An Open Benchmark for Evaluating Time Series Forecasting Methods across Financial Markets

Jeremiah Bejarano*† Viren Desai‡ Kausthub Keshava‡ Arsh Kumar‡
Zixiao Wang‡ Vincent Hanyang Xu‡ Yangge Xu‡

November 3, 2025

Abstract

Financial regulators and researchers have emphasized forward-looking risk monitoring to address systemic vulnerabilities. This paper benchmarks state-of-the-art global time series forecasting methods, which have proven superior in the time series literature, on a wide-ranging suite of financial datasets. Benchmarks drive progress, and our systematic evaluation of over a dozen forecasting methods ranging from classical models to modern deep learning architectures reveals which approaches best capture early warning signals across different market segments. We evaluate these methods on critical financial stability metrics including arbitrage basis spreads that signal funding market stress, banking indicators that reveal institutional vulnerabilities, and asset returns across multiple markets. To enable reproducible research and continuous improvement in financial forecasting, we develop an open-source *financial time series forecasting repository* that standardizes these datasets according to canonical academic methodologies. Our results provide financial stability authorities with evidence-based guidance on which forecasting approaches most reliably detect emerging risks in specific market segments, directly enhancing the toolkit for macroprudential surveillance and systemic risk monitoring. Consistent with decades of empirical finance, returns remain extremely difficult to forecast, yet machine-learning-based global models yield meaningful accuracy gains for basis spreads, liquidity metrics, and other supervisory indicators where classical baselines fall short.

1 Introduction

Financial regulators and researchers increasingly emphasize forward-looking risk monitoring to preemptively address systemic vulnerabilities.[1] (Adrian, Covitz, and Liang, 2015). In this context, improving

*Office of Financial Research, U.S. Department of the Treasury and the Financial Mathematics program at the University of Chicago

†All mistakes are my own. Views and opinions expressed are those of the authors and do not necessarily represent official positions or policy of the Office of Financial Research (OFR) or the U.S. Department of the Treasury. Please address correspondence to jeremiah.bejarano@ofr.treasury.gov. We thank seminar participants at the Office of Financial Research. We are grateful for feedback and discussions with Mark Carey, Reed Douglas, Melanie Friedrichs, Salil Gadgil, Corey Garriott, Francisco E. Ilabaca, William D. Larson, Mark Paddrik, and Sriram Rajan. We thank Guanyu Chen, Raiden Egbert, Bailey Meche, Om Mehta, Duncan Park, Kyle Parran, Raul Renteria, Kunj Shah, Fernando Urbano, and Haoshu Wang for their contributions.

‡Independent. Most of this work was completed as students in the Financial Mathematics program at the University of Chicago

[1]See the following examples. The Office of Financial Research's (OFR) Annual Report of 2022 states "The OFR has and will continue to monitor and analyze risks to financial stability, remaining agile to identify and examine emerging threats as they arise now and in the coming years." According to the Financial Stability Oversight Council's (FSOC) 2024 Annual Report, "The Systemic Risk Committee 'supports the Council's efforts in identifying risks and responding to emerging threats ... and has been using the Analytic Framework to identify and evaluate vulnerabilities and build consensus regarding risk priorities.' " (See https://home.treasury.gov/system/files/261/FSOC2024AnnualReport.pdf.) According to their the Federal Reserve's Financial Stability documentation, "The Federal Reserve maintains a flexible, forward-looking financial stability monitoring program to help inform policymakers of the financial system's vulnerabilities to a range of potential adverse events or shocks." (See https://www.federalreserve.gov/financial-stability/proactive-monitoring-of-markets-and-institutions.htm.) As another example, the Large Institution Supervision

our ability to forecast key financial metrics is not a theoretical exercise. Rather, it is critical for early warning signals and timely policy responses. By developing an open-source financial time series forecasting benchmark, we directly enhance the toolkit available for financial stability monitoring. This benchmark brings together a wide range of datasets—spanning asset returns, arbitrage (basis) spreads, and banking indicators—that are crucial for assessing vulnerabilities in different corners of the financial system. Importantly, it evaluates state-of-the-art "global" forecasting methods (which learn across many time series) and compares them to classical models, shedding light on which approaches best capture early signs of stress in these datasets. In sum, better forecasting can help regulators and market participants spot trouble on the horizon, supporting measures to safeguard the economy.

The time-series forecasting community increasingly recognizes the need for standardized evaluation frameworks. Benchmarks drive progress. Until recently, the absence of standardized benchmark datasets meant that most studies evaluated their methods on limited, arbitrarily selected time series or domain-specific data, making meaningful comparisons across methods virtually impossible. As Prater, Hanne, and Dornberger (2024) note, this lack of standardization has resulted in "poor quality evaluations" and irreproducible comparisons across forecasting studies. While several benchmark repositories have emerged to address this gap, they remain limited in their coverage of financial markets. The FRED-MD database (McCracken and Ng, 2016) provides a valuable collection of 134 U.S. macroeconomic series, but focuses primarily on real economic indicators rather than financial market data. The FinTSB benchmark (Hu et al., 2018) includes equity returns, though it does not use CRSP, widely regarded as the gold standard for high-quality equity market data. Meanwhile, the UCR Time Series Classification Archive (Dau et al., 2019) and UEA Multivariate Time Series Classification Archive (Bagnall et al., 2018) and MUU Extrinsic Regression Repository (Tan et al., 2020) do not include coverage of times series from the financial domain. The Monash repository (Godahewa et al., 2021) and recent efforts like TFB (Qiu et al., 2024) have made strides toward broader coverage, yet comprehensive representation of financial asset classes, from corporate bonds to credit derivatives, remains absent. (See Table 1 for a summary of existing benchmarks.) This gap is particularly problematic given that, according to the "No Free Lunch" Theorem (Wolpert and Macready, 1997), there is no single forecasting method that performs best for all time series. As such, to improve financial forecasting, we need to evaluate forecasting methods on a suite of datasets from the financial domain and, importantly, these datasets need to reflect the form of the data as they commonly appear in the relevant finance literature. That is, they should use the same data sources and the same domain-specific

Coordinating Committee (LISCC) was established "based on lessons learned from the 2007-09 global financial crisis that revealed deficiencies in how large, systemically important firms had been supervised. These lessons underscored the need for the supervision of the largest firms to be more forward-looking, consistent, and informed by analysis from multiple perspectives and disciplines". (See https://www.federalreserve.gov/supervisionreg/large-institution-supervision.htm.)

cleaning and normalization procedures established in the literature.

Our proposed benchmark creates a standardized, literature-compliant repository specifically designed for financial forecasting. By providing open-source scripts that automate the download and cleaning of data from institutional sources like WRDS and Bloomberg, we enable researchers to reproduce exact datasets used in canonical finance papers. Our aim is to fill a gap in quantitative finance research by providing the first comprehensive forecasting benchmark tailored to financial markets. In particular, our paper has the following main contributions:

- We introduce the first comprehensive time series forecasting repository focused specifically on financial markets. This archive provides standardized datasets across multiple asset classes and market segments:
  - Asset return series spanning equities, corporate bonds, U.S. Treasuries, foreign exchange rates, commodity futures, credit default swaps (CDS), and options
  - Arbitrage basis spreads including covered interest parity (CIP) deviations, CDS-bond basis, Treasury-swap spreads, TIPS-Treasury spreads, and Treasury spot-futures basis.
  - Specialized datasets critical for financial stability monitoring, including bank Call Report data, financial intermediary risk factors, and yield curve dynamics
- We provide validated implementations of canonical data cleaning procedures from seminal finance papers, each following the exact methodology of the original studies. Crucially, while these foundational papers often lack publicly available replication code, our implementations successfully reproduce their key results, establishing both the correctness of our data processing and creating a reusable framework for future research.
- We emphasize domain-specific data curation, recognizing that financial data requires specialized treatment reflecting market microstructure, regulatory requirements, and established academic conventions. Each dataset is processed using the precise subsample definitions, filters, and transformations specified in the corresponding literature, ensuring that forecasting evaluations reflect the actual data as used by experts rather than generic time series.
- We present comprehensive baseline forecasting results across all datasets using a suite of 12 methods ranging from classical statistical models (ARIMA, ETS, Theta) to modern machine learning approaches (CatBoost, neural networks) and state-of-the-art deep learning architectures (Transformer variants, N-BEATS, TimesFM). These baseline results enable standardized performance comparisons and establish benchmarks for future research.

- We provide novel insights into market-specific predictability by systematically evaluating forecasting performance across different asset classes and market segments. This analysis contributes to our understanding of where predictability exists in financial markets and can inform both academic research and policy decisions related to financial stability monitoring.
- All implementations are fully open source on GitHub[2] with best practices for reproducibility, including an automated data pipeline that pulls data from the appropriate data sources and cleans it according to the canonical methods established in the literature, and virtual environments ensuring perfect reproducibility across different computing environments.

By providing this resource, we enable the kind of standardized, apples-to-apples comparisons that have been lacking in financial forecasting research, while also contributing replication code for key papers and insights into predictability across financial markets.

Our own basesline forecasting estimates confirms two stylised facts. For asset returns—whether CRSP equities, TRACE corporate bonds, CDS, FX, or SPX options—forecasting remains extraordinarily difficult: even the best auto deep-learning models achieve out-of-sample R^2 values near zero, leaving the historical average as the practical standard. In contrast, machine-learning-based global models deliver economically meaningful gains for the basis spreads and supervisory indicators that underpin liquidity and funding surveillance. This illustrates why a domain-specific benchmark matters: the answer to "which model works best?" depends on the task. Thus, the choice of forecasting model might depend on, e.g., whether we are forecasting returns or detecting stress in funding and balance-sheet data.

Taken together, the datasets and baseline results enhance the infrastructure for developing the next generation of forecasting methods tailored to financial markets, with direct implications for risk management, asset allocation, and financial stability monitoring.

2 Background

2.1 Forecasting for Financial Stability and Many-Predictor Methods

Accurate time-series forecasting is central to macroprudential policy and supervision. Early-warning systems and composite risk indicators help authorities detect vulnerabilities with enough lead time to act (Oet et al., 2011). Supervisory stress tests also hinge on multi-quarter projections of earnings, losses, and capital ratios, underscoring the operational role of forecasting in setting buffers and calibrating policy tools.[3] Complementing firm-level exercises, broad cyclic risk gauges, such as the ECB's

[2] All code is open source and available at https://github.com/jmbejara/ftsfr.

[3] Federal Reserve, *Dodd-Frank Act Stress Test 2020: Supervisory Stress Test Framework and Model Methodology.* See https://www.federalreserve.gov/publications/

Table 1: Existing Time Series Forecasting Benchmark Data Repositories

Source	Name	Coverage of Finance Domain	Website
McCracken and Ng (2016)	FRED-MD	US Macroeconomic Series	Link
Hu et al. (2018)	FinTSB	Equities Returns only	Link
Dau et al. (2019)	UCR Time Series Classification Archive	None	Link
Bagnall et al. (2018)	UEA Multivariate Time Series Classification Archive	None	Link
Tan et al. (2020)	Monash, UEA & UCR Time Series Extrinsic Regression Repository	None	Link
Godahewa et al. (2021)	Monash Time Series Forecasting Repository	Fred-MD is one component	Link
Bauer et al. (2021)	Libra	Anonymous data, unclear	Link
Qiu et al. (2024)	TFB	NN5 Bank Cash Withdrawals, Equity (NYSE/NASDAQ), Foreign Exchange Rates	Link
Aksu et al. (2024)	GIFT-EVAL	Anonymous data, unclear	Link

Source: Authors' analysis

cyclical systemic risk indicator (CSRI), show that parsimonious, transparent signals constructed from multiple sectors can forecast the likelihood and severity of crises several years ahead.[4]

A large literature shows that exploiting many predictors improves forecast performance when common factors drive co-movements across series. In approximate factor models, principal components (diffusion indexes) summarize large panels into a few latent indexes that deliver competitive, often superior, forecasts relative to small VARs and univariate benchmarks (Stock and Watson, 2002a,b, 2010). The Office of Financial Research's own Financial Stress Index (FSI) applies a closely related approach: a factor model that is essentially uses the first principal component of a broad, daily panel of market indicators (see Monin (2019) and (Bejarano, 2023)). Another example is the Systemic Assessment of Financial Environment (SAFE) early warning system monitors, by researchers from the Federal Reserve Bank of Cleveland, which integrates supervisory and market data to forecast episodes of systemic stress (Oet et al., 2011). Together, these results motivate evaluating modern, panel-based forecasting methods on financial-stability-relevant datasets.

2.2 Return Predictability and Asset Pricing

Furthermore, financial forecasting has a more broad importance. Return predictability is central to modern asset pricing. Traditional efficient market views held that prices primarily varied with expected dividend growth, implying little scope for forecasting returns. Subsequent evidence overturned this perspective: variation in price-dividend ratios corresponds almost entirely to changes in discount rates—expected returns—and not to expected cash flows (Cochrane, 2011). High valuations reliably

june-2020-supervisory-stress-test-framework-and-model-methodology.htm.

[4]Detken, Fahr, and Lang (2018), ECB Financial Stability Review (May 2018) Special Feature. See https://www.ecb.europa.eu/press/financial-stability-publications/fsr/special/html/ecb.fsrart201805_2.en.html.

precede periods of low subsequent returns across asset classes, including equities, bonds, currencies, credit, and real estate. This common pattern underscores discount-rate variation as the organizing principle of contemporary asset pricing research.

Related to the many-predictor methods discussed previously, Kelly and Pruitt (2013) show that cross-sectional valuation ratios contain even more forecasting power than aggregate predictors. Using a partial least squares framework, they extract latent factors from the cross-section of book-to-market ratios, yielding substantial out-of-sample predictive power for both market returns and dividend growth. Out-of-sample R^2 values reach 13% at the annual horizon, magnitudes rarely achieved in predictive regressions. Their approach demonstrates that cross-sectional information can sharpen aggregate forecasts, highlighting how high-dimensional predictor sets can be distilled into robust factors that capture time-varying risk premia.

Recent survey work extends these insights into the era of financial machine learning. Kelly and Xiu (2023) argue that asset prices are themselves forecasts, discounted expectations of future payoffs, making return predictability the central empirical task of asset pricing. Because the conditioning information set available to investors is vast and the functional form of return dynamics is ambiguous, machine learning methods are particularly well suited. Penalized linear models, tree-based methods, and neural networks can systematically extract predictive signals from large panels of firm-level and macroeconomic variables, often outperforming traditional specifications. This perspective bridges classical evidence on discount-rate variation with modern global forecasting approaches, emphasizing that return prediction is both statistically feasible and economically important. This paper seeks to aid research in this area by developing a standardized benchmark that enables rigorous comparison of global forecasting methods on financial time series data.

2.3 Global Time Series Forecasting Methods

When working with many variables observed across multiple time series, such as returns across hundreds of assets or risk indicators across different market segments, researchers face a natural panel data structure. Traditional time series approaches to such panels typically estimate separate time series models for each cross-sectional unit, potentially missing valuable information embedded in the cross-sectional dimension or estimate multivariate models that don't scale well to high dimensions. An alternative approach recognizes that these many variables often share common underlying drivers and can benefit from joint estimation while managing the high dimensionality of the data.

This insight motivates global time series forecasting methods, which treat the entire panel as a single modeling problem. Rather than fitting separate models to individual series, global models leverage the full cross-sectional and temporal structure simultaneously, learning patterns that are both series-

specific and shared across the panel. This approach naturally captures spillover effects and cross-series dependencies while providing a framework for forecasting all series within a unified statistical model.

The empirical success of this approach is well documented in major forecasting competitions. The M-competition series is one of the most popular time series forecasting competitions in the field (Makridakis et al., 1982; Makridakis and Hibon, 2000; Makridakis, Spiliotis, and Assimakopoulos, 2018, 2022). Notably, the winning approaches of recent M-competitions have consistently employed global forecasting strategies that train a single model across all series in the dataset. As Godahewa et al. (2021) observe, this pattern extends beyond the M-competitions: winners of the NN3 and NN5 Neural Network competitions and various Kaggle competitions have similarly leveraged global models to achieve state-of-the-art performance.

The advantages of global methods are particularly pronounced in financial applications. They can handle short series by borrowing strength from longer ones, provide robust parameter estimation through implicit regularization across series, and naturally capture spillover effects between markets. This enables them to learn common patterns while accounting for series-specific variations-especially valuable when forecasting related financial series that share underlying economic drivers. Recent implementations range from pooled regression approaches to sophisticated deep learning architectures, with many demonstrating superior performance over traditional univariate methods (Godahewa et al., 2021).

This paper contributes to the global forecasting literature by providing a suite of benchmark datasets specifically designed for financial markets. Our benchmark thus serves to improve the infrastructure for developing the next generation of forecasting methods tailored to the complexities of financial markets, potentially improving risk management, asset allocation, and financial stability monitoring.

2.4 Univariate vs Multivariate Forecasting and the Inclusion of Exogenous Variables

Following Godahewa et al. (2021), this paper focuses specifically on *univariate* global methods. Local methods fit a separate model to each individual time series, so parameters are estimated solely from the history of that series. Classical ARIMA specifications are a good example of a univariate local model. By contrast, global methods estimate a single model—with shared parameters or shared representation—over an entire panel of series. Deep learning architectures that pool all bond spreads into one training problem, or a pooled exponential smoothing model, are global because they borrow strength across cross-sectional units. Hybrid approaches also exist: a researcher might still run ARIMA on each series (local), but learn the hyperparameters or initial states from panel-wide principal

components, thereby moving toward a univariate global perspective.

Orthogonal to the local/global distinction is the choice between univariate and multivariate targets. Univariate models forecast each series separately, even if parameters are shared across series as in global pooling. Multivariate models, such as vector autoregressions (VARs) or state-space systems with multiple jointly estimated equations, explicitly model the evolution of several dependent series simultaneously so that the forecast for one variable can depend on the lags of others. A global multivariate model would combine both ideas, fitting one high-dimensional system across the entire panel, whereas the univariate global models considered here forecast each series individually using parameters learned collectively from all series.

Many forecasting algorithms can also ingest exogenous information. In the classical literature, ARIMAX and VARX extensions incorporate external regressors. Modern deep learning methods such as DeepAR or Temporal Fusion Transformers can accept static attributes (such as sector, rating, tenor for bonds) or contemporaneous macro variables as covariates. These exogenous variables can improve forecasts by providing leading indicators or by capturing heterogeneity that the time-series dynamics alone miss. Allowing such augmentations, however, requires additional modeling decisions about feature engineering, alignment, and availability across datasets.

To maintain clarity and comparability, the benchmark focuses on univariate global forecasting without any exogenous regressors. Each dataset (basis spreads, asset returns, other supervisory indicators) is modeled and evaluated in isolation, and the taxonomy (e.g., the one we will provide in Table 2) is meant purely to organize data access rather than to impose cross-panel interactions. Researchers interested in extending these models with multivariate structures or exogenous signals can do so. We leave this for future research

2.5 The Importance Of Open Source And Replicability

A critical component of this paper is the development of an open-source benchmarked data repository that can serve as a standardized evaluation framework for financial time series forecasting. While other benchmarks in machine learning provide datasets openly, financial data is typically subject to licensing restrictions and copyright limitations that prevent direct redistribution. To address this challenge, we create a reproducible analytical pipeline that streamlines the process of downloading, formatting, and cleaning financial data in a standardized way, making the entire workflow open source and accessible to the research community.

Finance faces an active debate about a potential "replication crisis." Jensen, Kelly, and Pedersen (2023) find that predictors' average strength is only 54% of the original strength outside the original sample, with 32% becoming insignificant. While much of the debate centers around data snooping,

some of it involves coding and similar such bugs. There have been several high profile retractions[5]. Despite disagreement on the crisis's extent, consensus exists at least on one way to help the situation: standardized, open source data infrastructure. Chen et al. (2022) demonstrate with their "Open Source Cross-Sectional Asset Pricing" project that transparent data construction, version control, and community validation can eliminate arbitrary researcher degrees of freedom. Such infrastructure prevents simple errors from invalidating years of research. For policymakers and financial regulators, standardized open source data helps them more effectively monitor systemic financial risks. This paper addresses this need by developing an open-source financial time series forecasting repository that standardizes datasets across multiple asset classes and market segments, enabling reproducible research and evidence-based guidance for financial stability monitoring.

3 Benchmark Datasets

In this section, we describe the datasets included in the benchmark. The datasets are organized into three main groups: asset class datasets, basis spread datasets, and other financial data. We provide brief description of each dataset below. Each dataset is constructed based off of a cleaning procedure from a well-known paper in the academic literature. To validate our cleaning and transformations, we replicate a key plot or table from each paper. We give brief details of this process below. Full details of the cleaning and replications are provided in the appendix. The code that automates the data pull, cleaning, and formatting is available on GitHub.[6]

3.1 Dataset Selection Rationale

Our dataset selection is grounded in the intermediary asset pricing literature, which provides both theoretical justification and practical relevance for financial stability monitoring. Following He, Kelly, and Manela (2017) (HKM), we focus on assets that are primarily accessible to financial intermediaries rather than retail investors, who typically invest only in equities. This distinction is crucial because intermediary budget constraints create risk factors that drive pricing across asset classes, with stronger effects for assets that only intermediaries can trade—such as credit default swaps, options, corporate bonds, and commodity futures. These assets are therefore essential indicators for monitoring financial stability, as distress in intermediary balance sheets manifests first and most strongly in these markets.

