backbone is a stabilized Transformer using Pre-Layer Normalization (Pre-LN), RoPE, FlashAttention, and a key–value (KV) cache for efficient causal self-attention over patch tokens. Probabilistic forecasting is trained end-to-end with TimeFlow loss, a flow-matching objective that conditions a small MLP (FM-Net) on the lookback representation to model each next patch's continuous predictive distribution. For rollout, Sundial uses multi-patch prediction to reduce autoregressive steps and performs repeated sampling at inference to estimate statistics such as medians and quantiles while reusing the same lookback representation for speed.

Table D.1 and Table D.2 together provide a comprehensive overview of Chronos, TimesFM, and the above-described TSFMs. Table D.1 focuses on the institutional origin, backbone architecture, loss function, and embedding of each model, highlighting their key features and official repositories. Meanwhile, Table D.2 summarizes the pre-training datasets used by these models, detailing their composition in terms of real and synthetic data, as well as the total number of observations involved, and whether they include financial-related data in their pre-training. Together, these tables contextualize how architectural and data-scale choices differ across the current landscape of TSFMs.

Table D.1: TSFMs: Institutions, Architectures, Embedding, Loss Functions, and Key Features

Model	Issuing Institution	Backbone Architecture	Embedding	Website	Loss Function	Key Features
Chronos	Amazon	Transformer	Quantization	GitHub Link	Cross-entropy (token NLL)	Employs a discrete-value quantization tokenizer; the Chronos-Bolt variant accelerates rollout and enhances long-horizon inference
TimesFM	Google	Transformer (decoder-style)	Patching	GitHub Link	MSE	Utilizes patch tokens with multi-horizon decoding; demonstrates strong zero-shot performance on long sequences
Moirai	Salesforce	Transformer (+MoE variants)	Patching	GitHub Link	Distributional NLL (v1.x); Quantile loss (v2.0+)	Incorporates multi-resolution patching; applies Mixture-of-Experts for scalability
Kairos	ShanghaiTech University, Ant Group	Transformer encoder–decoder	Patching (dynamic, multi-scale)	GitHub Link	MSE	Uses Mixture-of-Size Dynamic Patching (adaptive size with null experts); includes instance-adaptive RoPE; enables multi-patch decoding for efficient long-range forecasting
Moment	Carnegie Mellon University, University of Pennsylvania	Encoder Transformer + MLP head	Patching	GitHub Link	Reconstruction loss	Applies reversible instance normalization (RevIN) with masked patch reconstruction; combines relative and absolute positional encodings; lightweight reconstruction head for stable scaling
Lag-Llama	University of Montreal, McGill University	Decoder-only Transformer	Lag values with mixed frequency	GitHub Link	Student-t NLL	Represents lag values (past observations and calendar features) as embedding; causal attention with RoPE; Student-t probabilistic head
TiRex	NX-AI, Johannes Kepler University Linz	xLSTM (decoder-style)	Patching	GitHub Link	Quantile (pinball) loss	Stacked xLSTM blocks with lightweight I/O; patch-to-horizon mapping; supports missing-value encoding for multi-patch forecasting
FlowState	IBM Research, ETH Zurich	State-space (S5) encoder + Functional Basis Decoder	Causal normalization	GitHub Link	Quantile (pinball) loss	Continuous-time modeling; horizon-agnostic resampling via basis decoder; parallel forecasting and time-scale adaptation
TTM	IBM Research	MLP/TSMixer + gated mixing	Patching (multi-resolution)	GitHub Link	MSE	Compact models (1–5M params) using channel–time mixing; resolution-prefix tuning; CPU-efficient zero-/few-shot performance
Toto	Datadog AI, Carnegie Mellon University	Decoder-only Transformer	Patching	GitHub Link	Student-t NLL + robust loss	Causal per-patch scaling for nonstationarity; factorized attention across time/variates; outputs Student-t mixtures for heavy tails
Sundial	Tsinghua University	Decoder-only Transformer	Re-normalization + Patching	GitHub Link	TimeFlow (flow-matching) loss	Instance re-normalization and patching to construct continuous embedding; multi-patch prediction with RoPE for efficient long sequences

