WORKING PAPER 26-05

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

Jeremiah Bejarano, Office of Financial Research
Viren Desai, Independent
Kausthub Keshava, Independent
Arsh Kumar, Independent
Zixiao Wang, Independent
Vincent Hanyang Xu, Independent
Yangge Xu, Independent

OFR
OFFICE OF FINANCIAL RESEARCH

August 25, 2026

Why These Findings Are Important

Forecasting financial markets, including stock returns, bond prices, and signs of funding stress, guides decisions from monetary policy to bank supervision. Yet, studies disagree on the method that works best, mostly because of the data. When researchers studying the same market clean their data differently, their conclusions cannot be compared. The authors build an open benchmark that fixes data across methods, spanning equities, bonds, Treasuries, currencies, commodities, and other markets. Holding the data constant, they find that modern machine learning and deep learning methods deliver real forecasting gains for indicators important to financial stability monitoring.

Key Findings

1. **Modern machine learning and deep learning methods deliver clear forecasting gains measures of funding stress, capturing patterns that classical methods miss.**
2. **These methods stay competitive with the best classical models on bank and intermediary indicators central to financial stability monitoring.**
3. **No single method wins everywhere, so using a domain-specific benchmark is essential to choosing an appropriate forecasting method.**

How the Authors Reached These Findings

The authors assemble a standardized, open-source collection of financial datasets spanning seven asset classes (equities, corporate bonds, Treasuries, foreign exchange, commodities, credit default swaps, and options), five funding stress basis spread measures, and bank and intermediary indicators. Each dataset is cleaned according to best practices for that market, and every cleaning step is validated against published results. The authors then test about a dozen forecasting methods, from classical statistical models to modern deep learning, on identical data.

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[‡]

August 25, 2026

Abstract

Accurate time series forecasts underpin asset pricing, risk management, monetary and macroprudential policy, and other applications. The set of available forecasting methods is expanding rapidly, driven by new machine learning models. This raises a practical question: Do these methods deliver real forecasting power gains on financial data? Since no single method is best across all data generating processes, the question can be answered only by direct evaluation on domain-specific data, and a fair comparison requires holding the data fixed across methods. Yet, new methods are rarely evaluated on the financial data commonly used in academic finance; when they are, each is assessed on its own dataset and cleaning conventions. Apparent method rankings entangle method skill with cleaning choices that themselves require domain expertise. We address this by assembling, in one place, canonical cleaning procedures for many financial datasets to hold the data fixed across forecasting experiments. We introduce a standardized open-source dataset covering equities, corporate bonds, U.S. Treasuries, foreign exchange, commodities, credit default swaps, options, five basis spread datasets, and bank and intermediary indicators, each cleaned per the canonical paper for that asset class. Holding the data fixed and evaluating roughly a dozen univariate methods without exogenous regressors, we find that asset returns remain near unforecastable across every method family and that hybrid and machine learning methods exhibit additional forecasting power on basis spreads and bank indicators.

1 Introduction

Accurate time series forecasts drive consequential decisions across economics and finance. In asset pricing, for example, the predictability of aggregate returns reveals the extent to which price variation reflects time varying risk premia rather than shifting expectations of future cash flows, making return forecasting the central empirical task of contemporary asset pricing (Cochrane, 2011; Kelly and Xiu, 2023). In macroeconomic policy, central banks rely on factor models and large panel forecasts of output, inflation, and financial conditions to inform policy in real time (Stock and Watson, 2002a;

[*]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, Jared Szajkowski, 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

McCracken and Ng, 2016). Supervisory stress tests project bank losses, earnings, and capital ratios over multiple quarterly horizons to calibrate buffers and intervene before vulnerabilities crystallize; although these projections come from scenario-based supervisory models rather than the time series forecasting methods studied here, were such methods to inform these exercises, method choice would carry similar stakes. Also, macroprudential regulators monitor early warning indicators of systemic stress (Adrian, Covitz, and Liang, 2015; Oet et al., 2011). Across these domains, the empirical question is the same: Given a panel of financial time series, which forecasting method should one use?