Beyond the asset returns themselves, we include a comprehensive set of basis spreads following Siriwardane, Sunderam, and Wallen (2021), as these spreads serve as critical early warning indicators of stress in the financial system. When intermediaries face funding constraints or balance sheet pres-

[5] See Lee (2023) at Bloomberg, and Dickerson, Robotti, and Rossetti (2024).
[6] See `https://github.com/jmbejara/ftsfr`

sures, arbitrage opportunities persist and basis spreads widen, signaling potential systemic stress. Our methodological approach follows the principle established by He, Kelly, and Manela (2017) of using canonical cleaning procedures from seminal papers in each asset class, ensuring that our datasets reflect established best practices without reinventing data processing methods. This approach, combined with additional financial stability indicators such as the HKM intermediary factors and bank regulatory data, provides comprehensive coverage of the assets and indicators most critical for understanding and predicting financial stability risks. Table 2 provides a comprehensive overview of all datasets included in our benchmark, organized by asset class and methodology.

Table 3 shows the specific data sources required for each dataset in our benchmark. The table illustrates the diverse range of financial data providers needed to construct the datasets.

Table 4 provides detailed statistics for each dataset in our benchmark, including entity counts, time series characteristics, and temporal coverage. The table shows substantial variation across datasets, with disaggregated series (individual bonds, stocks, options contracts) containing thousands of entities, while portfolio-level datasets typically contain 10-50 series. The cutoff dates indicate the train/test split boundary used in our forecasting evaluation, with most datasets providing substantial out-of-sample periods for robust model comparison.

3.2 Asset Class Datasets

Our repository provides comprehensive datasets across seven major asset classes, each cleaned according to canonical methods established in the academic literature. We replicate each of the asset classes examined in He, Kelly, and Manela (2017), which are commodities, corporate bonds, credit default swaps (CDS), equities, foreign exchange markets, options, and US government bonds (Treasuries). This paper references a seminal paper in each asset class and uses the data cleaning procedures from that paper to construct test portfolios within that asset class.

For each asset class, we provide both aggregated portfolio returns (matching the test portfolios used in He, Kelly, and Manela (2017)) and disaggregated security-level data. The disaggregated datasets apply identical filtering criteria, subsampling, and transformations as specified in the original papers, but preserve individual security information rather than aggregating into portfolios. To validate our data cleaning procedures, we replicate key summary statistics and cross-sectional patterns from each source paper cited within He, Kelly, and Manela (2017). For instance, our corporate bond portfolios match the credit spread patterns in Nozawa (2017) and our options portfolios reproduce the volatility risk premium patterns found in Constantinides et al. (2013). These validation exercises ensure that our standardized datasets faithfully represent the canonical cleaning methods established in the literature while providing a unified framework for cross-asset forecasting research. This approach

Table 2: Overview of Datasets in the FTSFR Benchmark

Dataset Name	Description	Citation
Returns Data		
CDS Contract	Monthly returns for individual CDS contracts. The definition of returns for CDS follows Palhares (2012).	Palhares (2012)
CDS Portfolio	Similar to contract-level, but aggregated into 20 CDS portfolios by tenor and credit quality	He, Kelly, and Manela (2017)
Commodity	Monthly returns for commodity futures	Yang (2013)
Corporate Bond	Monthly returns for individual corporate bonds from TRACE. Authors cleaning builds on Nozawa (2017).	Dickerson, Robotti, and Rossetti (2024)
Corporate Portfolio	Monthly returns for corporate bond portfolios by credit spread	Nozawa (2017)
CRSP Stock	Monthly stock returns from CRSP database	Fama and French (1993)
FF25 Size-BM	Daily Fama-French 25 portfolios: size and book-to-market	ibid.
FX	Daily foreign exchange returns vs USD	Lettau, Maggiori, and Weber (2014)
SPX Options Portfolios	Monthly returns for individual SPX option contracts	Constantinides et al. (2013)
Treasury Bond	Monthly returns for individual Treasury bonds from CRSP	Gürkaynak, Sack, and Wright (2007)
Treasury Portfolio	Monthly returns for Treasury bond portfolios by maturity	ibid.
Basis Spread Data		
CDS-Bond	Monthly CDS-bond basis spreads	Siriwardane, Sunderam, and Wallen (2021)
CIP	Monthly covered interest parity deviations	Du, Tepper, and Verdelhan (2018)
TIPS-Treasury	Monthly TIPS-Treasury basis spreads	Fleckenstein, Longstaff, and Lustig (2014)
Treasury-SF	Monthly Treasury-SF arbitrage spreads	Fleckenstein and Longstaff (2020)
Treasury-Swap	Monthly Treasury-Swap arbitrage spreads	Siriwardane, Sunderam, and Wallen (2021)
Other Financial Data		
Bank Cash Liquidity	Quarterly cash liquidity from call report data	Drechsler, Savov, and Schnabl (2017)
Bank Leverage	Quarterly leverage ratios from call report data	ibid.
BHC Cash Liquidity	Quarterly bank holding company cash liquidity	ibid.
BHC Leverage	Quarterly bank holding company leverage ratios	ibid.
HKM Daily Factor	Intermediary risk factors, including capital ratio, capital risk factor, value-weighted investment return, and leverage ratio squared	He, Kelly, and Manela (2017)
HKM Monthly Factor	Same as above, but monthly	ibid.
Treasury Yield Curve	Daily Nelson-Siegel-Svensson zero-coupon yields, 1-30 years	Gürkaynak, Sack, and Wright (2007)

Source: Authors' analysis

Table 3: Data Sources by Dataset

Dataset Name	Data Sources
Returns Data	
CDS Contract	S&P Global CDS (formerly Markit)
CDS Portfolio	ibid.
Commodity	Bloomberg Terminal
Corporate Bond	WRDS TRACE, following Open Source Bond Asset Pricing[a]
Corporate Portfolio	ibid.
CRSP Stock	Center for Research in Security Prices (CRSP)
CRSP Stock (ex-div)	ibid.
FF25 Size-BM	CRSP and Compustat, following Fama and French (2023)[b]
FX	Bloomberg Terminal
SPX Options Portfolios	OptionMetrics IvyDB
Treasury Bond	Center for Research in Security Prices
Treasury Portfolio	ibid.
Basis Spread Data	
CDS-Bond	WRDS TRACE, following Open Source Bond Asset Pricing; S&P Global CDS and RED Entity (formerly Markit)
CIP	Bloomberg Terminal
TIPS-Treasury	ibid.
Treasury-SF	ibid.
Treasury-Swap	ibid.
Other Financial Data	
Bank Cash Liquidity	WRDS Bank Regulatory Call Reports, following Drechsler, Savov, and Schnabl (2017)[c]
Bank Leverage	ibid.
BHC Cash Liquidity	ibid.
BHC Leverage	ibid.
HKM Daily Factor	CRSP and Compustat, following He, Kelly, and Manela (2017)[d]
HKM Monthly Factor	ibid.
HKM All Factor	ibid.
Treasury Yield Curve	Yield Curve Data from Board of Governors of the Federal Reserve System[e]

[a] See https://openbondassetpricing.com/

[b] See the Ken French Data Library, https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html and the instructions to replicate it with CRSP and Compustat in Fama and French (2023).

[c] See https://pages.stern.nyu.edu/~pschnabl/data/data_callreport.htm

[d] See https://asafmanela.github.io/data/

[e] See https://www.federalreserve.gov/data/nominal-yield-curve.htm

Source: Authors' analysis

Table 4: Dataset Statistics Summary

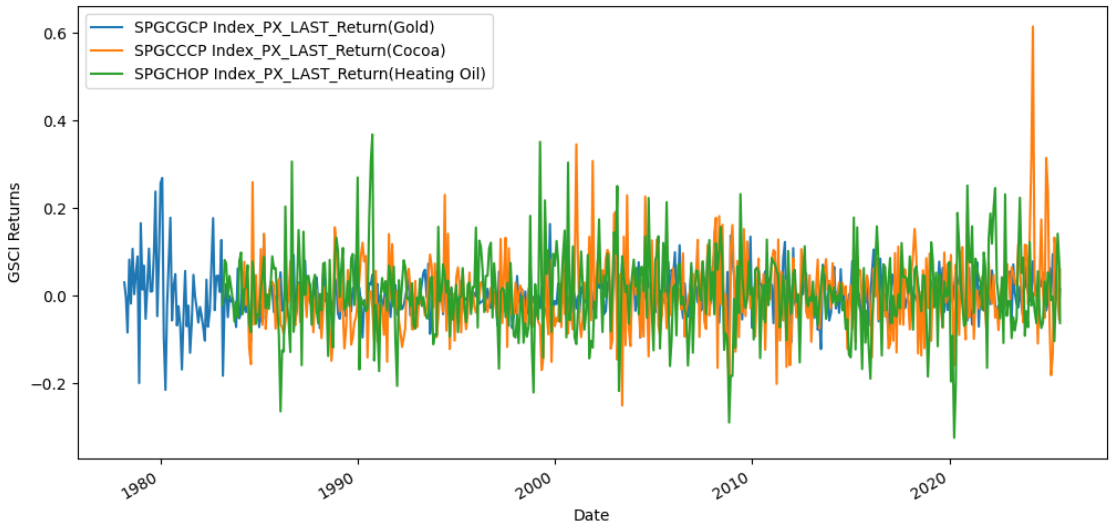
	Frequency	Unique Entities	Min Length	Median Length	Max Length	Min Date	Max Date
Basis Spreads							
CDS-Bond	Monthly	3402	1	16	169	2002-09-30	2022-09-30
CIP	Monthly	8	3997	5732	6030	2001-12-04	2025-02-28
TIPS-Treasury	Monthly	4	5126	5162	5197	2004-07-21	2025-05-30
Treasury-SF	Monthly	5	3783	5185	5192	2004-06-23	2025-01-08
Treasury-Swap	Monthly	7	1353	4482	6164	2001-12-20	2025-08-11
Returns (Portfolios)							
CDS Portfolio	Monthly	20	275	275	276	2001-01-01	2023-12-01
Corporate Portfolio	Monthly	10	242	242	242	2002-08-31	2022-09-30
FF25 Size-BM	Daily	25	26023	26023	26023	1926-07-01	2025-06-30
SPX Options Portfolios	Monthly	18	288	288	288	1996-01-31	2019-12-31
Treasury Portfolio	Monthly	10	659	666	668	1970-01-31	2025-08-31
Returns (Disaggregated)							
CDS Contract	Monthly	6552	1	25	96	2001-01-01	2023-12-01
CRSP Stock	Monthly	26757	1	85	1188	1926-01-30	2024-12-31
CRSP Stock (ex-div)	Monthly	26757	1	85	1188	1926-01-30	2024-12-31
Commodity	Monthly	23	283	511	668	1970-01-30	2025-08-12
Corporate Bond	Monthly	23473	1	36	242	2002-08-31	2022-09-30
FX	Monthly	9	4029	5991	6789	1999-02-09	2025-02-28
Treasuries	Monthly	2054	1	37	364	1970-01-31	2025-08-31
Other							
BHC Cash Liquidity	Quarterly	13770	1	46	177	1976-03-31	2020-03-31
BHC Leverage	Quarterly	13761	1	46	177	1976-03-31	2020-03-31
Bank Cash Liquidity	Quarterly	23862	1	66	177	1976-03-31	2020-03-31
Bank Leverage	Quarterly	22965	1	67	177	1976-03-31	2020-03-31
HKM All Factor	Monthly	4	516	516	516	1970-01-01	2012-12-01
HKM Daily Factor	Daily	4	4765	4766	4766	2000-01-03	2018-12-11
HKM Monthly Factor	Monthly	4	587	587	587	1970-01-01	2018-11-01
Treasury Yield Curve	Daily	30	9936	12230	16026	1961-06-14	2025-09-12

Sources: Bloomberg, Board Of Governors Of The Federal Reserve System, Center for Research in Security Prices, U.S. Call Reports, WRDS TRACE, OptionMetrics, S&P Global, Authors' analysis

allows researchers to both replicate existing studies (which typically use the aggregated portfolios) as well as to use the disaggregated data for richer analysis using the global time series forecasting methods.

Commodities Commodity futures follow the protocol of Yang (2013), but we are unable to access the same Commodity Research Bureau data that the authors used. Our replication therefore adopts the approach of Koijen et al. (2018), who provide Bloomberg tickers for the Goldman Sachs Commodity Index (GSCI) with directly computed monthly returns for 24 commodities. These GSCI-based return series constitute the entirety of our dataset, ensuring transparency and consistency with a well-documented and externally validated source. Because the GSCI returns are directly provided by Bloomberg, no further futures-chain construction or interpolation is required in our baseline analysis. An example of a few of the commodity futures returns are shown in Figure 1.

Figure 1: GSCI Commodity Returns



Sources: Bloomberg, Authors' analysis

Corporate Bonds We obtain corporate bond returns, we use data from the TRACE dataset, available via WRDS. We use the same cleaning procedure from the Open Source Bond Asset Pricing (OSBAP) project[7], which begins with the full TRACE transaction tape and then applies the market–microstructure–noise (MMN) correction procedure of Dickerson, Robotti, and Rossetti (2024). Using MMN–adjusted "clean" prices is essential: the bid–ask–averaged quotes in raw TRACE embed a mechanical reversal that can be mis-interpreted as illiquidity and lead researchers to overstate corporate bond excess returns. With the cleaned data we replicate and extend the ten value-weighted

[7]See https://openbondassetpricing.com/

credit-spread deciles of Nozawa (2017), sorting on option-adjusted spreads each month. Relative to unadjusted TRACE figures, the MMN correction eliminates roughly half of the apparent return spread and aligns liquidity estimates with quote-based ICE data, illustrating that much of the perceived illiquidity premium is an artefact of noise rather than compensation for trading frictions. The OSBAP procedure also furnishes auxiliary fields such as modified duration, amount outstanding, coupon rate, and matched-Treasury yields, which we exploit later for duration-matched excess-return calculations. We validate our cleaning procedure by matching the construction of corporate bond portfolios in He, Kelly, and Manela (2017).

Credit Default Swaps Our Credit Default Swap (CDS) data originate from S&P Global CDS database (formerly Markit), providing daily dealer-contributed curves, reference entity identifiers, and rich quote metadata for all USD-denominated contracts. Our CDS sample follows the constant-risky-duration construction of Palhares (2012), a commonly used protocol that neutralises maturity roll-over noise introduced by the 2009 "Big Bang" contract change. Raw Markit XR quotes are first filtered to exclude zero-bid or non-standard contracts, reconciled with auction recovery data, and rescaled by the risky annuity so that spreads are comparable across tenors. The procedure then interpolates missing maturities, align observations to common month-end fixing dates, and drop quotes that violate no-arbitrage bounds or sit outside the 1st-99th percentile of the cross-sectional spread distribution. Open-sourcing this transformation pipeline provides researchers with CDS excess-return series that are free of roll discontinuities, stale quote reversals, and documentation clause inconsistencies.

Equities For equities we adopt the same filters applied in Fama and French (1993). We begin with data from the Center for Research in Security Prices (CRSP) and Compustat by Standard & Poor's. These data sets are the gold standard for research in equity markets. The filtering methodology applies multiple layers of restrictions to ensure data quality and consistency with canonical equity research. We filter to the common stock universe restrict to U.S. incorporated firms with corporate issuer types, and limit to actively traded stocks on major exchanges (NYSE, AMEX, NASDAQ).

Foreign Exchange We provide daily returns from the individual foreign currencies, against the US dollar. The specified structure is based upon implied returns of the US dollar if converted to foreign currency then investing in the foreign currencies overnight repo rate (we use the interest rate).

Options Our monthly SPX options portfolio returns series follows the data cleaning and portfolio construction methodology of Constantinides et al. (2013). This framework forms forms the foundation of the approach used by He, Kelly, and Manela (2017) to construct the options portfolios used in

their paper. The Constantinides et al. (2013) portfolios are organized by option type (call or put), moneyness (9 levels), and maturity (3 levels), leading to $2 \times 9 \times 3 = 54$ distinct portfolios. The 18 portfolios in He, Kelly, and Manela (2017) were constructed by taking an equal weight average across the 3 maturities for CJS portfolios with the same moneyness, adding $2 \times 9 = 18$ distinct portfolios to the dataset, for a total of 72 distinct SPX option portfolios. Our dataset is comprised of monthly leverage-adjusted portfolio returns for all 72 SPX option portfolios, spanning 23 years from January 1996 to December 2019. The 72 portfolio returns series were constructed from a raw daily dataset of approximately 19 million individual SPX option contracts. These portfolios are constructed using leverage-adjustments and daily dynamic rebalancing to maintain constant risk exposures over time.

The leverage adjustments and dynamic portfolio rebalancing process outlined in Constantinides et al. (2013) had the overarching objective of constructing monthly portfolio returns that were roughly normally distributed over time, and only moderately skewed. We document our good faith reproduction of these procedures, and if a particular process was not sufficiently detailed in the original paper, we acknowledge these areas and made educated guesses about the authors' intent. For convenience, we provide the user with generalized functions that operate on any set of options data that is structured the same as the dataset we utilized (OptionMetrics).

We also provide a comprehensive overview of the data cleaning and preprocessing steps we undertook to prepare the raw SPX options data for analysis and portfolio construction, which includes technical and mathematical details on volatility estimation, kernel methods, and other relevant techniques. These details are given in the appendix as well as in the code module provided online.

Treasuries Our US Treasury bond portfolios utilize the CRSP Treasury database and the same filters used in the methodology of Gürkaynak, Sack, and Wright (2007), which underpins one of the various yield curve data publications available on the Federal Reserve Board's website[8]. At a high level, we keep only non-callable notes and bonds, strip out STRIPS/TIPS and other exotic issues, correct returns for accrued interest, and use strictly month-end transaction quotes. These filters remove optionality, stale prices, and auction-cycle noise that otherwise create spurious risk-premia.

3.3 Basis Spread Datasets

In well-functioning markets, basis spreads—deviations from classical no-arbitrage conditions like covered interest parity (CIP) or put-call parity—should be essentially zero. Persistent or large basis spreads signal that arbitrageurs are unable or unwilling to close riskless profit opportunities, often due to funding frictions or balance sheet constraints. Since the Global Financial Crisis (GFC), researchers

[8]See https://www.federalreserve.gov/data/yield-curve-models.htm

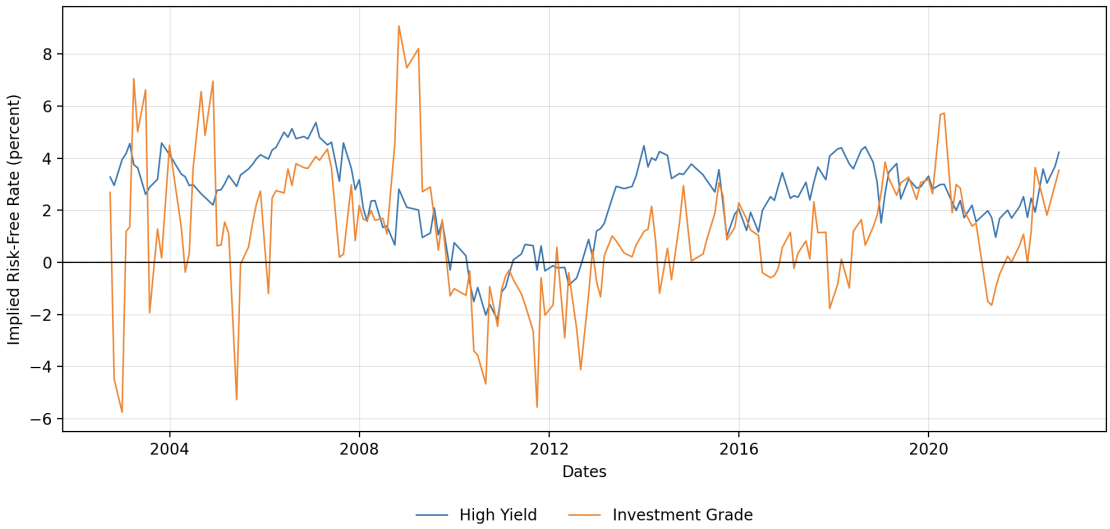
have documented several such anomalies across asset classes, and these have important implications for financial stability. For example, the Covered interest parity (CIP), once "the closest thing to a physical law in international finance," has been systematically violated since 2008 (Du, Tepper, and Verdelhan, 2018). During crisis episodes, these CIP deviations tend to spike sharply, revealing strains in global funding liquidity. For example, in 2008 and again in March 2020, the USD cross-currency basis for major currencies widened dramatically, indicating that foreign institutions were scrambling for dollar liquidity. (Board of Governors of the Federal Reserve System, 2020). Central banks and financial regulators closely monitor and respond to these metrics. And the Federal Reserve's swap lines are agreements with other central banks to exchange currencies, primarily to provide U.S. dollars to foreign institutions during times of market stress.