Note: This table summarizes leading time series foundation models (TSFMs) in terms of their issuing institutions, underlying backbone architectures, tokenization approaches, websites, loss function, and key features. The 'Issuing Institution' column lists the primary organizations responsible for developing each model, while the 'Backbone Architecture' column specifies the core neural architecture. The 'Embedding' column describes the type of embedding used. The 'Website' and 'Key Features' columns provide the corresponding repository URL and a concise summary of the model's main innovations. 'NLL' stands for negative log-likelihood.

Table D.2: TSFM Pre-Training Datasets: Provenance, Composition, Volume, and Financial Data Inclusion

Model	Pre-training Datasets	Composition (%)	Number of Observations	Finance-Related Datasets
Chronos	Publicly available datasets from domains including Monash, Kaggle, M-series, retail, weather, and finance, plus synthetic datasets generated via Gaussian Process (GP) methods (KernelSynth) and Time Series Mixup (TSMixup) augmentation	90% real, 10% synthetic	100 B	None
TimesFM	Real-world and synthetic time series datasets include Wikipedia Pageviews, Google Trends, M4 competition datasets, Electricity, Traffic, Weather, Favorita Sales, and LibCity, as well as ARMA-generated, trend, seasonal, and step-function series composed of mixtures of sine/cosine, trend, and step components	80% real, 20% synthetic	100 B	M4 competition dataset (yearly, quarterly, monthly, weekly, daily, and hourly frequencies)
Moirai	Nine domains: Energy, Transport, Climate, CloudOps, Web, Sales, Nature, Economics/Finance, and Healthcare. The dataset is aggregated from numerous open sources such as Monash, GluonTS, and various long-sequence benchmarks	100% real	27 B	The Monash archive: M-competitions (M1, M3, M4 and M5 with yearly, quarterly, monthly, weekly, daily, and hourly frequencies), NN5 (daily and weekly), CIF-2016 (monthly), FRED-MD (monthly macroeconomic series), and Bitcoin (daily) datasets. In addition, the GoDaddy (monthly) microbusiness dataset was included
Kairos	Predictability-Stratified Time Series (PreSTS) corpus built from diverse real-world datasets including ERA5 and NOAA climate data, LibCity and PeMS traffic series, financial and healthcare signals, IoT telemetry, and retail benchmarks such as Favorita and M4, with a small synthetic subset generated using Gaussian Process and ARMA-based methods	Primarily real with minor synthetic augmentation	300 B	Same as Moirai
Moment	Curated collection of open-source datasets aggregated from over five public repositories, encompassing diverse domains such as healthcare, energy, finance, engineering, traffic, and weather, and derived from sources including Monash, UCI, GluonTS, and other public archives	100% real	1.23 B	Same as Moirai
Lag-Llama	Corpus of 27 open time series datasets spanning six domains: energy, transportation, economics, nature, air quality, and cloud operations. Datasets drawn from Monash Time Series Forecasting Archive, GluonTS, and other open benchmarks	100% real	1 B	None
TiRex	Chronos training datasets; Synthetic gaussian process (GP) time series generated using a KernelSynth-style sampling procedure; and a subset of the GIFT-Eval pre-training corpus	92% Chronos + GP, 8% GIFT-Eval	2.5 B	None
FlowState	GIFT-Eval pre-training corpus; TiRex pre-training data (Chronos corpus and GIFT-Eval subset, excluding overlap); and additional synthetic GP time series generated via KernelSynth	Primarily real (approximately 90% real, 10% synthetic)	2.5 B	None
TTM	Public datasets from the Monash Time Series Forecasting Archive and LibCity traffic data repository, covering diverse domains such as weather, traffic, retail, and energy	100% real	1 B	The Monash archive: Bitcoin (daily) dataset
Toto	Internal sources include Datadog observability and system metrics, while open sources consist of the GIFT-Eval (General Industrial Forecasting Testbed), LSF (Large-Scale Forecasting), Chronos, and TimesFM datasets. Synthetic data is generated through procedural and GP methods	43% internal, 24% open, 33% synthetic	2,360 B	Same as Moirai
Sundial	TimeBench corpus collected from diverse real-world domains, including climate (ERA5), IoT, finance, ECG, LOTSA, and Chronos datasets	99.95% real, 0.05% synthetic	1,000 B	The relevant details have not been provided