Despite the stakes and a rapidly expanding toolkit of forecasting methods driven by machine learning, no method has emerged as a universal winner, and the leaderboard shifts from one study to the next. A large scale re-evaluation on the M3 series found that classical statistical models dominate machine learning alternatives both in accuracy and computational cost (Makridakis, Spiliotis, and Assimakopoulos, 2018a). The M4 competition's winner was a hybrid recurrent neural network and exponential smoothing system, with the strongest pure machine learning methods placed in the middle of the field (Makridakis, Spiliotis, and Assimakopoulos, 2020). Just two years later, the M5 competition was swept by pure machine learning methods, and each beat the best statistical baselines and combinations thereof (Makridakis, Spiliotis, and Assimakopoulos, 2022). The Monash Time Series Forecasting Repository likewise reports no uniform winner across its 25 datasets (Godahewa et al., 2021). These shifts are consistent with (i) genuine series-by-series heterogeneity in dynamics and (ii) the No Free Lunch theorem (Wolpert and Macready, 1997): No single algorithm is best for all data generating processes. The implication is practical in that to learn whether a new method genuinely improves financial forecasting, one must evaluate it based on financial data, with the data held fixed across the compared methods.

Holding the data fixed is precisely what extant comparisons fail to do, and variations in data can have material effects on measured forecasting power. We prepare standard datasets for each of many variables that have been the focus of forecast evaluations and apply a range of forecasting methods to each such dataset. Though exceptions exist, forecasting power differs more across the variable being forecast than across methods applied to a variable. For example, all the methods that we implement find poor predictability for asset returns, but predictability is more meaningful for basis spreads or bank financials.

As a concrete example, Dickerson, Robotti, and Rossetti (2024) shows that signals constructed from bid–ask averaged TRACE prices (credit spreads, yields, short-term reversals) are mechanically correlated with next month's returns because the same market microstructure noise (MMN) enters both the signal and the realized return. Their MMN correction enforces an implementation gap between the prices used to form the signal and those used to compute the return. With this correction,

the long–short premium on credit spread sorted portfolios falls by 50% to 65%, and the short-term reversal premium falls by over 90% (from roughly 0.90% per month to almost zero). What looked like predictability was instead an artifact of the cleaning convention. Any verdict on which method best forecasts corporate bond returns is therefore conditional on whether that cleaning recipe was applied. The broader finance literature points the same way: Jensen, Kelly, and Pedersen (2023) finds that 32% of published cross-sectional return predictors become statistically insignificant out of sample, and the Open Source Cross-Sectional Asset Pricing project of Chen et al. (2022) shows that transparent, version-controlled data construction reduces the researcher degrees of freedom that drive irreproducibility (we revisit these in Section 2). Hewamalage, Ackermann, and Bergmeir (2023) documents a parallel problem on the methodology side: Machine learning benchmarking papers routinely use inadequate baselines, ill-suited error metrics, and inappropriate data partitioning, producing rankings that do not survive careful re-evaluation (Prater, Hanne, and Dornberger, 2024). Without a fixed-data benchmark, method skill and cleaning choices are entangled such that no cross-study comparison can disentangle.

As a service, we provide the Financial Time Series Forecasting Repository (FTSFR), a standardized open-source benchmark for financial markets. The FTSFR is the first comprehensive time series forecasting repository focused specifically on financial markets. Coverage spans seven asset classes (equities, corporate bonds, U.S. Treasuries, foreign exchange, commodities, credit default swaps, and options), five basis spread datasets following Siriwardane, Sunderam, and Wallen (2021), and bank and intermediary indicators including bank Call Reports (Drechsler, Savov, and Schnabl, 2017) and the intermediary capital factors of He, Kelly, and Manela (2017). For each dataset, the cleaning procedure produces data that support replication of the canonical paper for that asset class, and we validate every replication against published figures or summary statistics.[1] Data curation is domain-specific, recognizing that financial data require specialized treatment reflecting market microstructure, regulatory requirements, and established academic conventions. This specialized treatment rests on judgments that only domain knowledge supplies, like which instruments are genuinely tradable, which quotes or contracts are faulty and must be dropped, whether a series carries enough liquidity to be meaningful, and which transformations a given market convention requires. Each dataset is processed using precise subsample definitions, filters, and transformations specified in the corresponding literature, ensuring that forecasting evaluations reflect the actual data used by experts rather than generic time series.

Because financial data are typically subject to licensing restrictions that prevent direct redistribution, we made the full pipeline (download, clean, format) open source rather than just the raw data.

[1] While the original papers often lack publicly available replication code, our implementations reproduce their key results, which both validates our data processing and provides a reusable framework for future research. The full pipeline is open source on GitHub at `https://github.com/ftsfr`. In this online code repository, an automated workflow downloads each series from its underlying source and cleans it according to the canonical methods in the literature, and pinned virtual environments yield consistent results across computing environments.