In this paper and its associated code repository, we provide code that will allow researchers to replicate a number of basis spreads across several different asset classes. We replicate many of the basis spreads in Siriwardane, Sunderam, and Wallen (2021), including the CDS-bond basis, covered interest parity (CIP), TIPS-Treasury basis, Treasury spot-futures basis, and Treasury-swap basis. Each of these basis spreads are, themselves, constructed following methodologies that are well-established in the literature. To validate out replication, we follow the papers recommended in Siriwardane, Sunderam, and Wallen (2021) and replicate key summary statistics from each source paper.

CDS-Bond Basis Leveraging the daily Markit pricing files used by Siriwardane, Sunderam, and Wallen (2021), we link cash bonds to matching single-name CDS and compute the basis after (i) retaining only USD-denominated senior unsecured issues with fixed coupons and 1-10-year maturities, (ii) discarding quotes with prices below 50¢, zero bids, or stale timestamps, and (iii) mapping each bond to a duration-matched CDS par spread via cubic-spline interpolation. We further exclude callable/convertible structures, winsorise extreme 100% bases, and require each bond to appear in our cleaned version of the public FINRA TRACE dataset to guarantee tradability, yielding a high-quality daily series that mirrors the investment-grade and high-yield bases reported in the original study. These spreads are shown in Figure 2.

Covered Interest Parity (CIP) We replicate the G10 series in Du, Tepper, and Verdelhan (2018), merging Bloomberg mid-quotes for spot rates, 3-month forwards, and maturity-matched OIS curves. Core filters rescale forward points (and invert USD-quoted pairs), synchronise all legs to the 17:00 ET close while dropping holiday or stale quotes, and winsorise the extreme 1% of annualised spreads with spline interpolation for any missing OIS tenors to remove scaling, timing, and rate-sourcing distortions. These spreads are shown in Figure 3.

Figure 2: CDS-Bond Basis spreads

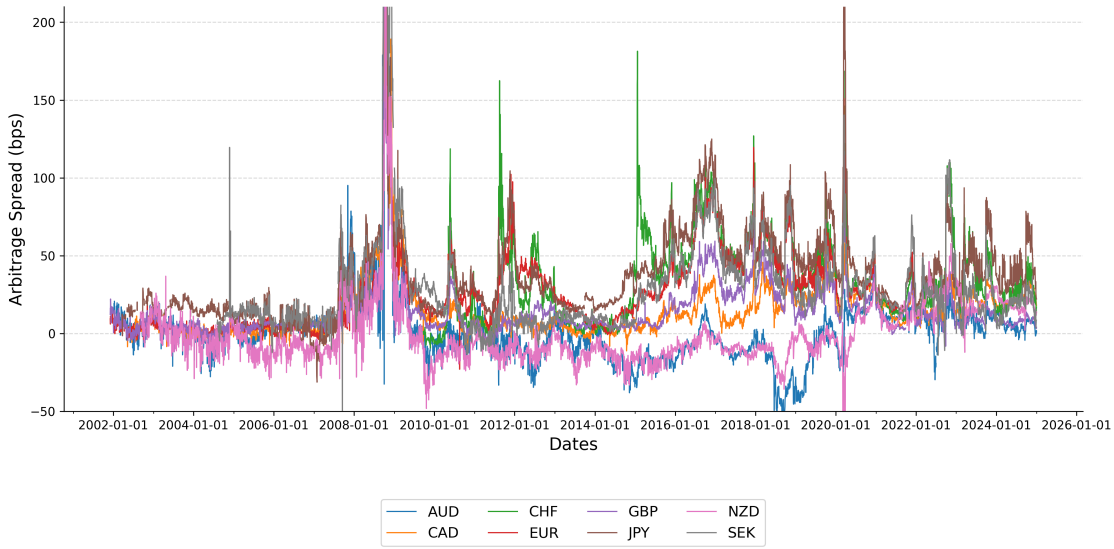


Sources: S&P Global, WRDS TRACE, Authors' analysis

TIPS-Treasury Basis Replicating Fleckenstein, Longstaff, and Lustig (2014) as implemented in Siriwardane, Sunderam, and Wallen (2021), we splice Federal Reserve zero-coupon TIPS yields with Bloomberg constant-maturity inflation-swap rates to create synthetic nominal yields, then difference these against equal-maturity zero-coupon Treasury yields for the 2-, 5-, 10- and 20-year tenors. We exclude days with missing swap quotes or illiquid TIPS issues, drop outliers beyond the extreme 1%, and harmonise all series to 17:00 ET closes to obtain a clean daily TIPS-Treasury basis series. These spreads are shown in Figure 4.

Treasury Spot-Futures Basis Following Fleckenstein and Longstaff (2020) as employed by Siriwardane, Sunderam, and Wallen (2021), we construct the Treasury spot-futures basis using Bloomberg cheapest-to-deliver implied repo rates for 2-, 5-, 10-, 20-, and 30-year Treasury futures contracts. For each tenor and date, we use only the first-deferred contract to avoid delivery option distortions documented by Burghardt et al. (2005). We compute days-to-maturity for each contract based on the last business day of the delivery month, then linearly interpolate OIS rates across available tenors (1-week through 1-year) to match the futures horizon. The basis is the difference between the futures-implied repo rate and the interpolated OIS rate. We restrict the sample to post-June-2004 observations, require positive trading volume in the deferred contract, and remove outliers using a rolling 45-day median absolute deviation filter with a threshold of 10 times the MAD. Missing values are forward-filled for up to 5 days. These spreads are shown in Figure 5.

Figure 3: Covered Interest Parity (CIP) spreads



Sources: Bloomberg, Authors' analysis

Treasury-Swap Basis We follow Siriwardane, Sunderam, and Wallen (2021) in constructing daily Treasury-Swap arbitrage spreads by pairing Bloomberg fixed-rate USD OIS quotes for tenors 1-, 2-, 3-, 5-, 10-, 20-, and 30-years with Bloomberg constant-maturity Treasury yields of identical maturities. For each tenor, the spread is computed as 100 times the difference between the swap rate and Treasury yield (Swap - Treasury), expressed in basis points. Observations with missing values are dropped, and the sample is restricted to 2000 onwards. The resulting series replicates the persistently negative swap spreads documented by Jermann (2020), Du, Hébert, and Li (2023), and Hanson, Malkhozov, and Venter (2023). These spreads are shown in Figure 6.

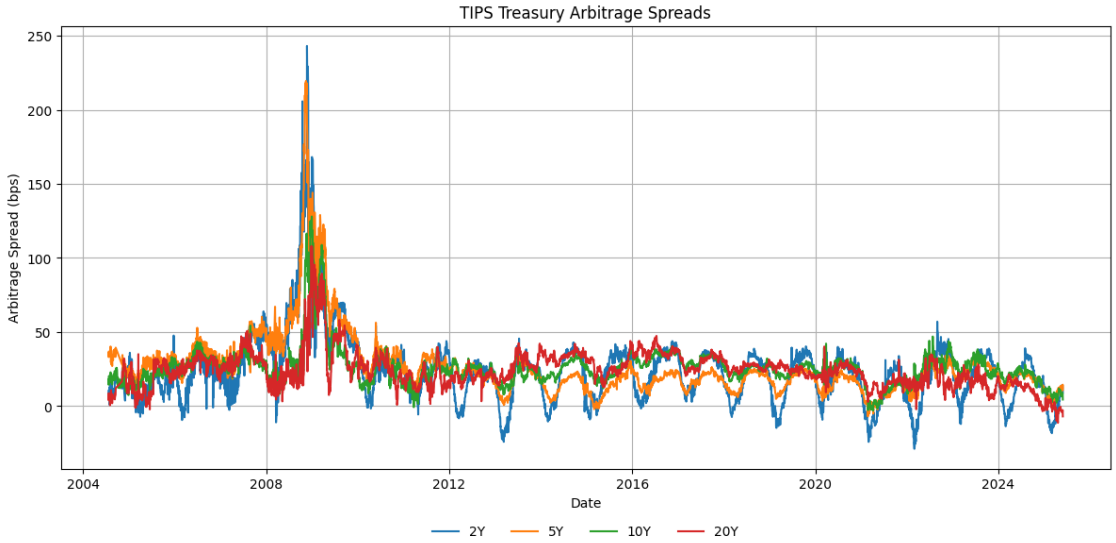
3.4 Other Financial Data

Call Report Bank variables come directly from the Drechsler-Savov-Schnabl Call Report panel, (Drechsler, Savov, and Schnabl, 2017). Their WRDS code downloads FFIEC 031/041 series and harmonizes item definitions across form changes (e.g., RCFD→RCON after 2011; time-deposit thresholds shifting from $100k to $250k in 2010), converts YTD income/expense to quarterly flows, handles pre-1982 semiannual filers, and builds consistent series for assets, loans, deposits, and interest expense. We use their released bank-level file (1976-2020) and construct our ratios (e.g., cash/liquidity, leverage).

Intermediary Capital Risk Factors We use the series released by He, Kelly, and Manela (2017)[9]. This is constructed from CRSP and Compustat data. He, Kelly, and Manela (HKM) construct the intermediary capital ratio for primary dealer holding companies as aggregate market equity divided

[9]See https://zhiguohe.net/data-and-empirical-patterns/intermediary-capital-ratio-and-risk-factor/

Figure 4: TIPS-Treasury Basis spreads



Sources: Bloomberg, Authors' analysis

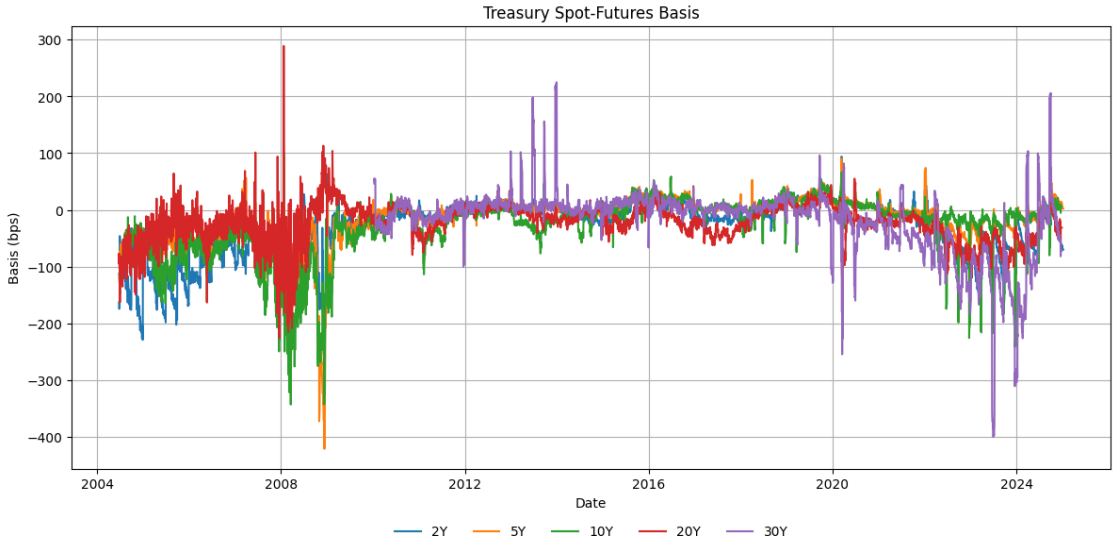
by aggregate market equity plus book debt, matching the dealer list to public holding companies in CRSP/Compustat. The "intermediary capital risk factor" is the AR(1) innovation in the capital ratio scaled by the lagged ratio. We simply align the published monthly (and post-2000 daily) series to our panel and do not re-estimate the factor.

Yield Curve We take the off-the-shelf Gürkaynak-Sack-Wright (GSW) nominal zero-coupon curve published by the Federal Reserve (Gürkaynak, Sack, and Wright, 2007). GSW fit a smooth Nelson-Siegel-Svensson curve to off-the-run Treasury notes and bonds (bills/FRNs and on-the-run issues excluded). The six-parameter Svensson model is used post-1980 and Nelson-Siegel before 1980. We simply reshape the released zero-coupon yields to our panel.

3.5 A Note about Potential Cross-Panel Interactions

We would like to note that these various datasets may also be combined to create richer forecasts. Information embedded in, say, covered interest parity (CIP) deviations could plausibly help forecast credit default swap (CDS) bond basis spreads, and liquidity stress in one asset class might foreshadow strains elsewhere. Similarly, macroeconomic announcements or regulatory indicators may serve as valuable covariates for certain panels while acting as leading indicators for others. In section 2.4, we explained that we while we acknowledge the opportunity to create richers forecasts in this way, forecasting methodologies that do so are outside of the scope of this paper. Consequently, the comparative results in Table 12 remain panel-specific by construction, even though the underlying data

Figure 5: Treasury Spot-Futures Basis (FTSFR Replication)



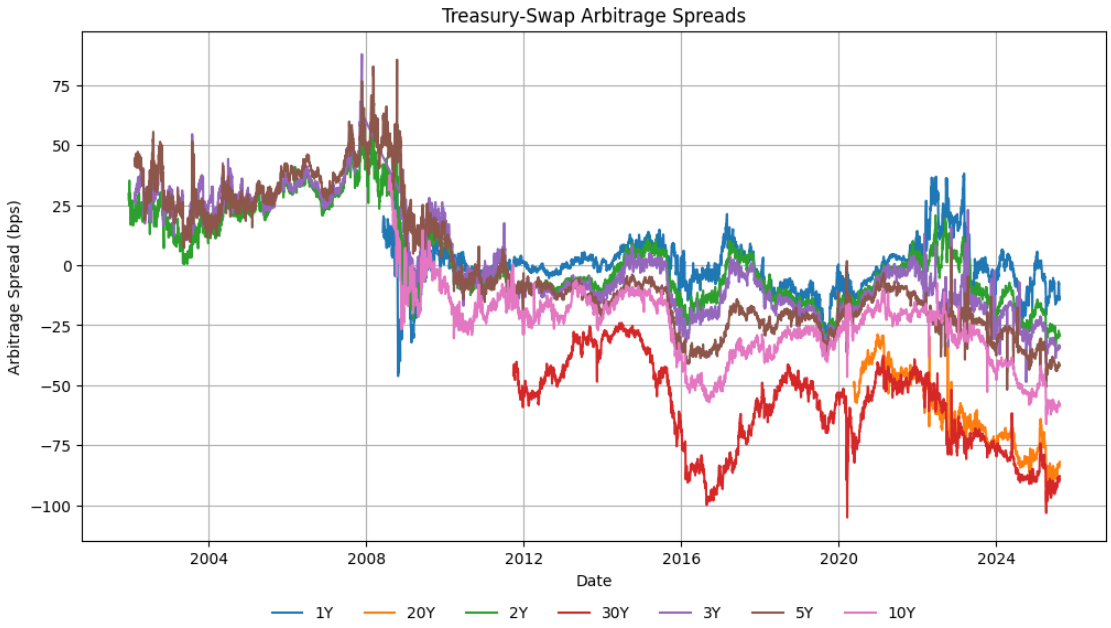
Sources: Bloomberg, Authors' analysis

resources support broader experimentation. Exploring these cross-panel dynamics and exogenous-variable augmentations is an important avenue for future work. Our goal in this release is to provide clean, panel-specific baselines. The modular data construction and standardized preprocessing allow researchers to layer on cross-panel pooling or covariate-informed architectures when tackling those broader questions. Out dataset design intentionally preserves the possibility of richer linkages and we hope that future research will address this.

4 Forecasting Methodology

In this section, we describe our forecasting methodology, including the baseline models used, data preprocessing procedures, and error metrics employed for evaluation. Our approach is designed to establish reliable baseline performance measures and provide fair comparisons across diverse financial time series. We first present our selection of forecasting models spanning classical statistical methods and modern machine learning approaches. We then detail our data preprocessing and filtering methodology, which ensures consistent treatment across all models and datasets. Finally, we define our comprehensive suite of error metrics, chosen to provide complementary perspectives on forecasting performance while enabling meaningful comparisons across different asset classes and time series characteristics.

Figure 6: Treasury-Swap Basis spreads



Sources: Bloomberg, Authors' analysis

4.1 Baseline Models

The models presented in this section are not intended to provide an exhaustive review of all available forecasting methods, nor are they estimated to their fully optimized specifications. Rather, we select a representative set of models from different areas of statistics and machine learning—including classical statistical methods, traditional machine learning approaches, and modern deep learning architectures—and apply them with minimal tuning to establish baseline forecasting performance. This approach allows us to demonstrate the utility of our benchmark datasets while providing reasonable comparative baselines that researchers can build upon. The primary emphasis of this paper is to provide high-quality, standardized benchmark datasets that others can readily use to evaluate their own forecasting models and methodologies. All classical statistical and modern machine learning models are implemented using the `statsforecast` and `neuralforecast` packages by Nixtla for Python, ensuring reproducible and standardized implementations.

Throughout, we rely on the *Auto* variants provided by Nixtla (e.g., `Theta`, `ARIMA`, `SES`, `NBEATS`, `NHITS`, `NLinear`). These wrappers automatically tune key hyperparameters using internal rolling-origin cross-validation, Bayesian search, or information-criterion-driven heuristics, and then refit the selected specification on the full training sample. Using the auto versions keeps the benchmark transparent—every model follows the same automated selection recipe—and mirrors realistic practitioner workflows where hand-tuning dozens of models across hundreds of series would be infeasible. Unless

otherwise noted, references to each model below therefore correspond to its Auto implementation.

- Theta (Theta): Splits a time-series into multiple Theta lines that are extrapolated separately and recombined, with the auto version searching over drift parameters and seasonal adjustments following Assimakopoulos and Nikolopoulos (2000).
- SES (Simple Exponential Smoothing): Automatically selects among simple, Holt, and Holt-Winters exponential smoothing specifications and their smoothing coefficients using information criteria, following Brown (2004) and Winters (1960).
- ARIMA: Autoregressive Integrated Moving Average model with automatic parameter selection, where the algorithm determines the optimal ARIMA(p, d, q) values through stepwise search and information criteria, following Box (2013) and Hyndman and Khandakar (2008).
- DeepAR: An autoregressive recurrent neural network (RNN) with Long Short-Term Memory (LSTM) cells, where the auto wrapper tunes the hidden size, dropout, and learning rate via Bayesian optimization, following Salinas et al. (2020).
- N-BEATS: A deep fully-connected architecture with forward and backward residual blocks; the auto procedure chooses stack types, block depth, and polynomial basis sizes, following Oreshkin et al. (2020).
- N-HiTS: Extends N-BEATS with multi-rate sampling and multi-scale interpolation, and the auto search allocates lookback windows, pooling sizes, and learning schedules, following Challu et al. (2022).
- DLinear: Implements seasonal-trend decomposition with linear layers and automatically selects window lengths and regularisation, following Zeng et al. (2022).
- NLinear: Applies the normalisation trick from Zeng et al. (2022) to handle distribution shifts, while the auto wrapper tunes the deseasonalisation and residual learning configuration.
- VanillaTransformer: A transformer-based architecture that utilises attention mechanisms to model dependencies in the input and output, with automatic tuning of heads, depth, and dropout, following Vaswani et al. (2017).
- TiDE: A time-series dense encoder with encoder/decoder blocks where the auto procedure searches over the number of layers, hidden units, and skip connections, following Das et al. (2023).

- KAN: Kolmogorov-Arnold Networks with spline-based activations that gain interpretability and accuracy; the auto routine selects spline granularity and network width, following Liu et al. (2025).

Table 5 summarises how these baselines differ across a set of practical design dimensions (Assimakopoulos and Nikolopoulos, 2000; Brown, 2004; Winters, 1960; Box, 2013; Hyndman and Khandakar, 2008; Salinas et al., 2020; Oreshkin et al., 2020; Challu et al., 2022; Zeng et al., 2022; Vaswani et al., 2017; Das et al., 2023; Liu et al., 2025). The comparison emphasises which models encode seasonality directly, which can absorb exogenous regressors, and which natively generate probabilistic forecasts. These distinctions may help interpret the performance tables in Section 5. These contrasts provide context for the error metrics we report in Tables 7 through 12.

Table 5: Key Properties of Baseline Forecasting Models

Model	Category	Seasonal	Exogenous	Prob.
Theta	Statistical	✓	–	–
SES/ETS	Statistical	✓	–	–
ARIMA	Statistical	✓	✓	✓
DeepAR	Deep Learning	–	✓	✓
N-BEATS	Deep Learning	✓	–†	–
N-HiTS	Deep Learning	✓	–	–
DLinear	Hybrid	✓	✓	–
NLinear	Hybrid	–	✓	–
Vanilla Transformer	Deep Learning	✓	✓	–
TiDE	Deep Learning	–	✓	–
KAN	Deep Learning	–	✓	–

Notes: "Seasonal" indicates built-in mechanisms for seasonal components; "Exogenous" denotes native support for additional regressors; "Prob." captures whether the implementation produces full predictive distributions rather than point forecasts. †N-BEATSx extends the architecture to exogenous inputs. Attributes are assessed from the original model publications and Nixtla documentation.[10]
Source: Authors' analysis.