Note: This table summarizes the pre-training datasets and their compositions for leading time series foundation models (TSFMs). The 'Composition' column reports the approximate proportions of real-world versus synthetically generated time series used during model pre-training. The 'Number of Observations' indicates the approximate number of samples (individual time series instances), in billions, that were used during model pre-training. The 'Finance-Related Datasets' column indicates whether any financial data sources were included among the pre-training datasets. Each model type may include multiple versions or checkpoints; however, all were pre-trained on the same underlying dataset.

References

Ansari, A. F., L. Stella, C. Turkmen, X. Zhang, P. Mercado, H. Shen, O. Shchur, S. S. Rangapuram, S. Pineda Arango, S. Kapoor, J. Zschiegner, D. C. Maddix, M. W. Mahoney, K. Torkkola, A. Gordon Wilson, M. Bohlke-Schneider, and Y. Wang (2024). Chronos: Learning the Language of Time Series. *Transactions on Machine Learning Research*.

Asness, C. S., T. J. Moskowitz, and L. H. Pedersen (2013). Value and Momentum Everywhere. *The Journal of Finance 68*(3), 929–985.

Auer, A., P. Podest, D. Klotz, S. Böck, G. Klambauer, and S. Hochreiter (2025). TiRex: Zero-Shot Forecasting Across Long and Short Horizons with Enhanced In-Context Learning.

Bali, T. G., R. F. Engle, and S. Murray (2016). *Empirical Asset Pricing: The Cross Section of Stock Returns*. John Wiley & Sons.

Barroso, P. and P. Santa-Clara (2015). Momentum Has Its Moments. *Journal of Financial Economics 116*(1), 111–120.

Beck, M., K. Pöppel, P. Lippe, R. Kurle, P. M. Blies, G. Klambauer, S. Böck, and S. Hochreiter (2025). xLSTM 7B: A Recurrent LLM for Fast and Efficient Inference.

Berk, J. (2023). Comment on "The Virtue of Complexity in Return Prediction". *Available at SSRN 4410125*.

Brock, W., J. Lakonishok, and B. LeBaron (1992). Simple Technical Trading Rules and the Stochastic Properties of Stock Returns. *The Journal of Finance 47*(5), 1731–1764.

Buncic, D. (2025). Simplified: A Closer Look at the Virtue of Complexity in Return Prediction. *Available at SSRN*.

Cartea, Á., Q. Jin, and Y. Shi (2025). The Limited Virtue of Complexity in a Noisy World. *Available at SSRN*.

Chen, L., M. Pelger, and J. Zhu (2024). Deep Learning in Asset Pricing. *Management Science 70*(2), 714–750.

Chen, T. and C. Guestrin (2016, August). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, New York, NY, USA, pp. 785–794.

Cohen, B., E. Khwaja, K. Wang, C. Masson, E. Ramé, Y. Doubli, and O. Abou-Amal (2024). Toto: Time Series Optimized Transformer for Observability. *arXiv preprint arXiv:2407.07874*.

Das, A., W. Kong, R. Sen, and Y. Zhou (2024). A Decoder-Only Foundation Model for Time-Series Forecasting. In *Forty-first International Conference on Machine Learning*.

Ehsani, S. and J. T. Linnainmaa (2022). Factor Momentum and the Momentum Factor. *The Journal of Finance 77*(3), 1877–1919.