Therefore, researchers with WRDS and Bloomberg subscriptions can reproduce every series exactly. FTSFR is the financial domain analogue of FRED-MD (McCracken and Ng, 2016) for macroeconomics and the Monash Time Series Forecasting Repository (Godahewa et al., 2021) for general time series, complementing more limited efforts such as FinTSB (Hu et al., 2025), TFB (Qiu et al., 2024), and several non-financial archives (Dau et al., 2019; Bagnall et al., 2018; Tan et al., 2020); Table 1 compares coverage. By holding the data constant across methods, FTSFR turns the question about which forecaster wins into a question that can be answered reproducibly.

We present comprehensive baseline forecasting results across all datasets using a suite of 12 methods ranging from classical statistical models (ARIMA, Simple Exponential Smoothing, Theta) and the Historical Average benchmark to hybrid linear architectures (DLinear, NLinear) and deep learning models (DeepAR, N-BEATS, N-HiTS, Vanilla Transformer, TiDE, KAN). We provide novel insights into market specific predictability by systematically evaluating forecast 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. Our results suggest that standardized comparisons depend not only on implementation of methods but on data for a variable that are the same across methods.

Our own baseline forecasting estimates document three stylized facts. First, for asset returns (Center for Research in Security Prices (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. We caveat this finding because the benchmark restricts attention to univariate global forecasts without exogenous regressors (Section 2.4), so this near-zero predictability holds within that model class. Admitting exogenous covariates or other model classes is a natural extension that we leave for future work. Second, on basis spreads, hybrid and deep learning global models deliver the most forecasting power, exploiting persistent seasonal funding-quarter structure that classical univariate methods do not encode. Third, on bank and intermediary indicators, classical statistical methods (ARIMA, Theta) and deep learning competitors are roughly competitive on median R^2_{oos}, but classical methods edge out deep learning on the mean by avoiding tail failures on disaggregated panels. Knowing which model works best therefore depends on the task because the appropriate default differs across return forecasting, basis spread monitoring, and bank balance sheet projection.

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.

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. (2025)	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

Sources: Authors' analysis

2 Background

2.1 Return Predictability and Asset Pricing

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 the perspective that variation in price-to-dividend ratios corresponds almost entirely to changes in discount rates (expected returns) and not to expected cash flows (Cochrane, 2011). High valuations reliably precede periods of low subsequent returns across asset classes, including equities, bonds, currencies, credit, and real estate. This common pattern underscores variation in discount rates as the organizing principle of contemporary asset pricing research.

Kelly and Pruitt (2013) shows 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.

2.2 Forecasting for Financial Stability and Many-Predictor Methods

Beyond asset pricing, 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 involve multiple quarterly projections of earnings, losses, and capital ratios.[2] In practice, these projections are produced by supervisory models rather than by the time series forecasting methods evaluated here. However, if such methods were to be used to inform stress testing exercises, a standardized benchmark of their forecasting performance on financial data would be an important input. Complementing firm-level exercises, broad cyclic risk gauges, such as the ECB's 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.[3]

The 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 Financial Stress Index (FSI) applies a closely related approach in that it is a factor model that essentially uses the first principal component of a broad, daily panel of market indicators (see Monin (2019) and Bejarano (2023)). Another example was the Systemic Assessment of Financial Environment (SAFE) early warning system, developed by researchers from the Federal Reserve Bank of Cleveland, which integrated 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 datasets relevant to financial stability.

[2] Federal Reserve, *Dodd-Frank Act Stress Test 2020: Supervisory Stress Test Framework and Model Methodology*. See https://www.federalreserve.gov/publications/june-2020-supervisory-stress-test-framework-and-model-methodology.htm.

[3] 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.

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 on 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 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, 2018b, 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) observes, this pattern extends beyond the M-competitions as 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, which is 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 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 Versus 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 might also exist when a researcher runs ARIMA on each series (local) but learns 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 classical literature, ARIMAX and VARX extensions incorporate external regressors. Modern deep learning methods, such as DeepAR or Temporal Fusion Transformers, can accept static attributes (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, bank and intermediary indicators) is modeled and evaluated in isolation, and the taxonomy (e.g., the one provided 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

Other machine learning benchmarks distribute their datasets openly, but financial data are typically subject to licensing restrictions and copyright limitations that prevent direct redistribution. We work around this by making the full download, cleaning, and formatting pipeline available as open source rather than the data itself so that any researcher with the appropriate vendor subscriptions can reproduce each series exactly.