4.2 Data Preprocessing

To ensure fair model comparisons, our forecasting system applies a unified preprocessing pipeline before any model is trained. Table 6 summarizes the effect of this pipeline, reporting entity counts, retention rates, median series lengths, and the adaptive minimum observation thresholds that each dataset must satisfy.

Filtering Methodology and Fairness Guarantees. The filtering rules are motivated by the different data appetites of statistical and neural models, but we apply them uniformly so that every estimator sees the same vetted panel. Classical methods such as Theta or SES can in principle run on short or irregular series, whereas the neural architectures rely on a meaningful train/validation split

and stable variance. Modern neural models are global, sharing parameters across thousands of entities, so they can borrow strength from cohorts of short series. Nevertheless, each individual series still needs enough history to contribute to the shared model without destabilising the split. Rather than maintain bespoke pipelines, we enforce a common set of quality screens: entity-level forecast horizons adapt to each dataset (with a minimum six-observation buffer for short panels), small datasets with ten or fewer entities are grandfathered to avoid wiping out entire categories, and the resulting filtered panel is passed unchanged to both StatsForecast and NeuralForecast implementations.

Technical Implementation. The pipeline first normalises timestamps (e.g., aligns month-end series to true month-ends) and builds a canonical grid using per-entity gap filling. We then split train/test windows per series using frequency-specific horizons: daily data forecast roughly a calendar month ahead (30 calendar or 21 trading days), monthly data forecast one month ahead, and quarterly data forecast one quarter ahead. The adaptive horizon feeds into the coverage requirements in `get_data_requirements`: monthly panels still need roughly 16 observations in total, but only a single hold-out point, while daily/business-day panels must deliver 21–30 observations in the test window. After splitting we record the original and retained entity counts, add gap indicators, and perform light forward-only imputation on the training slice; any series failing post-imputation validation is discarded. The resulting train/test slices form the single source of truth for every model family.

Forecast Horizons and Cross-Validation. All models—statistical and neural—are evaluated with rolling-origin cross-validation. The hold-out horizon matches the per-frequency targets above, with the step size set equal to the horizon. We allow up to six windows, but the actual count is capped by the shortest surviving series in each dataset; some monthly panels therefore supply six end-of-sample forecasts, whereas thin monthly or asset-level panels produce only one. The same window schedule is passed to both StatsForecast and NeuralForecast so baseline and neural estimates remain comparable, and the Auto wrappers reuse these folds during their internal hyperparameter searches. This design captures "one-month-ahead" predictability while avoiding prohibitively long evaluation windows that would exhaust shorter financial histories and provides repeated, out-of-sample checks that guard against overfitting.

Impact and Necessity. Portfolio-level aggregates are mostly unaffected. CIP spreads, TIPS/Treasury bases, and Treasury swap spreads retain 100% of the underlying series, while the CDS bond basis panel keeps 1,516 of 3,402 entities (44.6%). Monthly portfolio returns also remain intact apart from the CDS portfolio, which retains 4 of 20 entities (20%). Disaggregated panels bear the brunt of the length screens: individual CDS contracts shrink from 6,552 to 234 series (3.6% retention), corporate

bonds from 23,473 to 16,719 (71.2%), and Treasury bond securities from 2,054 to 1,912 (93.1%). Regulatory filings sit in between: BHC cash-liquidity series fall from 13,770 to 6,351 entities (46.1%), BHC leverage from 13,761 to 6,653 (48.3%), while bank-level leverage retains roughly three quarters (22,965 to 17,295; 75.3%). These filters strike a balance—protecting neural models from degenerate inputs, preventing some models from exploiting overly short histories, and keeping comparisons focused on genuine forecasting skill rather than artefacts of uneven preprocessing.

Table 6: Impact of Robust Forecasting Preprocessing on Dataset Statistics

	Frequency	Entities Before	Entities After	Retention (%)	Median Len Before	Median Len After	Date Range
Basis Spreads							
CDS-Bond	Monthly	3402	1516	44.6%	16	57	2002-09-30 – 2022-09-30
CIP	Monthly	8	8	100.0%	5732	272	2001-12-31 – 2025-02-28
TIPS-Treasury	Monthly	4	4	100.0%	5162	251	2004-07-31 – 2025-05-31
Treasury-SF	Monthly	5	5	100.0%	5185	247	2004-06-30 – 2025-01-31
Treasury-Swap	Monthly	7	7	100.0%	4482	207	2001-12-31 – 2025-08-31
Returns (Portfolios)							
CDS Portfolio	Monthly	20	4	20.0%	275	276	2001-01-31 – 2023-12-31
Corporate Portfolio	Monthly	10	10	100.0%	242	242	2002-08-31 – 2022-09-30
FF25 Size-BM	Daily	25	25	100.0%	26023	36160	1926-07-01 – 2025-06-30
SPX Options Portfolios	Monthly	18	18	100.0%	288	288	1996-01-31 – 2019-12-31
Treasury Portfolio	Monthly	10	10	100.0%	666	668	1970-01-31 – 2025-08-31
Returns (Disaggregated)							
CDS Contract	Monthly	6552	234	3.6%	25	54	2001-02-28 – 2023-12-31
CRSP Stock	Monthly	26757	25095	93.8%	85	94	1926-01-31 – 2024-12-31
CRSP Stock (ex-div)	Monthly	26757	25095	93.8%	85	94	1926-01-31 – 2024-12-31
Commodity	Monthly	23	23	100.0%	511	511	1970-01-31 – 2025-08-31
Corporate Bond	Monthly	23473	16719	71.2%	36	52	2002-08-31 – 2022-09-30
FX	Monthly	9	9	100.0%	5991	276	1999-02-28 – 2025-02-28
Treasuries	Monthly	2054	1912	93.1%	37	49	1970-01-31 – 2025-08-31
Other							
BHC Cash Liquidity	Quarterly	13770	6351	46.1%	46	69	1976-03-31 – 2020-03-31
BHC Leverage	Quarterly	13761	6653	48.3%	46	67	1976-03-31 – 2020-03-31
Bank Cash Liquidity	Quarterly	23862	17383	72.8%	66	82	1976-03-31 – 2020-03-31
Bank Leverage	Quarterly	22965	17295	75.3%	67	82	1976-03-31 – 2020-03-31
HKM All Factor	Monthly	4	4	100.0%	516	516	1970-01-01 – 2012-12-01
HKM Daily Factor	Daily	4	4	100.0%	4766	6918	2000-01-03 – 2018-12-11
HKM Monthly Factor	Monthly	4	4	100.0%	587	587	1970-01-01 – 2018-11-01
Treasury Yield Curve	Daily	30	30	100.0%	12230	17902	1961-06-14 – 2025-09-12

Sources: Bloomberg, Board Of Governors Of The Federal Reserve System, Center for Research in Security Prices, U.S. Call Reports, WRDS TRACE, OptionMetrics, S&P Global, Authors' analysis

4.3 Error Metrics

We evaluate forecasting performance using two complementary error metrics: Mean Absolute Scaled Error (MASE) and out-of-sample R^2 (R^2_{oos}). We use MASE because it is the standard accuracy measure in time series forecasting (Hyndman and Koehler, 2006), while R^2_{oos} is standard in finance for evaluating predictive models (Campbell and Thompson, 2008). Both metrics are scale-free and benchmark model performance against simple baselines, but they differ in their treatment of errors: MASE uses absolute errors (robust to outliers), while R^2_{oos} uses squared errors (emphasizing large misses).

Notation. Let $\{y_t\}_{t=1}^{T}$ be targets in the test window, $\{\hat{y}_t\}_{t=1}^{T}$ the corresponding forecasts, and $\bar{y}_{\text{train}}$ the training sample mean.

Mean Absolute Scaled Error (MASE).

$$\text{MASE} = \frac{\frac{1}{T}\sum_{t=1}^{T}|y_t - \hat{y}_t|}{\frac{1}{N-s}\sum_{t=s+1}^{N}|y_t - y_{t-s}|}$$

The denominator is the in-sample MAE of a seasonal naïve forecast on the training window of length N (with seasonal period s; s=1 for nonseasonal data). MASE values < 1 indicate the model outperforms the naïve benchmark; values > 1 indicate underperformance.

Relative MASE.

$$\text{Relative MASE} = \frac{\text{MASE}_{\text{model}}}{\text{MASE}_{\text{HA}}}$$

This compares each model's MASE to the Historic Average (HA) baseline on the same dataset. Values below 1.0 indicate a model improves upon the Historic Average, while values above 1.0 indicate weaker performance than the baseline.

Out-of-Sample R^2 (R^2_{oos}).

$$R^2_{\text{oos}} = 1 - \frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{T}(y_t - \bar{y}_{\text{train}})^2}$$

This measures the percentage reduction in MSE achieved by the model relative to predicting the training sample mean. Positive values indicate the model outperforms the historical average benchmark; zero or negative values indicate underperformance.

5 Baseline Results and Analysis

We present comprehensive forecasting results across all datasets and models using two primary error metrics: Mean Absolute Scaled Error (MASE) and Root Mean Square Error (RMSE). MASE provides scale-free comparison across different time series, while RMSE captures the magnitude of forecasting errors in original units.

Table 7 shows MASE results for all model-dataset combinations. The table is organized with datasets as rows and forecasting models as columns, allowing for easy comparison of model performance within each dataset and identification of datasets that present particular challenges for forecasting.

To better understand the relative performance of sophisticated models compared to the Historic Average baseline, Table 8 presents MASE ratios where each model's MASE is divided by the Historic Average model's MASE for the same dataset. Values less than 1.0 indicate that the model outperforms

Table 7: MASE Results by Dataset and Model

	HistAvg	ARIMA	SES	Theta	DLinear	DeepAR	KAN	NBEATS	NHITS	NLinear	TiDE	Transformer
Basis Spreads												
CDS-Bond	4.67	1.43	2.81	1.39	2.31	3.23	1.57	1.29	**1.27**	1.40	1.45	1.42
CIP	0.60	0.34	0.45	0.41	0.43	0.47	0.46	**0.26**	0.40	0.39	0.44	0.34
TIPS-Treasury	1.78	0.45	0.94	**0.37**	0.45	0.57	0.43	0.59	0.46	0.39	0.49	0.42
Treasury-SF	0.95	0.99	1.15	0.89	0.73	1.08	0.91	0.99	0.85	**0.71**	1.60	1.07
Treasury-Swap	2.63	0.29	0.69	0.30	**0.26**	0.39	0.45	0.30	0.33	0.29	0.32	0.36
Returns (Portfolios)												
CDS Portfolio	0.64	0.64	0.81	0.81	0.62	0.74	0.63	0.78	0.68	0.74	**0.56**	0.68
Corporate Portfolio	2.29	2.47	2.14	2.13	2.25	1.95	2.15	2.09	**1.77**	2.16	1.98	1.95
FF25 Size-BM	1.41	**1.41**	1.45	1.45	1.45	1.46	1.46	1.44	1.46	1.46	1.45	1.44
SPX Options Portfolios	1.44	0.72	0.70	0.73	0.69	0.95	0.69	**0.69**	0.70	0.75	0.84	0.71
Treasury Portfolio	0.52	0.66	0.56	0.55	**0.52**	0.53	0.73	0.58	0.80	0.55	0.55	0.55
Returns (Disaggregated)												
CDS Contract	1.90	1.81	1.77	**1.65**	1.73	1.80	1.68	1.69	1.71	1.80	1.71	1.69
CRSP Stock	0.87	0.90	0.89	0.91	0.88	1.21	0.87	0.89	0.87	0.92	0.88	**0.86**
CRSP Stock (ex-div)	0.87	0.90	0.89	0.91	0.87	1.15	0.87	0.87	0.87	0.91	0.88	**0.86**
Commodity	0.52	0.54	0.52	0.52	**0.51**	0.52	0.70	0.57	0.55	0.60	0.54	0.52
Corporate Bond	0.84	0.85	0.79	0.77	0.79	1.72	**0.74**	0.75	0.88	0.78	0.79	0.77
FX	18.49	**1.17**	7.21	1.86	1.27	12.47	3.43	1.72	1.45	1.25	1.26	3.12
Treasuries	0.41	0.35	0.29	0.25	0.29	0.33	0.25	**0.24**	0.25	0.25	0.27	0.26
Other												
BHC Cash Liquidity	1.82	0.84	1.14	0.82	0.83	1.73	0.82	**0.79**	0.79	0.84	0.83	0.85
BHC Leverage	2.66	0.99	1.62	0.97	1.48	4.18	0.96	0.95	**0.93**	1.06	1.02	1.00
Bank Cash Liquidity	1.83	0.82	1.09	0.80	0.80	1.09	0.84	0.78	**0.77**	0.81	0.92	0.80
Bank Leverage	3.19	1.35	2.04	1.33	1.61	2.99	1.42	**1.29**	1.30	1.36	1.41	1.34
HKM All Factor	2.64	0.75	1.02	**0.69**	0.84	0.93	1.16	1.33	0.81	0.84	0.84	0.92
HKM Daily Factor	7.31	1.46	**1.42**	1.45	1.47	1.52	1.64	1.57	1.49	1.49	1.48	1.50
HKM Monthly Factor	2.35	**0.43**	0.50	0.45	0.53	0.55	0.86	0.57	0.66	1.01	0.80	0.65
Treasury Yield Curve	11.26	0.77	1.09	0.76	0.84	1.20	0.85	**0.75**	0.76	0.84	0.92	0.83

Note: Values show Mean Absolute Scaled Error (MASE). Lower values indicate better performance. – indicates missing results.
Sources: Bloomberg, Board Of Governors Of The Federal Reserve System, Center for Research in Security Prices, U.S. Call Reports, WRDS TRACE, OptionMetrics, S&P Global, Authors' analysis

the Historic Average baseline, while values greater than 1.0 indicate underperformance. This normalization allows for clear interpretation of whether the additional complexity of advanced forecasting models provides meaningful improvements over the simple benchmark.

Figure 7 provides a visual representation of the relative MASE results using a diverging color scheme where blue indicates better performance than the Historic Average baseline (values < 1.0), white represents performance equal to the baseline, and red shows worse performance (values > 1.0). This visualization makes it easy to identify which models consistently outperform the simple baseline and which datasets benefit most from sophisticated forecasting approaches.

Table 9 presents the corresponding RMSE results using the same organization. While MASE provides scale-free comparisons, RMSE offers insights into the actual magnitude of forecasting errors, which can be particularly relevant for risk management applications.

Table 10 presents the out-of-sample R^2 (R^2_{oos}) results, providing a complementary perspective to the MASE and RMSE metrics. While MASE focuses on absolute scaled errors and RMSE captures error magnitudes, R^2_{oos} measures the percentage reduction in mean squared error achieved by each model relative to predicting the historical average. This metric is particularly valuable in finance as it directly quantifies predictive performance against the standard benchmark of using historical means for forecasting.

Figure 8 visualizes the out-of-sample R^2 results, highlighting which datasets and models deliver the

Table 8: Relative MASE Results by Dataset and Model

	ARIMA	SES	Theta	DLinear	DeepAR	KAN	NBEATS	NHITS	NLinear	TiDE	Transformer
Basis Spreads											
CDS-Bond	0.31	0.60	0.30	0.49	0.69	0.34	0.28	**0.27**	0.30	0.31	0.31
CIP	0.57	0.76	0.69	0.72	0.78	0.78	**0.44**	0.67	0.65	0.73	0.57
TIPS-Treasury	0.25	0.53	**0.21**	0.26	0.32	0.24	0.33	0.26	0.22	0.28	0.24
Treasury-SF	1.04	1.21	0.94	0.77	1.14	0.96	1.05	0.90	**0.75**	1.69	1.13
Treasury-Swap	0.11	0.26	0.12	**0.10**	0.15	0.17	0.11	0.13	0.11	0.12	0.14
Returns (Portfolios)											
CDS Portfolio	1.00	1.27	1.27	0.97	1.16	0.98	1.21	1.07	1.16	**0.87**	1.07
Corporate Portfolio	1.08	0.93	0.93	0.98	0.85	0.94	0.91	**0.77**	0.94	0.86	0.85
FF25 Size-BM	**1.00**	1.03	1.03	1.02	1.04	1.03	1.02	1.03	1.04	1.02	1.02
SPX Options Portfolios	0.50	0.49	0.51	0.48	0.66	0.48	**0.48**	0.49	0.52	0.59	0.49
Treasury Portfolio	1.27	1.07	1.05	**0.99**	1.02	1.41	1.12	1.54	1.06	1.05	1.06
Returns (Disaggregated)											
CDS Contract	0.95	0.93	**0.87**	0.91	0.95	0.88	0.88	0.90	0.95	0.90	0.89
CRSP Stock	1.03	1.02	1.04	1.01	1.39	1.00	1.02	0.99	1.05	1.00	**0.99**
CRSP Stock (ex-div)	1.03	1.02	1.04	0.99	1.31	0.99	1.00	1.00	1.04	1.01	**0.99**
Commodity	1.04	1.00	1.00	**0.98**	1.00	1.33	1.09	1.05	1.15	1.04	1.00
Corporate Bond	1.01	0.93	0.91	0.93	2.04	**0.88**	0.89	1.04	0.92	0.93	0.91
FX	**0.06**	0.39	0.10	0.07	0.67	0.19	0.09	0.08	0.07	0.07	0.17
Treasuries	0.86	0.71	0.60	0.70	0.81	0.62	**0.58**	0.60	0.61	0.65	0.65
Other											
BHC Cash Liquidity	0.46	0.63	0.45	0.46	0.95	0.45	**0.43**	0.43	0.46	0.46	0.47
BHC Leverage	0.37	0.61	0.37	0.55	1.57	0.36	0.36	**0.35**	0.40	0.38	0.38
Bank Cash Liquidity	0.45	0.60	0.44	0.44	0.59	0.46	0.43	**0.42**	0.44	0.50	0.44
Bank Leverage	0.42	0.64	0.42	0.50	0.94	0.45	**0.40**	0.41	0.43	0.44	0.42
HKM All Factor	0.28	0.39	**0.26**	0.32	0.35	0.44	0.50	0.30	0.32	0.32	0.35
HKM Daily Factor	0.20	**0.19**	0.20	0.20	0.21	0.22	0.22	0.20	0.20	0.20	0.21
HKM Monthly Factor	**0.18**	0.22	0.19	0.23	0.24	0.37	0.25	0.28	0.43	0.34	0.28
Treasury Yield Curve	0.07	0.10	0.07	0.07	0.11	0.08	**0.07**	0.07	0.07	0.08	0.07

Note: Values show MASE ratios relative to the Historic Average baseline. Values < 1.0 indicate better performance than the baseline. Numbers in bold indicate the best performing model for each dataset.
Sources: Bloomberg, Board Of Governors Of The Federal Reserve System, Center for Research in Security Prices, U.S. Call Reports, WRDS TRACE, OptionMetrics, S&P Global, Authors' analysis

Table 9: RMSE Results by Dataset and Model

	HistAvg	ARIMA	SES	Theta	DLinear	DeepAR	KAN	NBEATS	NHITS	NLinear	TiDE	Transformer
Basis Spreads												
CDS-Bond	2.81	1.31	1.84	1.26	1.65	2.09	1.42	1.23	**1.22**	1.27	1.27	1.26
CIP	12.74	8.01	9.89	9.28	9.88	10.22	10.60	**6.07**	9.54	8.92	9.62	7.39
TIPS-Treasury	22.78	6.58	13.18	**5.47**	6.59	8.04	6.94	8.33	6.54	6.34	7.00	6.02
Treasury-SF	42.23	42.25	51.26	39.92	32.99	46.04	38.75	40.95	37.92	**31.19**	67.34	48.22
Treasury-Swap	29.19	4.26	8.08	4.11	**3.82**	4.77	6.15	4.04	4.69	4.24	4.66	4.70
Returns (Portfolios)												
CDS Portfolio	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.00**	0.00
Corporate Portfolio	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	**0.03**	0.04	0.04	0.04
FF25 Size-BM	0.02	**0.02**	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
SPX Options Portfolios	0.02	0.02	0.02	0.02	0.02	0.02	**0.02**	0.02	0.02	0.02	0.02	0.02
Treasury Portfolio	0.00	0.01	0.01	0.01	0.01	**0.00**	0.01	0.01	0.01	0.01	0.01	0.01
Returns (Disaggregated)												
CDS Contract	0.02	0.02	0.02	0.02	**0.02**	0.02	0.02	0.02	0.02	0.02	0.02	0.02
CRSP Stock	0.21	0.21	0.21	0.21	0.21	0.26	0.21	0.21	0.21	0.22	0.21	**0.21**
CRSP Stock (ex-div)	0.21	0.21	0.21	0.21	0.21	0.25	0.21	0.21	0.21	0.22	0.21	**0.21**
Commodity	**0.05**	0.05	0.05	0.05	0.05	0.05	0.07	0.06	0.05	0.06	0.05	0.05
Corporate Bond	0.03	0.03	0.03	0.03	0.03	0.05	**0.03**	0.03	0.03	0.03	0.03	0.03
FX	1.33	0.08	0.47	0.14	0.08	0.97	0.30	0.13	0.11	**0.08**	0.08	0.25
Treasuries	0.00	0.00	0.00	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Other												
BHC Cash Liquidity	0.04	0.02	0.03	0.02	0.02	0.04	0.02	**0.02**	0.02	0.02	0.02	0.02
BHC Leverage	5.34	3.50	4.38	3.41	4.21	6.81	3.40	3.36	**3.35**	3.48	3.48	3.51
Bank Cash Liquidity	0.05	0.03	0.03	0.02	0.02	0.03	0.03	0.02	**0.02**	0.02	0.03	0.03
Bank Leverage	9.46	6.97	8.19	**6.84**	7.47	9.28	7.10	6.94	6.95	7.15	7.07	6.96
HKM All Factor	41.59	13.38	22.93	**12.04**	16.22	21.12	13.84	28.19	12.21	16.06	16.63	14.61
HKM Daily Factor	32.16	3.04	3.27	**2.97**	3.16	3.22	3.23	3.17	3.04	3.18	3.20	3.16
HKM Monthly Factor	47.30	**3.23**	3.80	3.23	3.69	4.37	12.55	4.79	6.74	8.70	7.04	4.50
Treasury Yield Curve	0.98	0.09	0.11	0.09	0.09	0.12	0.09	**0.08**	0.09	0.09	0.10	0.09