Ekambaram, V., A. Jati, P. Dayama, S. Mukherjee, N. Nguyen, W. M. Gifford, C. Reddy, and J. Kalagnanam (2024). Tiny Time Mixers (TTMs): Fast Pre-trained Models for Enhanced Zero/Few-Shot Forecasting of Multivariate Time Series. *Advances in Neural Information Processing Systems 37*, 74147–74181.

Feng, K., S. Lan, Y. Fang, W. He, L. Ma, X. Lu, and K. Ren (2025). Kairos: Towards Adaptive and Generalizable Time Series Foundation Models. *arXiv preprint arXiv:2509.25826*.

Frazzini, A., R. Israel, and T. J. Moskowitz (2012). Trading Costs of Asset Pricing Anomalies. *Fama-Miller Working Paper, Chicago Booth Research Paper* (14-05).

Goswami, M., K. Szafer, A. Choudhry, Y. Cai, S. Li, and A. Dubrawski (2024). MOMENT: A Family of Open Time-series Foundation Models. *arXiv preprint arXiv:2402.03885*.

Graf, L., T. Ortner, S. Woźniak, and A. Pantazi (2025). FlowState: Sampling Rate Invariant Time Series Forecasting.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies 33*(5), 2223–2273.

Gupta, T. and B. T. Kelly (2018). Factor Momentum Everywhere.

He, S., L. Lv, A. Manela, and J. Wu (2025). Chronologically Consistent Large Language Models. *arXiv preprint arXiv:2502.21206*.

Hoerl, A. E. and R. W. Kennard (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics 12*(1), 55–67.

Hoffmann, J., S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. (2022). Training Compute-Optimal Large Language Models. *arXiv preprint arXiv:2203.15556*.

Huang, A. H., H. Wang, and Y. Yang (2023). FinBERT: A Large Language Model for Extracting Information from Financial Text. *Contemporary Accounting Research 40*(2), 806–841.

Huang, H., M. Chen, and X. Qiao (2024). Generative Learning for Financial Time Series with Irregular and Scale-Invariant Patterns. In *The Twelfth International Conference on Learning Representations*.

Huber, P. J. (2011). Robust Statistics. In *International Encyclopedia of Statistical Science*, pp. 1248–1251. Springer.

Jegadeesh, N. (1990). Evidence of Predictable Behavior of Security Returns. *The Journal of Finance 45*(3), 881–898.

Jegadeesh, N. and S. Titman (1993, March). Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *The Journal of Finance 48*(1), 65–91.

Jensen, T. I., B. Kelly, and L. H. Pedersen (2023). Is There a Replication Crisis in Finance? *The Journal of Finance 78*(5), 2465–2518.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 30. Curran Associates, Inc.

Kelly, B., S. Malamud, and K. Zhou (2024). The Virtue of Complexity in Return Prediction. *The Journal of Finance 79*(1), 459–503.

Kelly, B. T., B. Kuznetsov, S. Malamud, and T. A. Xu (2025). Artificial Intelligence Asset Pricing Models. Technical report, National Bureau of Economic Research.

Kelly, B. T. and S. Malamud (2025). Understanding The Virtue of Complexity. *Available at SSRN 5346842*.

Kingma, D. P. and J. Ba (2015). Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*.

Kudo, T. (2018). Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Leippold, M., Q. Wang, and W. Zhou (2022, August). Machine Learning in the Chinese Stock Market. *Journal of Financial Economics 145*(2, Part A), 64–82.

Li, B., A. G. Rossi, X. S. Yan, and L. Zheng (2025). Machine Learning from a "Universe" of Signals: The Role of Feature Engineering. *Journal of Financial Economics 172*, 104138.

Liang, Y., H. Wen, Y. Nie, Y. Jiang, M. Jin, D. Song, S. Pan, and Q. Wen (2024). Foundation Models for Time Series Analysis: A Tutorial and Survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6555–6565.

Liao, S., H. Ni, M. Sabate-Vidales, L. Szpruch, M. Wiese, and B. Xiao (2024). Sig-Wasserstein GANs for Conditional Time Series Generation. *Mathematical Finance 34*(2), 622–670.