Note: Values show Root Mean Square Error (RMSE). Lower values indicate better performance. – indicates missing results.
Sources: Bloomberg, Board Of Governors Of The Federal Reserve System, Center for Research in Security Prices, U.S. Call Reports, WRDS TRACE, OptionMetrics, S&P Global, Authors' analysis

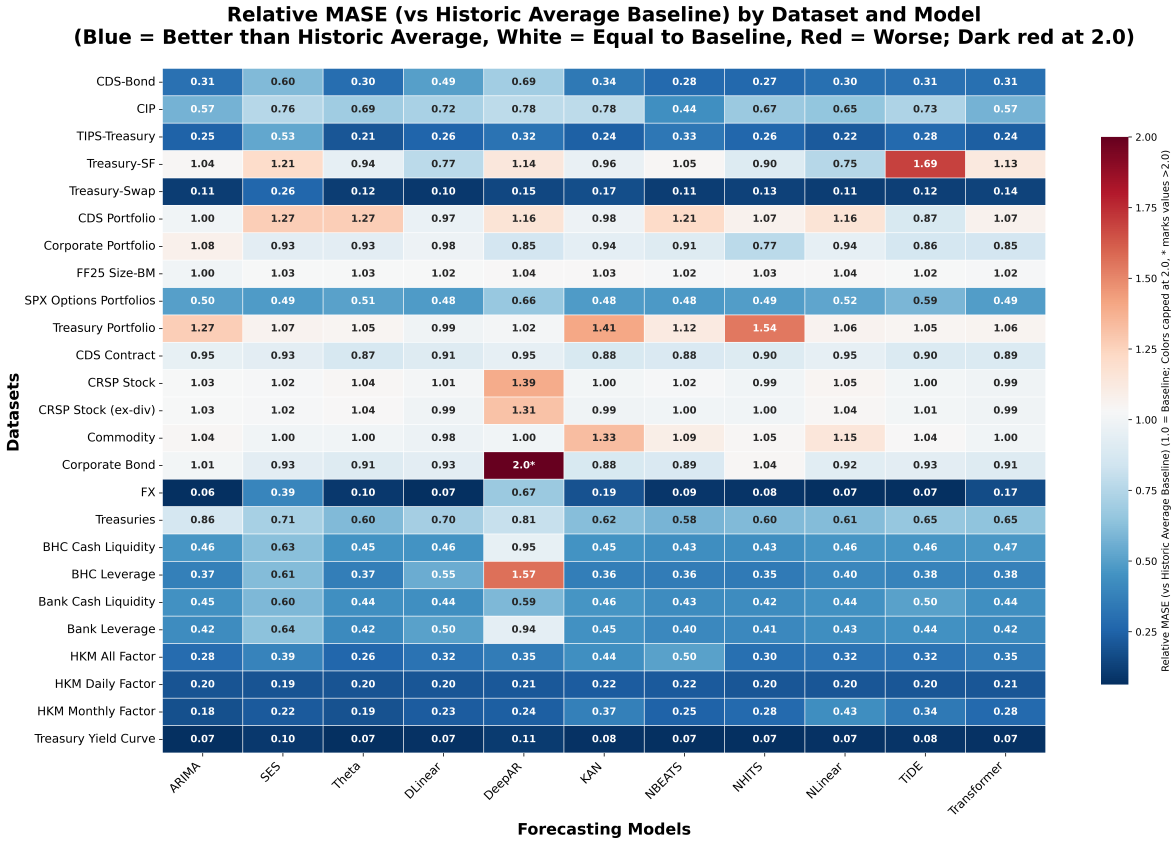


Figure 7: Relative MASE Results Heatmap by Dataset and Model. Values represent MASE ratios relative to the Historic Average baseline, with blue indicating better performance than the baseline (< 1.0), white indicating equal performance, and red indicating worse performance (> 1.0). The color scale is centered at 1.0 for optimal visual balance, and extreme outliers are marked with asterisks (*).

Sources: Bloomberg, Board Of Governors Of The Federal Reserve System, Center for Research in Security Prices, U.S. Call Reports, WRDS TRACE, OptionMetrics, S&P Global, Authors' analysis

Table 10: Out-of-Sample R^2 Results by Dataset and Model

	HistAvg	ARIMA	SES	Theta	DLinear	DeepAR	KAN	NBEATS	NHITS	NLinear	TiDE	Transformer
Basis Spreads												
CDS-Bond	0.00	0.54	0.02	0.54	0.13	-0.95	0.41	**0.61**	0.58	0.51	0.52	0.58
CIP	0.00	0.53	0.35	0.14	0.39	0.19	0.32	**0.71**	0.33	0.44	0.34	0.55
TIPS-Treasury	0.00	0.91	0.65	**0.94**	0.91	0.87	0.90	0.84	0.91	0.91	0.90	0.93
Treasury-SF	0.00	0.04	-0.65	0.07	0.43	-0.56	0.04	-0.34	0.01	**0.46**	-1.77	-0.16
Treasury-Swap	0.00	0.92	0.89	0.95	0.95	**0.96**	0.88	0.95	0.91	0.93	0.93	0.92
Returns (Portfolios)												
CDS Portfolio	0.00	-0.01	-0.30	-0.30	-0.05	-0.17	-0.05	-0.71	-0.19	-0.73	**0.13**	-0.32
Corporate Portfolio	0.00	-0.09	0.18	0.18	0.12	0.23	-0.16	0.09	**0.37**	0.10	0.23	0.20
FF25 Size-BM	0.00	**0.00**	-0.00	-0.00	-0.01	-0.01	-0.03	0.00	-0.02	-0.03	-0.01	-0.01
SPX Options Portfolios	0.00	0.54	0.43	0.39	0.53	0.15	0.54	**0.54**	0.48	0.46	0.28	0.49
Treasury Portfolio	0.00	-0.42	-0.03	**0.00**	-0.06	-0.02	-1.32	-0.12	-1.75	-0.11	-0.13	-0.28
Returns (Disaggregated)												
CDS Contract	**0.00**	-0.05	-0.12	-0.03	-0.29	-0.49	-0.09	-0.93	-0.12	-0.32	-0.20	-0.06
CRSP Stock	**0.00**	-0.19	-2.99	-4.26	-5.10	-27.92	-0.59	-2.14	-1.56	-34.73	-6.07	-0.84
CRSP Stock (ex-div)	**0.00**	-0.19	-3.04	-4.32	-1.57	-35.59	-1.06	-3.72	-3.41	-14.60	-3.65	-1.26
Commodity	**0.00**	-0.06	-0.02	-0.03	-0.00	-0.08	-1.00	-0.30	-0.23	-0.76	-0.16	-0.05
Corporate Bond	0.00	-0.37	-0.02	-0.08	-0.21	-29.04	-0.06	0.05	-2.84	-0.26	-0.29	**0.06**
FX	0.00	**0.98**	0.48	0.96	0.98	-0.79	0.64	0.97	0.98	0.98	0.98	0.92
Treasuries	0.00	-0.25	0.19	0.11	0.02	-0.07	-0.03	0.10	0.19	**0.20**	-0.02	-0.04
Other												
BHC Cash Liquidity	0.00	0.36	0.17	0.38	0.37	-2.75	0.34	**0.47**	0.44	0.34	0.38	0.39
BHC Leverage	0.00	0.57	0.14	0.58	0.01	-15.78	0.62	**0.63**	0.63	0.43	0.42	0.61
Bank Cash Liquidity	0.00	0.41	0.25	0.43	0.39	-0.19	-0.71	0.45	**0.45**	0.35	0.23	0.39
Bank Leverage	0.00	0.49	0.15	0.56	0.36	-2.94	0.61	**0.65**	0.62	0.44	0.40	0.63
HKM All Factor	0.00	0.35	0.33	0.40	0.45	0.27	-1.43	-1.58	-0.15	0.37	**0.47**	-0.69
HKM Daily Factor	0.00	0.48	0.48	0.48	0.45	0.43	**0.49**	0.43	0.48	0.44	0.47	0.48
HKM Monthly Factor	0.00	**0.50**	0.46	0.46	0.31	0.45	-0.14	-0.16	-0.00	-0.66	-0.35	0.25
Treasury Yield Curve	0.00	0.96	0.93	0.96	0.96	0.93	0.95	**0.96**	0.96	0.95	0.95	0.95

Note: Values show out-of-sample R^2 (R^2_{oos}). Positive values indicate better performance than the historical average baseline; negative values indicate underperformance. Higher values indicate better performance. Bold values highlight the best performing model for each dataset. *Sources: Bloomberg, Board Of Governors Of The Federal Reserve System, Center for Research in Security Prices, U.S. Call Reports, WRDS TRACE, OptionMetrics, S&P Global, Authors' analysis*

greatest gains over the historical average benchmark. The diverging color scale emphasizes positive predictive power while clearly flagging underperformance.

To synthesize performance across datasets, Table 11 reports median and mean statistics for MASE, Relative MASE, and out-of-sample R^2. Because MASE is the workhorse metric in the forecasting literature, we use it to anchor the discussion: `NBEATS` delivers the lowest median MASE, while `NHITS` dominates by mean MASE. Both auto-deep models handily outperform classical baselines such as the Historic Average, the seasonal naïve benchmark, and simple exponential smoothing (all of which sit well above 1.0 on Relative MASE). Their automated hyperparameter search concentrates capacity on recurring seasonal windows, which is particularly valuable for the high-frequency panels in our benchmark. When we pivot to R^2_{oos}, however, the ranking changes: `Theta` and `ARIMA` produce the strongest medians, highlighting that classical specifications still deliver the largest reductions in mean-squared error relative to the historical mean benchmark.

Table 12 disaggregates these results by dataset type. Each panel focuses on the models' performance across basis spreads, returns, or other datasets, highlighting how relative rankings shift as the data generating process changes and underscoring that the "best" model depends both on the metric and the market segment.

Error-metric choice is especially consequential for the returns category. Out-of-sample R^2 is the most informative measure here because the Historic Average is already a hard-to-beat benchmark for

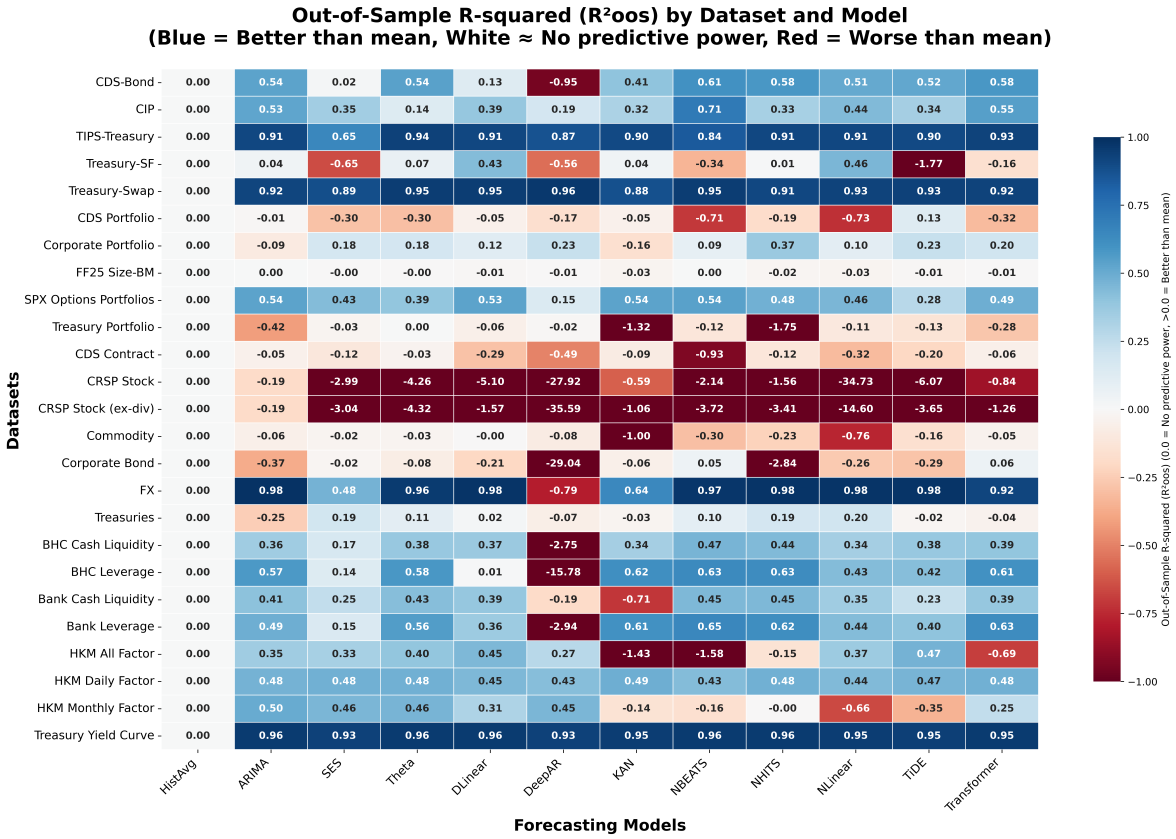


Figure 8: Out-of-Sample R^2 (R^2_{oos}) Heatmap by Dataset and Model. Warmer colors indicate larger improvements over the Historic Average baseline, while cooler colors indicate underperformance. Cells with extreme values are annotated to preserve readability.

Sources: Bloomberg, Board Of Governors Of The Federal Reserve System, Center for Research in Security Prices, U.S. Call Reports, WRDS TRACE, OptionMetrics, S&P Global, Authors' analysis

Table 11: Overall Model Performance Summary: MASE, Relative MASE, and R^2 Across All Datasets

	N	Med MASE	Mean MASE	Med Rel MASE	Mean Rel MASE	Med R^2	Mean R^2
HistAvg	25	1.819	2.957	.	.	0.000	0.000
ARIMA	25	0.841	0.933	0.498	0.622	0.359	**0.278**
SES	25	1.020	1.359	0.638	0.700	0.169	-0.043
Theta	25	0.814	0.927	0.505	**0.599**	**0.379**	-0.019
DLinear	25	0.829	0.977	0.554	0.606	0.307	0.019
DeepAR	25	1.148	1.791	0.851	0.837	-0.081	-4.514
KAN	25	0.861	1.063	0.481	0.641	-0.030	0.003
NBEATS	25	**0.786**	0.950	**0.477**	0.606	0.098	-0.062
NHITS	25	0.806	**0.912**	0.488	0.611	0.333	-0.077
NLinear	25	0.837	0.944	0.518	0.612	0.353	-1.755
TiDE	25	0.875	0.969	0.586	0.634	0.234	-0.200
Transformer	25	0.852	0.998	0.491	0.602	0.245	0.185

Note: Summary statistics across all datasets for each model. MASE shows absolute performance (lower is better), Relative MASE shows performance relative to Historic Average (lower is better), and R^2 shows out-of-sample predictive power (higher is better). Models are sorted by median MASE.
Sources: Bloomberg, Board Of Governors Of The Federal Reserve System, Center for Research in Security Prices, U.S. Call Reports, WRDS TRACE, OptionMetrics, S&P Global, Authors' analysis

Table 12: Model Performance by Dataset Category

Category	Model	N	Med MASE	Mean MASE	Med Rel MASE	Mean Rel MASE	Med R^2	Mean R^2
Basis Spreads	ARIMA	5	0.446	0.700	0.306	0.457	0.539	0.588
	DLinear	5	0.455	0.836	0.495	0.466	0.431	0.562
	DeepAR	5	0.567	1.147	0.693	0.615	0.186	0.102
	HistAvg	5	1.784	2.126	.	.	0.000	0.000
	KAN	5	0.465	0.766	0.337	0.497	0.405	0.509
	NBEATS	5	0.587	0.685	0.329	0.441	**0.706**	0.552
	NHITS	5	0.464	0.664	**0.272**	0.446	0.580	0.550
	NLinear	5	**0.395**	**0.637**	0.299	**0.407**	0.514	**0.651**
	SES	5	0.939	1.208	0.601	0.672	0.345	0.250
	Theta	5	0.411	0.672	0.297	0.449	0.544	0.529
	TiDE	5	0.491	0.862	0.311	0.627	0.517	0.185
	Transformer	5	0.421	0.723	0.305	0.475	0.581	0.563
Returns	ARIMA	12	0.873	1.035	1.003	0.902	-0.073	-0.009
	DLinear	12	0.827	0.988	0.974	0.837	-0.029	-0.469
	DeepAR	12	1.181	2.071	1.007	1.075	-0.125	-7.815
	HistAvg	12	0.873	2.519	.	.	**0.000**	**0.000**
	KAN	12	**0.803**	1.183	0.959	0.893	-0.073	-0.267
	NBEATS	12	0.825	1.024	0.955	0.858	-0.059	-0.513
	NHITS	12	0.869	1.000	0.996	0.881	-0.156	-0.675
	NLinear	12	0.845	1.015	0.992	0.876	-0.184	-4.149
	SES	12	0.853	1.501	0.965	0.899	-0.021	-0.437
	Theta	12	0.860	1.045	0.965	0.862	-0.015	-0.614
	TiDE	12	0.859	**0.975**	**0.913**	**0.833**	-0.077	-0.742
	Transformer	12	0.815	1.119	0.948	0.840	-0.043	-0.099
Other	ARIMA	8	0.830	0.926	0.328	0.305	**0.481**	0.514
	DLinear	8	0.839	1.049	0.378	0.346	0.380	0.411
	DeepAR	8	1.359	1.773	0.473	0.619	0.043	-2.446
	HistAvg	8	2.653	4.132	.	.	0.000	0.000
	KAN	8	0.912	1.069	0.404	0.353	0.413	0.090
	NBEATS	8	0.867	1.004	0.380	0.331	0.459	0.232
	NHITS	8	**0.796**	0.937	0.327	0.308	0.465	0.429
	NLinear	8	0.925	1.031	0.412	0.344	0.398	0.333
	SES	8	1.118	1.242	0.491	0.421	0.292	0.364
	Theta	8	0.811	**0.908**	**0.313**	**0.298**	0.474	**0.531**
	TiDE	8	0.918	1.028	0.362	0.341	0.414	0.372
	Transformer	8	0.888	0.988	0.363	0.326	0.437	0.375

Note: Metrics are computed within each dataset category (Basis Spreads, Returns, Other). Lower MASE/Relative MASE values indicate better performance; higher R^2_{oos} values indicate better performance.

asset returns. The table shows that nearly every model, including the auto deep networks, posts R^2_{oos} values clustered around zero for equity, bond, FX, and option returns. In practice this means that simply forecasting the historical mean is nearly optimal, an outcome that mirrors decades of empirical asset-pricing research and reinforces how difficult it is to extract persistent return predictability.

Basis spreads tell a different story. Across MASE and Relative MASE, `NLinear` and the multi-resolution `NHITS` and `NBEATS` variants dominate. These architectures excel because they decompose the series into trend and seasonal components while borrowing strength across entities through global training. Basis spreads exhibit pronounced seasonal structure (e.g., quarter-end funding pressure) and medium-term mean reversion, and the auto wrappers concentrate the search on those horizons. As a result, they deliver large improvements over the Historic Average and simple exponential smoothing baselines, widening the gap over classical contenders such as `Theta` or `ARIMA` on these spreads.

The "Other" category—which includes bank regulatory filings, volatility indicators, and macro-financial composites—leans back toward classical approaches. `Theta` delivers the strongest mix of low MASE and positive R^2_{oos}, with `NHITS` and `ARIMA` close behind. These datasets are typically longer and smoother than the high-volatility returns series but lack the strong seasonal spikes of basis spreads, so the flexible trend extrapolation in Theta and the parsimonious ARIMA dynamics fit naturally. The auto cross-validation also helps these simpler models stay calibrated without overfitting idiosyncratic shocks.

Taken together, the cross-metric evidence confirms that performance is conditional: MASE-based rankings highlight the strength of NBEATS/NHITS on error scales commonly used by practitioners, while out-of-sample R^2 favours Theta and ARIMA when the economic question is whether we beat the historical mean. Analysts selecting a forecast should therefore align the evaluation metric with the economic decision—R^2_{oos} for return forecasting where the historical mean is a credible trading benchmark, and scaled absolute errors for arbitrage spreads or supervisory indicators where seasonal accuracy matters more than mean-squared scorekeeping.

5.1 Analysis by Model Type

Traditional statistical models like Theta, SES, and ARIMA remain competitive across many datasets, particularly for univariate series with clear temporal patterns. These models offer computational efficiency and interpretability advantages that may outweigh marginal accuracy gains from more complex alternatives.

Among deep learning models, the results are mixed. Transformer-based models demonstrate strong performance on disaggregated datasets but struggle when data is limited. NBEATS and NHITS perform well on certain datasets but fail to consistently justify their additional computational costs. The

newer KAN architecture shows promise but requires further investigation to understand its strengths and limitations.

Interestingly, linear deep models like DLinear and NLinear achieve competitive results on many datasets despite their simplicity. This supports recent findings in the time series literature suggesting that complex architectures may be unnecessary for many forecasting tasks. The performance of these simpler models indicates that the inductive biases of traditional deep learning architectures may not align well with the characteristics of financial time series.

A notable pattern emerges in DeepAR's performance: the model exhibits relatively weak out-of-sample results across several datasets, particularly when compared to other deep learning architectures. This underperformance merits discussion in the context of our experimental design. All neural models in this benchmark use Nixtla's Auto implementations,[11] which automate hyperparameter selection through Bayesian optimization and related search strategies. The Auto wrappers—AutoDeepAR, AutoNBEATS, AutoNHITS, and others—systematically search over model-specific hyperparameters (e.g., hidden size, dropout, learning rate for DeepAR; stack types and block depth for NBEATS) using cross-validation on the training data. While these Auto methods employ sophisticated search algorithms and have been extensively tested on large-scale datasets, they cannot guarantee optimal performance for every individual dataset, particularly in the diverse and challenging landscape of financial time series.

The central objective of this paper is to provide a comprehensive benchmark of financial datasets that can be used by both practitioners and time-series forecasting researchers. We establish baseline results using off-the-shelf Auto methods from Nixtla to demonstrate feasible performance levels and to highlight which model families show promise on different types of financial data. However, we emphasize that achieving the absolute best forecast for any given dataset typically requires domain-specific expertise and careful manual tuning. Each model architecture embodies different inductive biases, and extracting maximum performance often demands understanding the specific characteristics of both the model and the data-generating process. For instance, DeepAR's autoregressive RNN structure with LSTM cells may require careful calibration of sequence length, distributional assumptions, and regularization strategies that generic hyperparameter searches might not fully explore. The benchmark's value lies not in claiming to provide optimal forecasts for every model, but in offering standardized datasets and reproducible pipelines that enable researchers to systematically investigate such questions. Future work can leverage these datasets to conduct focused hyperparameter sensitivity analyses, explore ensemble methods, or incorporate domain knowledge. However, these investigations fall outside the scope of establishing the set of datasets that form this benchmark.

[11] https://nixtlaverse.nixtla.io/

6 Conclusions and Future Work

This paper introduces an open-source benchmark specifically designed for financial time series forecasting research. By providing standardized datasets across multiple asset classes with reproducible data cleaning procedures, we address a critical gap in the forecasting literature and enable rigorous empirical comparison of forecasting methods on financial data.

Our work makes three primary contributions to the field. First, we provide a comprehensive financial forecasting benchmark covering seven major asset classes with both aggregated portfolio and disaggregated security-level data. Each dataset follows established academic cleaning procedures, ensuring consistency with canonical finance research while enabling modern forecasting methods that leverage cross-sectional information.

Second, we establish reproducible baselines using twelve forecasting models ranging from classical statistical methods to state-of-the-art deep learning architectures. Our results reveal that sophisticated models provide modest improvements of 3-10% over simple baselines, with gains concentrated in disaggregated datasets where cross-sectional learning is possible. These findings challenge the notion that complex models automatically yield superior performance in financial forecasting.

Third, we deliver open-source infrastructure that streamlines the entire pipeline from raw data acquisition to model evaluation. By making all code publicly available, we lower barriers to entry for researchers and enable rapid iteration on new forecasting methods. The modular architecture facilitates extension while maintaining compatibility with existing components.

The empirical results underscore that model choice must respect data generation processes. For asset returns, even the most sophisticated auto deep-learning architectures rarely beat the historical average; out-of-sample R^2 readings remain near zero, reaffirming how thin the exploitable signal is. Conversely, basis spreads and supervisory indicators exhibit structure that machine-learning-based global models can exploit: NBEATS, NHITS, and NLinear substantially improve MASE relative to DAR, historic averages, and classical smoothing, while Theta and NHITS capture persistent dynamics in bank liquidity and leverage ratios. Regulators interested in forecasting returns should therefore temper expectations, whereas those monitoring funding stress or balance-sheet metrics can extract tangible gains by deploying these global methods.

Acknowledgments

We would like to thank the following individuals. With their permission, we have adapted and used pieces of their code in this repository: Om Mehta and Kunj Shah for their replication of the Covered Interest Rate Parity (CIP) arbitrage spreads; Kyle Parran and Duncan Park for their replication of

commodity futures returns; Haoshu Wang and Guanyu Chen for their replication of the Treasury Spot-Futures basis; Arsh Kumar and Raiden Egbert for their replication of the Treasury Swap basis; and Bailey Meche and Raul Renteria for their replication of the TIPS-Treasury basis.

References

Adrian, Tobias, Daniel Covitz, and Nellie Liang. 2015. "Financial Stability Monitoring." *Annual Review of Financial Economics* 7 (Volume 7, 2015):357–395.

Aksu, Taha, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. 2024. "GIFT-Eval: A Benchmark For General Time Series Forecasting Model Evaluation." .

Assimakopoulos, V. and K. Nikolopoulos. 2000. "The theta model: a decomposition approach to forecasting." *International Journal of Forecasting* 16 (4):521–530.

Bagnall, Anthony, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. 2018. "The UEA multivariate time series classification archive, 2018." .

Bauer, André, Marwin Züfle, Simon Eismann, Johannes Grohmann, Nikolas Herbst, and Samuel Kounev. 2021. "Libra: A Benchmark for Time Series Forecasting Methods." *ICPE 2021 - Proceedings of the ACM/SPEC International Conference on Performance Engineering* :189–200.

Bejarano, Jeremiah. 2023. "The Transition to Alternative Reference Rates in the OFR Financial Stress Index." *Office of Financial Research Working Paper Series* 23-07.

Board of Governors of the Federal Reserve System. 2020. "Financial Stability Report, May 2020." Tech. rep., Board of Governors of the Federal Reserve System.

Box, George. 2013. "Box and Jenkins: Time Series Analysis, Forecasting and Control." In *A Very British Affair: Six Britons and the Development of Time Series Analysis During the 20th Century*. Palgrave Macmillan, London, 161–215.

Brown, RG. 2004. *Smoothing, forecasting and prediction of discrete time series*. Courier Corporation.

Burghardt, Galen D., Terrence M. Belton, Morton Lane, and John Papa. 2005. *The Treasury Bond Basis: An In-Depth Analysis for Hedgers, Speculators and Arbitrageurs*. McGraw Hill Library of Investment and Finance, 3rd ed.

Campbell, John Y. and Samuel B. Thompson. 2008. "Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?" *The Review of Financial Studies* 21 (4):1509–1531.

Challu, Cristian, Kin G. Olivares, Boris N. Oreshkin, Federico Garza Ramirez, Max Mergenthaler-Canseco, and Artur Dubrawski. 2022. "N-HiTS: Neural Hierarchical Interpolation for Time Series Forecasting." *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023* 37:6989–6997.

Chen, Andrew Y., Tom Zimmermann, Andrew Y. Chen, and Tom Zimmermann. 2022. "Open Source Cross-Sectional Asset Pricing." *Critical Finance Review* 11 (2):207–264.

Cochrane, John H. 2011. "Presidential Address: Discount Rates." *The Journal of Finance* LXVI (4):1047–1108.

Constantinides, George M, Jens Carsten Jackwerth, Alexi Savov, Gu¨nter Franke, Ben Golez, Bruce Grundy, Christopher Jones, Ralph Koijen, Stefan Ruenzi, and Amir Yaron. 2013. "The Puzzle of Index Option Returns." *The Review of Asset Pricing Studies* 3 (2):229–257.

Das, Abhimanyu, Weihao Kong, Andrew Leach, Shaan Mathur, Rajat Sen, and Rose Yu. 2023. "Long-term Forecasting with TiDE: Time-series Dense Encoder." *Transactions on Machine Learning Research* 2023.

Dau, Hoang Anh, Anthony Bagnall, Kaveh Kamgar, Chin Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Annh Ratanamahatana, and Eamonn Keogh. 2019. "The UCR time series archive." *IEEE/CAA Journal of Automatica Sinica* 6 (6):1293–1305.

Dickerson, Alexander, Cesare Robotti, and Giulio Rossetti. 2024. "Common pitfalls in the evaluation of corporate bond strategies." *SSRN Electronic Journal* .

Drechsler, Itamar, Alexi Savov, and Philipp Schnabl. 2017. "The Deposits Channel of Monetary Policy." *The Quarterly Journal of Economics* 132 (4):1819–1876.

Du, Wenxin, Benjamin Hébert, and Wenhao Li. 2023. "Intermediary balance sheets and the treasury yield curve." *Journal of Financial Economics* 150 (3):103722.

Du, Wenxin, Alexander Tepper, and Adrien Verdelhan. 2018. "Deviations from Covered Interest Rate Parity." *The Journal of Finance* 73 (3):915–957.

Fama, Eugene F. and Kenneth R. French. 1993. "Common risk factors in the returns on stocks and bonds." *Journal of Financial Economics* 33 (1):3–56.

———. 2023. "Production of U.S. Rm-Rf, SMB, and HML in the Fama-French Data Library." *SSRN Electronic Journal* .

Fleckenstein, Matthias and Francis A. Longstaff. 2020. "Renting Balance Sheet Space: Intermediary Balance Sheet Rental Costs and the Valuation of Derivatives." *The Review of Financial Studies* 33 (11):5051–5091.

Fleckenstein, Matthias, Francis A. Longstaff, and Hanno Lustig. 2014. "The TIPS-Treasury bond puzzle." *Journal of Finance* 69 (5):2151–2197.

Godahewa, Rakshitha, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. 2021. "Monash Time Series Forecasting Archive." .

Gürkaynak, Refet S., Brian Sack, and Jonathan H. Wright. 2007. "The U.S. Treasury yield curve: 1961 to the present." *Journal of Monetary Economics* 54 (8):2291–2304.

Hanson, Samuel Gregory, Aytek Malkhozov, and Gyuri Venter. 2023. "Demand-and-Supply Imbalance Risk and Long-Term Swap Spreads." *SSRN Electronic Journal* .

He, Zhiguo, Bryan Kelly, and Asaf Manela. 2017. "Intermediary asset pricing: New evidence from many asset classes." *Journal of Financial Economics* 126 (1):1–35.

Hu, Yifan, Yuante Li, Peiyuan Liu, Yuxia Zhu, Naiqi Li, Tao Dai, Dawei Cheng, Changjun Jiang, and Shu-tao Xia. 2018. "FinTSB: A Comprehensive and Practical Benchmark for Financial Time Series Forecasting Shu-tao Xia." *Proceedings of (Conference)* 1.

Hyndman, Rob J. and Yeasmin Khandakar. 2008. "Automatic Time Series Forecasting: The forecast Package for R." *Journal of Statistical Software* 27 (3):1–22.

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

Jensen, Theis Ingerslev, Bryan Kelly, and Lasse Heje Pedersen. 2023. "Is There a Replication Crisis in Finance?" *Journal of Finance* 78 (5):2465–2518.

Jermann, Urban J. 2020. "Negative Swap Spreads and Limited Arbitrage." *The Review of Financial Studies* 33 (1):212–238.

Kelly, Bryan and Seth Pruitt. 2013. "Market Expectations in the Cross-Section of Present Values." *The Journal of Finance* 68 (5):1721–1756.

Kelly, Bryan and Dacheng Xiu. 2023. "Financial Machine Learning." *Foundations and Trends® in Finance* 13 (3-4):205–363.

Koijen, Ralph S.J., Tobias J. Moskowitz, Lasse Heje Pedersen, and Evert B. Vrugt. 2018. "Carry." *Journal of Financial Economics* 127 (2):197–225.

Lee, Justina. 2023. "A Grad-School Number-Cruncher Shakes Up the World of Bond Quants."

Lettau, Martin, Matteo Maggiori, and Michael Weber. 2014. "Conditional risk premia in currency markets and other asset classes." *Journal of Financial Economics* 114 (2):197–225.

Liu, Ziming, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljači ´ Soljači ´c, Thomas Y Hou, and Max Tegmark. 2025. "KAN: Kolmogorov-Arnold Networks." .

Makridakis, S., A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler. 1982. "The accuracy of extrapolation (time series) methods: Results of a forecasting competition." *Journal of Forecasting* 1 (2):111–153.

Makridakis, Spyros and Michèle Hibon. 2000. "The M3-Competition: results, conclusions and implications." *International Journal of Forecasting* 16 (4):451–476.

Makridakis, Spyros, Evangelos Spiliotis, and Vassilios Assimakopoulos. 2018. "The M4 Competition: Results, findings, conclusion and way forward." *International Journal of Forecasting* 34 (4):802–808.

———. 2022. "M5 accuracy competition: Results, findings, and conclusions." *International Journal of Forecasting* 38 (4):1346–1364.

McCracken, Michael W. and Serena Ng. 2016. "FRED-MD: A Monthly Database for Macroeconomic Research." *Journal of Business and Economic Statistics* 34 (4):574–589.

Monin, Phillip J. 2019. "The OFR Financial Stress Index." *Risks 2019, Vol. 7, Page 25* 7 (1):25.

Nozawa, Yoshio. 2017. "What Drives the Cross-Section of Credit Spreads?: A Variance Decomposition Approach." *Journal of Finance* 72 (5):2045–2072.

Oet, Mikhail V., Ryan Eiben, Timothy Bianco, Dieter Gramlich, Stephen J. Ong, and Jing Wang. 2011. "SAFE: An Early Warning System for Systemic Banking Risk." *Working Paper* (11-29).

Oreshkin, Boris N, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio Mila. 2020. "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting." In *Eighth International Conference On Learning Representations*.

Palhares, D. 2012. "Cash-flow maturity and risk premia in cds markets." *Unpublished working paper* .

Prater, Ryan, Thomas Hanne, and Rolf Dornberger. 2024. "Generalized Performance of LSTM in Time-Series Forecasting." *Applied Artificial Intelligence* 38 (1).

Qiu, Xiangfei, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, Zhenli Sheng, and Bin Yang. 2024. "TFB: Towards Comprehensive and Fair Benchmarking of Time Series Forecasting Methods." *Proceedings of the VLDB Endowment* 17 (9):2363–2377.

Salinas, David, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 2020. "DeepAR: Probabilistic forecasting with autoregressive recurrent networks." *International Journal of Forecasting* 36 (3):1181–1191.

Siriwardane, Emil, Aditya Sunderam, and Jonathan Wallen. 2021. "Segmented Arbitrage." *Available at SSRN* 3960980.

Stock, James and Mark Watson. 2010. "Forecasting with Many Predictors." *Handbook of economic forecasting* 1 (August):1–5.

Stock, James H. and Mark W. Watson. 2002a. "Forecasting Using Principal Components From a Large Number of Predictors." *Journal of the American Statistical Association* 97 (460):1167–1179.

———. 2002b. "Macroeconomic forecasting using diffusion indexes." *Journal of Business and Economic Statistics* 20 (2):147–162.

Tan, Chang Wei, Christoph Bergmeir, Francois Petitjean, and Geoffrey I. Webb. 2020. "Time Series Extrinsic Regression." *Data Mining and Knowledge Discovery* 35 (3):1032–1060.

Vaswani, Ashish, Google Brain, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is All you Need." *Advances in Neural Information Processing Systems* 30.

Winters, Peter R. 1960. "Forecasting Sales by Exponentially Weighted Moving Averages." *Management Science* 6 (3):324–342.

Wolpert, David H. and William G. Macready. 1997. "No free lunch theorems for optimization." *IEEE Transactions on Evolutionary Computation* 1 (1):67–82.

Yang, Fan. 2013. "Investment shocks and the commodity basis spread." *Journal of Financial Economics* 110 (1):164–184.

Zeng, Ailing, Muxi Chen, Lei Zhang, and Qiang Xu. 2022. "Are Transformers Effective for Time Series Forecasting?" *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023* 37:11121–11128.

Appendices

A Data Cleaning Procedures and Replications

A.1 Asset Class Datasets

Following He, Kelly, and Manela (2017), we include comprehensive datasets across multiple asset classes, each cleaned according to canonical methods from the academic literature.

A.1.1 Equity Markets

The equity dataset implements the filtering methodology of Fama and French (1993) using CRSP's current CIZ format, providing a systematic approach to eliminate securities that would contaminate standard equity analysis. Our implementation applies multiple layers of restrictions: first, we filter to common stock universe using `sharetype=='NS'`, `securitytype=='EQTY'`, and `securitysubtype=='COM'`; second, we restrict to U.S. incorporated firms (`usincflg=='Y'`) with corporate issuer types (`issuertype` in ['ACOR', 'CORP']) to exclude foreign entities; third, we limit to actively traded stocks on major exchanges (NYSE, AMEX, NASDAQ) using `primaryexch` in ['N', 'A', 'Q'], `conditionaltype=='RW'`, and `tradingstatusflg=='A'`.

The following figures show summary statistics of the CRSP universe of stocks, including the average returns of all stocks for each given date (Figure 6) and the average three month rolling standard deviation for a given date (Figure 7). The earlier years of the dataset are dominated by small stocks, which have higher average returns and higher volatility, while the later years of the dataset is more diverse, leading to lower average returns and lower volatility.

A.1.2 US Treasuries

The US Treasury dataset follows Gürkaynak, Sack, and Wright (2007), whose methodology has become a common approach for constructing Treasury yield curves and returns. The results of using this methodology are published on the Federal Reserve Board's own website and underpins many studies in fixed income research.

Before estimating the yield curve, Gürkaynak, Sack, and Wright (2007) filter and clean the data in several steps. First, they restrict the sample to noncallable bonds and notes, excluding bills and other security types. Callable bonds embed optionality that contaminates pure interest rate risk measurement. For example, a bond trading at a premium with an embedded call option will exhibit negative convexity and compressed returns, distorting any analysis of term structure dynamics.

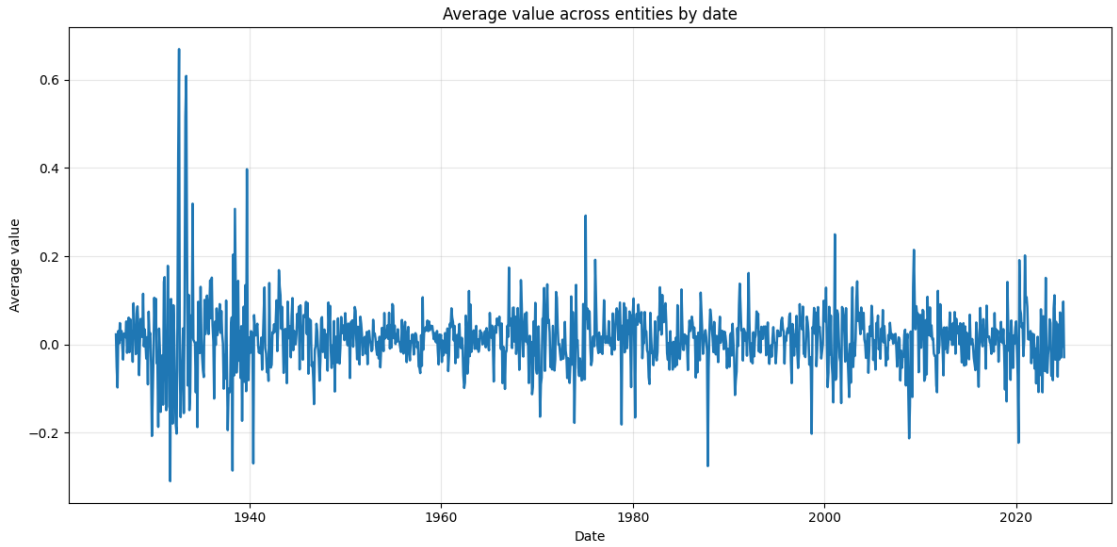


Figure 9: Average returns of all stocks for a given date

Second, the dataset focuses on bonds with remaining maturities between 0 and 5 years. This reflects market microstructure realities: bonds with longer maturities often suffer from illiquidity and fragmented trading, whereas the 0-5 year segment represents the most actively traded Treasury securities where price discovery is most efficient.

Third, the careful handling of accrued interest in return calculations is essential. Unlike equities, bond prices are quoted as clean prices (excluding accrued interest), but return calculations must incorporate coupon payments. Failing to account for this leads to dramatic miscalculations of returns, especially around coupon payment dates.