Liu, A., B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, et al. (2024). DeepSeek-V3 Technical Report. *arXiv preprint arXiv:2412.19437*.

Liu, Y., G. Qin, Z. Shi, Z. Chen, C. Yang, X. Huang, J. Wang, and M. Long (2025). Sundial: A Family of Highly Capable Time Series Foundation Models. *arXiv preprint arXiv:2502.00816*.

Liu, Y. and A. Tsyvinski (2021). Risks and Returns of Cryptocurrency. *The Review of Financial Studies 34*(6), 2689–2727.

Lo, A. W., H. Mamaysky, and J. Wang (2000). Foundations of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation. *The Journal of Finance 55*(4), 1705–1765.

Martin, I. W. and S. Nagel (2022). Market Efficiency in the Age of Big Data. *Journal of Financial Economics 145*(1), 154–177.

Massy, W. F. (1965). Principal Components Regression in Exploratory Statistical Research. *Journal of the American Statistical Association 60*(309), 234–256.

McIntosh-Smith, S., S. Alam, and C. Woods (2024). Isambard-AI: A Leadership-Class Supercomputer Optimised Specifically for Artificial Intelligence. In *Proceedings of the Cray User Group*, pp. 44–54.

Menkhoff, L., L. Sarno, M. Schmeling, and A. Schrimpf (2012). Currency Momentum Strategies. *Journal of Financial Economics 106*(3), 660–684.

Moskowitz, T. J., Y. H. Ooi, and L. H. Pedersen (2012). Time Series Momentum. *Journal of Financial Economics 104*(2), 228–250.

Nagel, S. (2025). Seemingly Virtuous Complexity in Return Prediction. *Chicago Booth Research Paper* (25-10).

Neely, C. J., D. E. Rapach, J. Tu, and G. Zhou (2014, March). Forecasting the Equity Risk Premium: The Role of Technical Indicators. *Management Science*.

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. (2022). Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems 35*, 27730–27744.

Prokhorenkova, L., G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin (2018). CatBoost: Unbiased Boosting with Categorical Features. In *Advances in Neural Information Processing Systems*, Volume 31.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research 21*(140), 1–67.

Rahimikia, E. and F. Drinkall (2024). Re(Visiting) Large Language Models in Finance. *Available at SSRN*.

Rasul, K., A. Ashok, A. R. Williams, H. Ghonia, R. Bhagwatkar, A. Khorasani, M. J. D. Bayazi, G. Adamopoulos, R. Riachi, N. Hassen, et al. (2023). Lag-Llama: Towards Foundation Models for Probabilistic Time Series Forecasting. *arXiv preprint arXiv:2310.08278*.

Sennrich, R., B. Haddow, and A. Birch (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology 58*(1), 267–288.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. (2023). LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems 30*.

Vuletić, M., F. Prenzel, and M. Cucuringu (2024). Fin-GAN: Forecasting and Classifying Financial Time Series via Generative Adversarial Networks. *Quantitative Finance 24*(2), 175–199.

Wiese, M., R. Knobloch, R. Korn, and P. Kretschmer (2020). Quant GANs: Deep Generation of Financial Time Series. *Quantitative Finance 20*(9), 1419–1440.

Woo, G., C. Liu, A. Kumar, C. Xiong, S. Savarese, and D. Sahoo (2024, 21–27 Jul). Unified Training of Universal Time Series Forecasting Transformers. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp (Eds.), *Proceedings of the 41st International Conference on Machine Learning*, Volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164. PMLR.

Yu, A., D. C. Maddix, B. Han, X. Zhang, A. F. Ansari, O. Shchur, C. Faloutsos, A. G. Wilson, M. W. Mahoney, and Y. Wang (2025). Understanding Transformers for Time Series: Rank Structure, Flow-of-ranks, and Compressibility. *arXiv preprint arXiv:2510.03358*.

Zou, H. and T. Hastie (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology 67*(2), 301–320.