Fourth, a requirement for complete price and maturity information ensures that bonds with sporadic trading activity or ambiguous terms are excluded. This prevents noise from entering the yield curve estimation and return construction process.

Fifth, the use of month-end observations for portfolio aggregation avoids substantial day-of-month effects in Treasury markets. These effects arise from auction cycles, end-of-month portfolio rebalancing, and futures contract rolls, all of which can distort return measures if not properly controlled.

These filters, applied in the Gürkaynak, Sack, and Wright methodology, yield stable estimates that closely match dealer quotes and futures-implied yields. A comparison of the treasury bond returns dataset with the FTSFR dataset is shown in Figure 11.

A.1.3 Corporate Bonds

The corporate bond returns data is sourced from the TRACE (Trade Reporting and Compliance Engine) dataset, available at `openbondassetpricing.com`, and follows the methodology outlined in

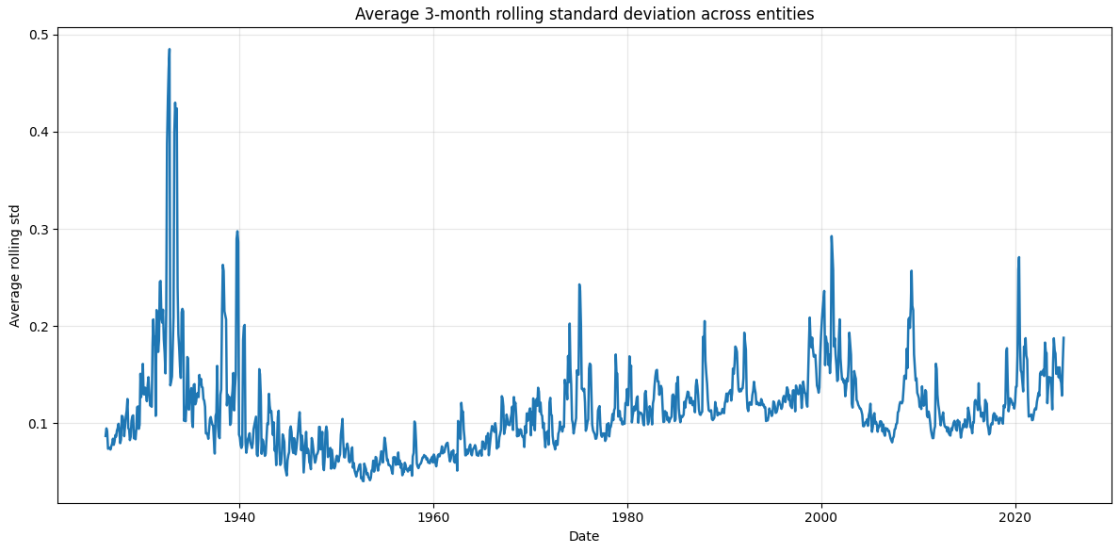


Figure 10: Average three month rolling standard deviation for a given date

Nozawa (2017). This dataset incorporates several key elements to ensure accuracy and relevance. First, it includes market microstructure adjustments to correct for noise and enhance the reliability of bond prices and returns. It also applies stringent data filters, such as including only U.S.-domiciled firms and excluding private placements and convertible bonds. The dataset covers corporate bonds with sufficient outstanding amounts and complete information, and includes essential fields such as MMN-adjusted clean prices, amount outstanding, monthly returns, and credit spreads.

The cleaning and standardization procedure adheres to the rigorous framework established by Nozawa (2017) and adopted by He, Kelly, and Manela (HKM). In terms of bond selection, floating rate bonds and those with put or convertible features are excluded, although callable bonds are retained. Bonds are removed if their prices exceed matched Treasury prices or fall below \$0.01 per \$1 face value. For return processing, return reversals are eliminated if the product of adjacent returns is less than -0.04, and monthly returns are computed to avoid assumptions about reinvestment. In synthetic Treasury construction, Treasury bonds with identical cash flow structures are constructed for each corporate bond using Federal Reserve constant-maturity yield data. Excess returns and credit spreads are then calculated in price terms rather than yield spreads.

For portfolio construction, bonds are sorted into deciles based on credit spreads for each date. Value-weighted portfolio returns are computed using bond values-defined as the product of MMN-adjusted clean price and amount outstanding-as weights. To ensure data quality, defaults are verified using Moody's Default Risk Service, while CRSP and Compustat data supplement the equity and accounting information. Callable bonds are also incorporated into regression models using fixed effects.

The final output is a dataset containing monthly corporate bond portfolio returns sorted by credit

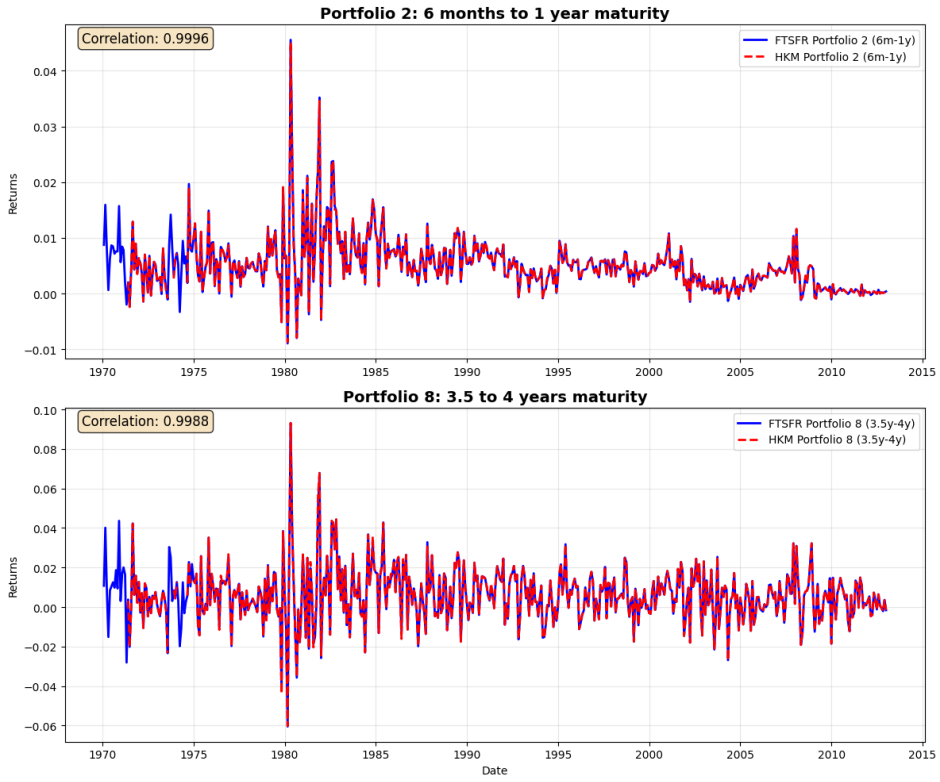


Figure 11: Comparison of He, Kelly, and Manela (2017) treasury bond returns dataset with FTSFR dataset

spread deciles, making it well-suited for credit risk analysis and for benchmarking against other bond market data. The constructed returns are validated against the dataset by He, Kelly, and Manela (2017), which serves as a benchmark for credit spread deciles. As shown in Figure 12, the time series comparison between the two datasets exhibits strong alignment in return patterns, especially during volatile periods such as the 2008 financial crisis.

A.1.4 Options

Our monthly SPX options portfolio returns series follows the data cleaning and portfolio construction methodology of Constantinides et al. (2013) ("CJS"), which has become the canonical approach for constructing option-based portfolios. This framework forms the foundation for numerous studies in derivatives pricing and risk management. He, Kelly, and Manela (2017) ("HKM") later adapted the CJS methodology to create a set of 18 option portfolios from the 54 portfolios in CJS. Our dataset includes both the original 54 CJS portfolios and the 18 HKM portfolios, for a total of 72 unique SPX option portfolios.

The original CJS paper used data from 1986 through 2012 (26 years of data). As of the time of writing, due to the unavailability of SPX option data from 1985 to 1995, we replicated the 54 CJS portfolios using data from January 1996 to December 2019 (23 years of data). As shown in Figure 13,

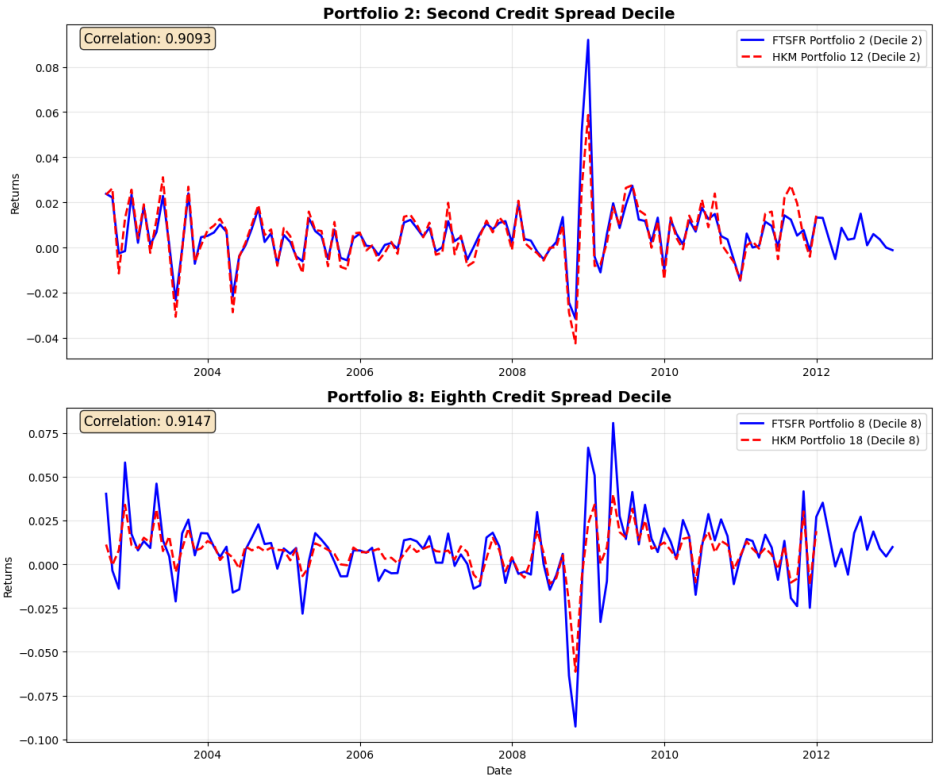


Figure 12: Comparison of He, Kelly, and Manela (2017) corporate bond returns dataset with FTSFR dataset

the number of SPX option observations has increased significantly over time, and since the volume of SPX options traded prior to 1996 was very low, we believe that our dataset from 1996 to 2019, while 3 years shorter is far richer in content and more relevant for current and future research.

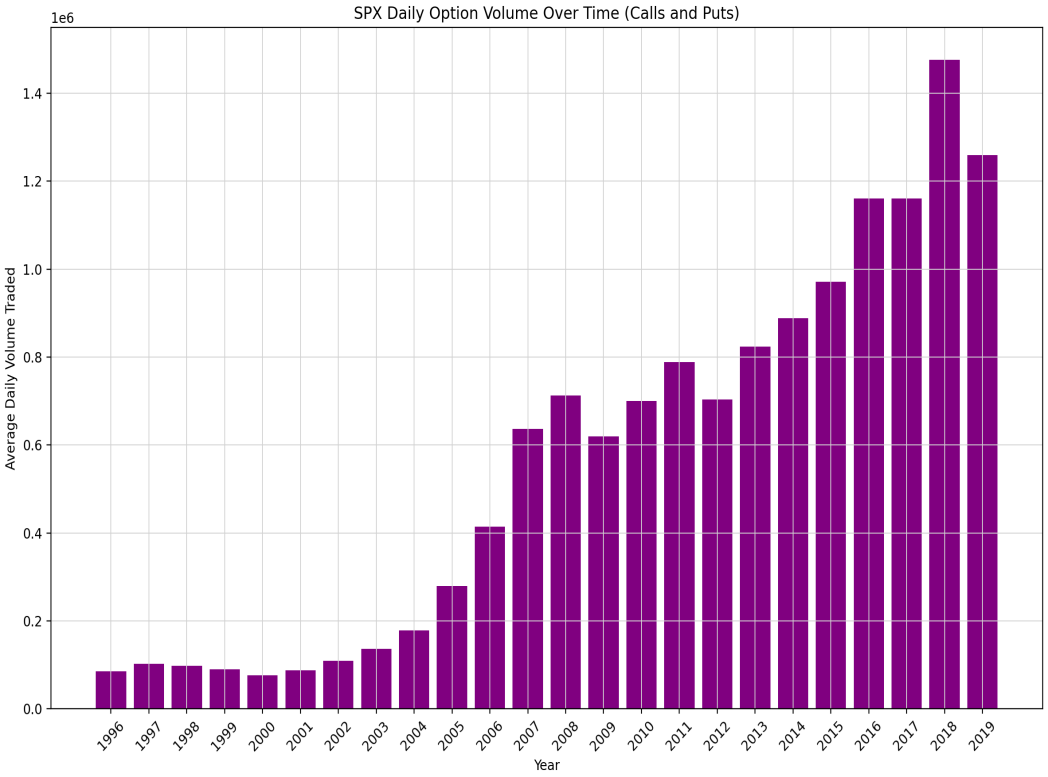
The process to construct our returns series involved two major phases: (1) Replicating with the highest practical fidelity the raw daily data filtration and monthly portfolio construction procedures outlined in Constantinides et al. (2013), and (2) Transforming these returns series into the 18 portfolios outlined in He, Kelly, and Manela (2017).

Options Data Series Access The final FTFSR data series comprise the monthly leverage-adjusted returns for call and put portfolios for both CJS 2013 (54 portfolios) and HKM 2017 (18 portfolios) for the period from Jan 1996–Dec 2019. Each portfolio is identified by a unique string of the form:

`{option type 'C' or 'P'}_{moneyness × 1000}_{maturity in days}`

Where "C" indicates a call option portfolio, "P" indicates a put option portfolio, "moneyness" is the strike price divided by the index price (e.g. 0.95, 1.00, 1.05), and "maturity in days" is the number of days to expiration (e.g. 30, 60, 90, 120, 150, 180). For example, the portfolio string "C_950_30" indicates a call option portfolio with moneyness of 0.95 and maturity of 30 days.

Figure 13: The SPX options market has grown dramatically over time



Sources: OptionMetrics, Authors' analysis

Data Cleaning Procedure To construct the SPX options portfolios, we implement a comprehensive cleaning procedure following Constantinides et al. (2013). The process applies three levels of filters to minimize quoting errors and remove anomalous options data. Level 1 filters eliminate duplicate quotes and zero-bid options. Level 2 filters restrict the dataset to options with 7-180 days to maturity, implied volatilities between 5% and 100%, and moneyness between 0.8 and 1.2. Level 3 filters remove volatility outliers and enforce put-call parity consistency.

Following filtration, portfolios are constructed using leverage-adjusted returns with daily dynamic rebalancing. Options are weighted using a bivariate Gaussian kernel in moneyness and maturity space, and returns are adjusted using Black-Scholes-Merton elasticity to maintain constant risk exposures over time. This procedure yields portfolio returns that are approximately normally distributed and only moderately skewed, making them suitable for empirical asset pricing tests.

The complete technical details of the cleaning procedure, including all filter specifications, mathematical formulations, and distributional analysis, are provided in Constantinides et al. (2013). Our implementation code is available in our GitHub repository, and a comprehensive methodological discussion is available in the online internet appendix.

A.1.5 Foreign Exchange

The Foreign Currencies returns series we generated uses spot rate changes and local repo rates to generate a USD-based foreign currency returns.

We are replicating a process where we convert our USD into the foreign currency i at end of day $t-1$, invest it in the overnight repo market, then switch the currency back to USD on day t.

$$ret_{t,i} = \frac{spot_{t-1,i}}{spot_{t,i}} \times fret_{t,i} \tag{1}$$

where

- i is the foreign currency
- t is the date of the implied foreign currency return
- ret is the return of USD invested in the foreign currency
- $fret$ is the return of the foreign currency when invested in their overnight repo market
- $spot$ is the spot price of the currency (how much 1 USD is worth in the foreign currency)

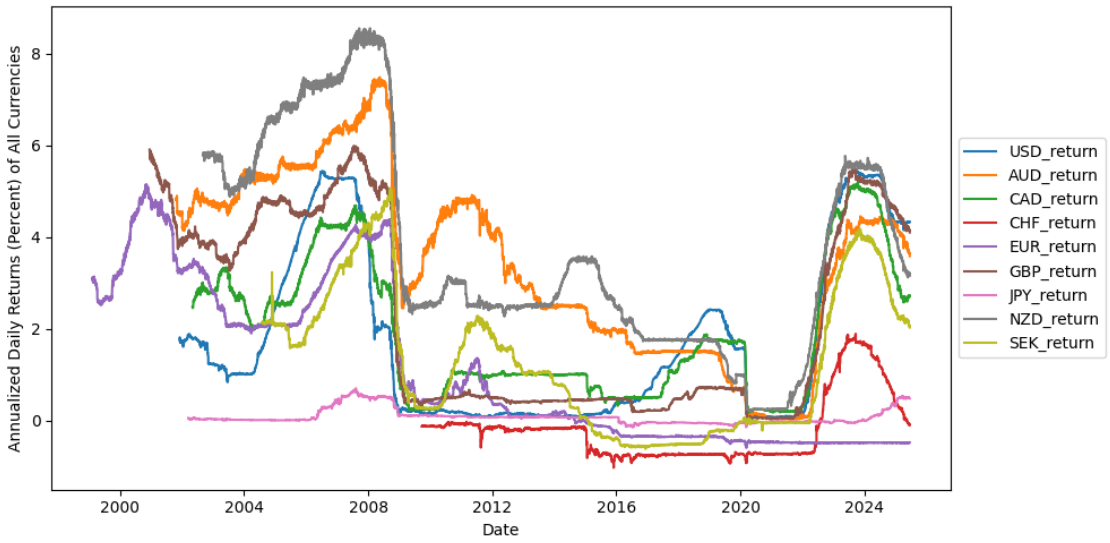


Figure 14: Foreign Currency Returns

A.1.6 Commodities

The commodity dataset follows Yang (2013), but in practice we adopt the implementation of Koijen et al. (2018), who provide Bloomberg tickers for the Goldman Sachs Commodity Index (GSCI). These GSCI indices have become the standard source for replication in commodity asset pricing, as they embed the official methodology for contract selection, roll schedules, and weighting across major futures markets. By relying on these pre-constructed Bloomberg series, we avoid the pitfalls of ad hoc roll choices or inconsistent maturity definitions that can bias commodity return factors.

Our processing of the GSCI data proceeds in several steps. First, we retrieve the designated Bloomberg excess return indices via the `xbbg` interface and store them as standardized parquet files. Second, we align all series to a monthly frequency by selecting the last observation in each calendar month and computing simple percentage returns. Third, we harmonize naming conventions and drop missing values, ensuring that each commodity return series is complete and comparable across the sample period. Finally, we compare these series with those published by He, Kelly, and Manela (2017). He, Kelly, and Manela (2017) publish returns series for 23 commodities from a broader pool of 31 commodities, but do not provide the names of the commodities nor the exact pool of commodities they used. To identify the commodities in their dataset, we computed the pairwise correlations between our Bloomberg GSCI-based returns and the corresponding series in He, Kelly, and Manela (2017). We then used the linear assignment algorithm to find the optimal one-to-one commodity matches, maximizing the total correlation between the two datasets.

The following figures illustrate the outcome of this replication exercise. Figure 15 reports the pairwise correlations between our Bloomberg GSCI-based returns and the corresponding series in He,

Kelly, and Manela (2017). While many commodities show very high alignment, several pairs exhibit unusually low correlations. We interpret these discrepancies as arising from ticker mismatches and differences in the underyling datasets used: the lists provided by Koijen et al. (2018) and Yang (2013) do not perfectly overlap, and it is unclear which exact subset was ultimately adopted in He, Kelly, and Manela (2017). Furthermore, He, Kelly, and Manela (2017) uses a different datasource than Bloomberg, which we use. In other words, low-correlation pairs likely reflect incorrect matches rather than genuine return differences. The accompanying heatmap in Figure 16 highlights this heterogeneity visually, with clusters of high-correlation matches alongside a handful of clear mismatches.

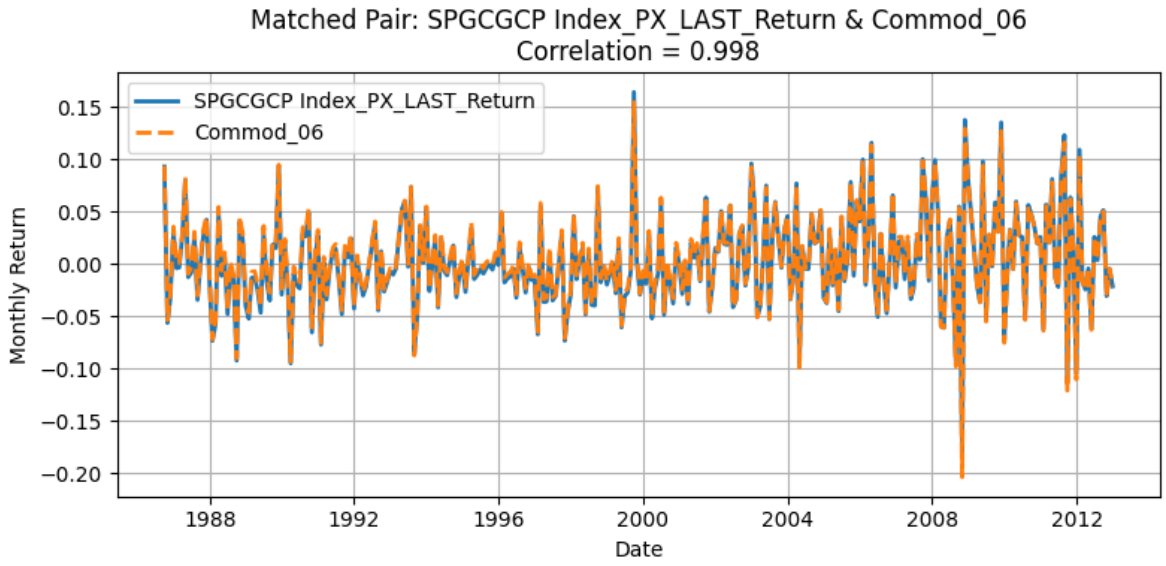


Figure 15: Pairwise Correlations between GSCI Commodity Returns and Yang (2013) Commodity Returns

A.1.7 Credit Default Swaps

Our Credit Default Swap (CDS) returns follow the methodology of Palhares (2012), implementing a constant-risky-duration construction that neutralizes maturity roll-over noise introduced by the 2009 "Big Bang" contract change. The cleaning procedure filters out zero-bid or non-standard contracts, reconciles with auction recovery data, and rescales by the risky annuity so that spreads are comparable across tenors. We interpolate missing maturities, align observations to common month-end fixing dates, and drop quotes that violate no-arbitrage bounds or sit outside the 1st-99th percentile of the cross-sectional spread distribution. This transformation pipeline provides CDS excess-return series that are free of roll discontinuities, stale quote reversals, and documentation clause inconsistencies.

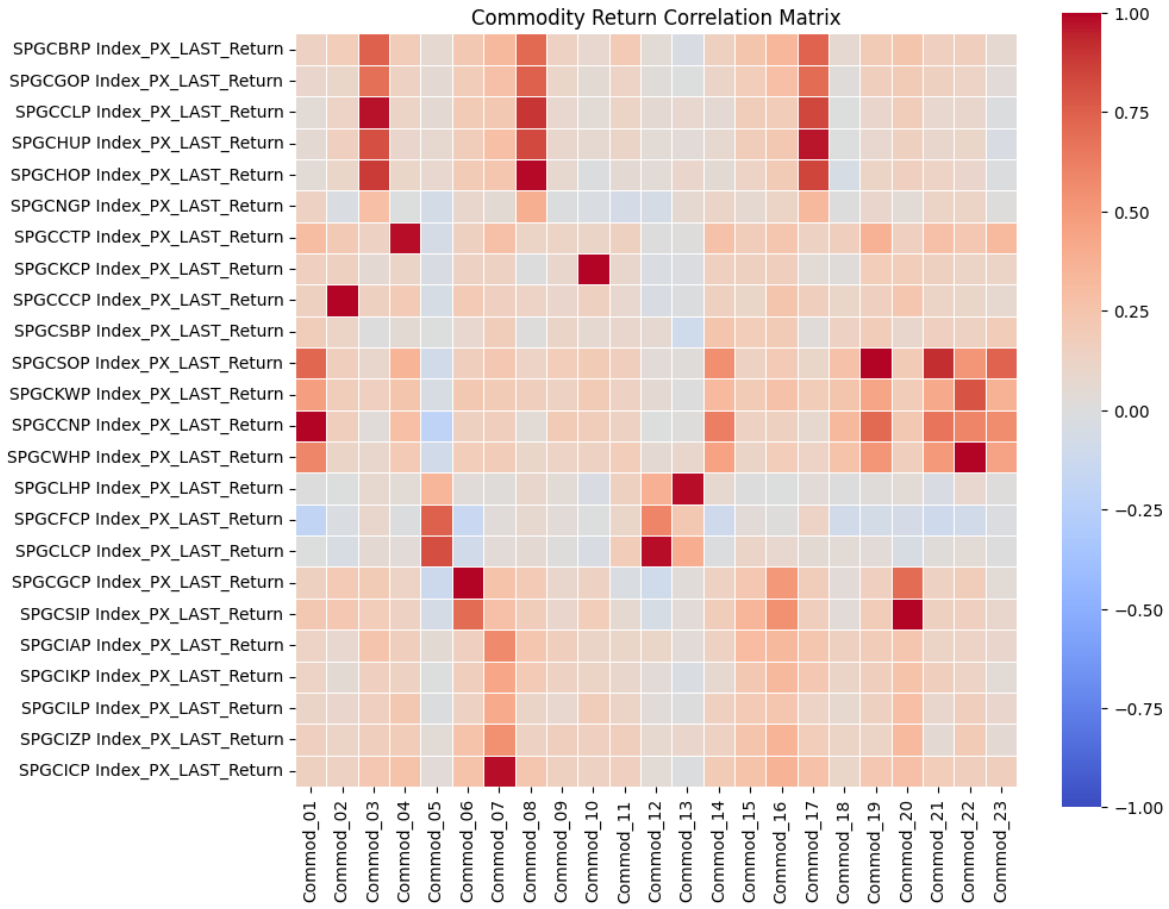


Figure 16: Heatmap of Pairwise Correlations between GSCI Commodity Returns and Yang (2013) Commodity Returns

A.2 Basis Spread Datasets

Following Siriwardane, Sunderam, and Wallen (2021), we construct various arbitrage spreads that measure market segmentation.

A.2.1 CDS-Bond Basis

Credit Default Swaps (CDS) are "insurance contracts" against the default of underlying corporate debt. The buyer of CDS protection pays a the CDS spread as a fixed annuity premium to the seller for a horizon τ. If there is a default before the time horizon τ, the buyer receives the difference between the bond's par value and its market value from the seller. As a result, the payoff of this portfolio should not deviate from the risk free bond. The resulting difference between CDS spread and floating rate spread (corporate bond rate - risk free rate) is defined by the authors as the CDS-basis.

The authors define the CDS basis (CB) as

$$CB_{i,t,\tau} \;=\; CDS_{i,t,\tau} \;-\; FR_{i,t,\tau}, \tag{2}$$

where:

- $FR_{i,t,\tau}$ = time t floating-rate spread implied by a fixed-rate corporate bond issued by firm i at tenor τ,
- $CDS_{i,t,\tau}$ = time t Credit Default Swap (CDS) par spread for firm i with tenor τ.

A negative basis implies an investor could earn a positive arbitrage profit by going long the bond and purchasing CDS protection. The investor would pay a lower par spread than the coupon of the bond itself and then receive value from the default.

The value of FR is substituted by the paper with Z-spread which we also modify in our construction. We address the substitution in detail later.

The value of CDS par spread is interpolated by the authors using a cubic spline function as not all necessary tenors are present.

Given the CDS spread from above, traditional construction of a risk-free rate for implied arbitrage implies the following return:

$$rfr^{CDS}_{i,t,\tau} \;=\; y_{t,\tau} \;-\; CB_{i,t,\tau}, \tag{3}$$

where:

- $y_{t,\tau}$ = maturity-matched Treasury yield at time t.

The implied risk-free arbitrage is then defined as the treasury yield in addition to the basis received when executing the CDS basis trade (investor benefits from negative basis).

Z-Spread (Zero-Volatility Spread)

Mathematical definition For a bond with cash-flows CF_t at times $t = 1, \ldots, N$ and Treasury spot rates s_t,

$$P = \sum_{t=1}^{N} \frac{CF_t}{\left(1 + s_t + Z\right)^t}.$$

The constant Z that solves this equation is the **Z-spread**.

Intuition Z is the uniform extra yield added to every point on the risk-free spot curve so that the discounted cash-flows equal the bond's dirty price P. It compensates investors for credit and liquidity risk relative to Treasuries.

Link to Yield-to-Maturity Setting the Z-spread pricing equation equal to the standard YTM equation gives

$$\sum_{t=1}^{N} \frac{CF_t}{(1+y)^t} = \sum_{t=1}^{N} \frac{CF_t}{\left(1 + s_t + Z\right)^t} \tag{4}$$

where y is the bond's yield-to-maturity. Except for the trivial flat-curve case ($s_t = s$), equation (4) has no algebraic solution-y or Z must be found numerically.

Continuous-Compounding Identity Rewrite discounts as e^{-rt}. With PV-weights

$$w_t = \frac{CF_t\, e^{-(s_t+Z)t}}{P}, \qquad \sum_t w_t = 1,$$

equation (4) yields the convenient mean-value relationship

$$y \;=\; \sum_{t=1}^{N} w_t \,(s_t + Z) \tag{A2}$$

Thus YTM is the PV-weighted average of the spot rates plus the Z-spread.

Practical Proxy: YTM Credit Spread Analysts often approximate Z with the **credit spread**

$$\Delta y = y_{\text{bond}} - y_{\text{Treasury-DM}},$$

where $y_{\text{Treasury-DM}}$ is the yield on a Treasury portfolio matched to the bond's (modified) duration.

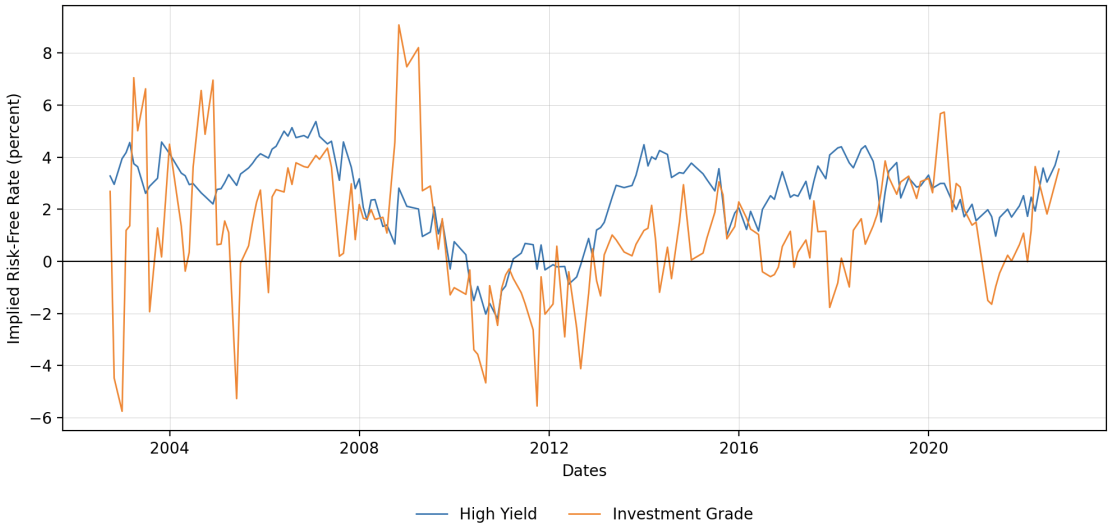


Figure 17: Comparison of CDS Arbitrage spreads

Why it works

1. A small parallel shift Z applied to all discount rates changes price by $-D_{\text{mod}}\ Z$. For modest spreads, this produces nearly the same price change as replacing the spot curve with a single rate shift Δy.

2. Duration-matching the Treasury benchmark neutralises curve-shape effects, so Δy isolates the average extra yield attributable to credit/liquidity risk.

3. Empirically, Δy tracks Z closely for plain-vanilla, option-free bonds, making it a "good-enough" proxy when full spot-curve data or iterative Z-spread calculations are impractical.

Note Z-spread is said to be populated by Markit in the CDS dataset but during the reconstruction process we found no proxy. Thus, we chose our own construction.

The CDS-Bond basis is plotted in Figure 17.

A.2.2 Covered Interest Parity (CIP)

During periods of market stress, such as the 2008 financial crisis and the 2020 COVID-19 pandemic, covered interest parity (CIP) may no longer hold as the forward—spot differential no longer exactly offsets interest-rate differentials. Factors contributing to this phenomenon include: Heightened Counterparty-Credit Risk, Liquidity Constraints, or Regulatory Pressures.

In other periods, deviations from CIP typically stem from market inefficiencies and are small in magnitude and are short-lived in timeframe due to arbitrage activities.

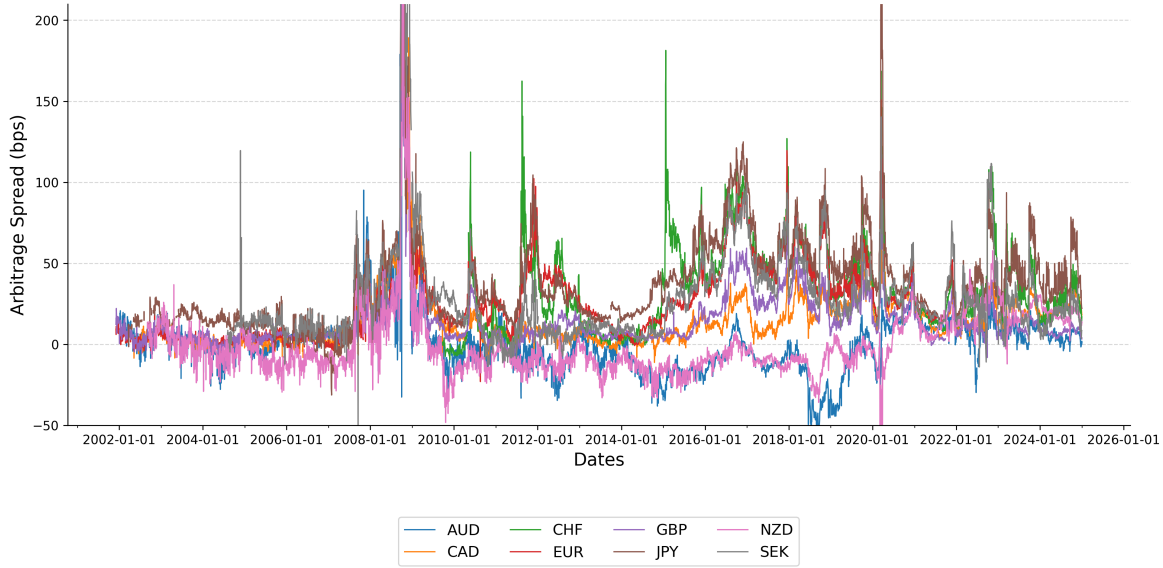


Figure 18: Comparison of CIP Arbitrage spreads

Our analysis examines CIP deviations across eight G10 currencies against the USD, using data from 1999 onwards sourced through Bloomberg Terminal.

The dataset includes:

- Spot exchange rates for each currency pair
- 3-month forward points for each currency pair
- 3-month Overnight Index Swap (OIS) rates for each currency

Data Standardization:

- Forward points are scaled appropriately (per 10,000 for most currencies, per 100 for JPY)
- Currencies conventionally quoted as USD-per-foreign-currency (EUR, GBP, AUD, NZD) are converted to reciprocal rates for consistency
- OIS rates serve as our risk-free benchmark to align with other arbitrage spread studies

We successfully replicate the CIP spreads as calculated by Siriwardane et al. (2023), as seen in our calculated currency respective arbitrage spreads versus those reported in their paper. If you compare the spreads in Figure 18, you will see that they are nearly identical to those reported in their paper.

During the overlapping periods of both datasets, the CIP spreads are nearly identical. Our findings confirm the presence of CIP deviations, particularly during periods of market stress, consistent with the conclusions drawn by Siriwardane et al. (2023).

A.2.3 TIPS-Treasury Basis

Following Fleckenstein, Longstaff, and Lustig (2014) as implemented in Siriwardane, Sunderam, and Wallen (2021), we construct the TIPS-Treasury arbitrage spread by creating synthetic nominal yields from TIPS real yields and inflation swaps, then comparing them to observed Treasury yields. The core intuition is that a TIPS bond plus an inflation swap replicates a nominal Treasury bond; persistent deviations from this equivalence reveal market segmentation or funding frictions. Our implementation combines Federal Reserve zero-coupon TIPS yields for 2-, 5-, 10-, and 20-year maturities with Federal Reserve zero-coupon nominal Treasury yields (using the Gürkaynak-Sack-Wright model) and Bloomberg zero-coupon inflation swap rates for matching maturities. For each tenor, we construct a TIPS-implied risk-free rate by combining the TIPS real yield (in continuously compounded decimal form) with the inflation swap rate via the formula $\text{tips\_treas} = 10{,}000 \times (\exp(r + \log(1 + \pi)) - 1)$, where r is the real yield and π is the inflation swap rate. The arbitrage spread is then the difference between this synthetic rate and the observed Treasury yield in basis points. We convert TIPS yields from percentage to decimal form and Treasury yields to basis points using the exponential transformation $10{,}000 \times (\exp(y) - 1)$ to properly handle the continuously compounded GSW model output. Inflation swap data availability varies by tenor and begins later than the TIPS and Treasury series. We implement quality filters to exclude observations with missing values for three or more of the four tenors, ensuring sufficient cross-sectional coverage. The resulting spreads closely match the patterns documented in the literature: substantial positive values during 2008-2009 (often exceeding 100 basis points), narrowing during the post-crisis period, and renewed spikes during the March 2020 liquidity crisis. The 10-year and 20-year spreads exhibit high correlation (typically above 0.95), while the 2-year spread shows more independent variation due to differing inflation dynamics and liquidity effects at the short end. These patterns confirm that Treasury bonds are consistently expensive relative to TIPS during stress periods, consistent with the flight-to-quality premium documented by Fleckenstein, Longstaff, and Lustig (2014).

A.2.4 Treasury Spot-Futures Basis

Following Fleckenstein and Longstaff (2020) and Siriwardane, Sunderam, and Wallen (2021), we construct the Treasury spot-futures basis by extracting implied repo rates from Bloomberg data for 2-, 5-, 10-, 20-, and 30-year Treasury futures contracts and comparing them to maturity-matched OIS rates. For each tenor and date, we pull both the near contract and the first-deferred contract, using Bloomberg's cheapest-to-deliver implied repo rate calculation along with trading volume and contract month specifications. To avoid the delivery option distortions documented by Burghardt et al. (2005), we compute the basis using only the first-deferred contract, which by construction is not in its delivery

month.

The core technical challenge lies in constructing a clean maturity-matched benchmark. We parse each contract's expiration month and year to compute days-to-maturity relative to the last business day of the delivery month, then linearly interpolate OIS rates across available tenors (1-week through 1-year) to match the futures contract horizon. The basis is simply the difference between the futures-implied repo rate and this interpolated OIS rate, expressed in basis points. We restrict to post-June-2004 data, require positive trading volume in the deferred contract, and remove outliers using a rolling 45-day median absolute deviation filter with a threshold of 10 times the MAD. Missing values are forward-filled for up to 5 days. The resulting series successfully replicates the persistent positive spreads documented in Siriwardane, Sunderam, and Wallen (2021), particularly during stress episodes in 2008-2009 and March 2020, when funding market frictions cause the futures-implied rate to exceed OIS by substantial margins.

A.2.5 Treasury-Swap Basis

Following Siriwardane, Sunderam, and Wallen (2021), we construct the Treasury-Swap arbitrage spread by comparing fixed-rate USD overnight indexed swap (OIS) yields to zero-coupon Treasury yields of matching maturities. The arbitrage opportunity arises when the swap spread is negative—when Treasury yields exceed swap rates. In this scenario, an investor purchases a Treasury financed via repo, pays fixed in a matched-tenor swap, and receives floating. The position earns positive carry because the Treasury coupon exceeds the swap fixed rate paid, while the floating rate received (LIBOR or SOFR) exceeds the repo financing cost. This constitutes a textbook arbitrage under frictionless conditions. The reverse trade does not work when spreads are positive because shorting Treasuries through reverse repo is expensive—LIBOR typically exceeds the reverse repo rate earned, making the floating leg unprofitable and the overall position uneconomical. This asymmetry explains why negative spreads violate no-arbitrage conditions while positive spreads simply reflect normal credit and liquidity premia. Despite the theoretical arbitrage, negative spreads have persisted since 2008 due to balance sheet costs, supplementary leverage ratio requirements, and margin constraints that limit dealer arbitrage capacity. Our implementation pulls Bloomberg constant-maturity Treasury yields for 1-, 2-, 3-, 5-, 10-, 20-, and 30-year tenors along with corresponding fixed-rate OIS quotes for identical maturities. For each tenor, the arbitrage spread is computed as 100 times the difference between the swap rate and the Treasury yield, expressed in basis points. We restrict the sample to observations beginning in 2000 and drop dates with missing values across all tenors. The replication successfully matches the persistent negative swap spreads documented by Jermann (2020), Du, Hébert, and Li (2023), and Hanson, Malkhozov, and Venter (2023), with the 10-year and 30-year tenors showing

particularly large deviations during the post-crisis period.