Finance faces an active debate about a potential replication crisis. Jensen, Kelly, and Pedersen (2023) finds that the average predictor's out-of-sample strength is only 54% of its original in-sample strength and that 32% become statistically insignificant out of sample. Much of the debate around these findings has centered on data snooping and the multiple testing problem. We want to draw attention to a parallel and arguably more tractable source of irreproducibility and focus on the construction of the underlying datasets themselves. Two researchers nominally drawing from the same source (TRACE corporate bonds, CRSP equities, OptionMetrics surfaces) routinely produce series that differ in their handling of bid–ask bounce, their survival and liquidity filters, their portfolio formation rules, and the sample windows they retain. Those choices propagate into every downstream statistical exercise. Forecasting evaluations are no exception. When corporate bond returns refer to one constructed series in one paper and a meaningfully different series in another, a comparison of forecasting methods across the two studies is not a comparison of methods.

The corporate bond literature illustrates the point starkly. Even a widely cited set of priced factors in the cross-section of corporate bond returns was later found to be sensitive to errors in the underlying data construction, errors that were invisible from the published text and surfaced only when the data pipeline was reconstructed. Dickerson, Robotti, and Rossetti (2024) documents the broader pattern that many published bond anomalies depend on bid–ask bounce contamination in the TRACE return series. Also, applying a market microstructure noise correction roughly halves the apparent credit spread return and substantially weakens the cross-section of priced anomalies.[4] Headline results, and forecasting evaluations built on the same return series are sensitive to data-construction choices that are largely invisible from the published text.

Two open-source efforts demonstrate that this irreproducibility is at least partly fixable by infrastructure rather than by tighter statistical discipline alone. Chen et al. (2022), in their Open Source Asset Pricing (OSAP) project, show that transparent, version-controlled construction of equity anomaly portfolios eliminates a large share of the researcher degrees of freedom that drive cross-study disagreement. The Open Source Bond Asset Pricing (OSBAP) project that grew out of Dickerson, Mueller, and Robotti (2023) plays an analogous role for fixed income: it releases public code for con-

[4] See Lee (2023) at Bloomberg for the popular account of the bond-data episode.

structing and pricing corporate bond portfolios, so that any subsequent claim about bond predictability can be traced back to a specific, auditable data pipeline. Our benchmark adopts the same philosophy and extends it across asset classes. We open-source every step of the download, clean, and format pipeline; only the raw data remain subject to vendor licensing. Any researcher can reproduce each series exactly. Running new methods on the same data then isolates the forecasting method from the cleaning method.

3 Benchmark Datasets

In this section, we describe the datasets included in the benchmark. The datasets are organized into asset class datasets, basis spread datasets, and other financial data. We provide brief description of each dataset below. Each dataset is constructed on a cleaning procedure from a well-known paper in 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.[5]

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 that typically invest only in equities. This distinction is crucial because, in the HKM theory, 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. Under this view, these assets are natural indicators for monitoring financial stability, as distress in intermediary balance sheets should manifest 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 this literature views these spreads as early warning indicators of stress in the financial system. In that account, when intermediaries face funding constraints or balance sheet pressures, arbitrage opportunities persist and basis spreads widen, signaling potential systemic stress (Du, Tepper, and Verdelhan, 2018; Siriwardane, Sunderam, and Wallen, 2021). Our methodological approach follows the principle established by HKM of using canonical cleaning pro-

[5] See https://github.com/ftsfr/ftsfr.

cedures 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 to 50 series. The date columns indicate each dataset's coverage span, with most datasets providing substantial histories for robust model comparison.

3.2 Asset Class Datasets

Our repository provides comprehensive datasets across seven major asset classes, each cleaned according to the canonical methods established in academic literature. We replicate each of the asset classes examined in HKM, which are commodities, corporate bonds, credit default swaps (CDS), equities, foreign exchange markets, options, and U.S. 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 HKM) 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 the securities into portfolios. To validate our data cleaning procedures, we replicate key summary statistics and cross-sectional patterns from each source paper cited within HKM. 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 allows researchers to both replicate existing studies (which typically use the aggregated portfolios) as well as use the disaggregated data for richer analysis with the global time series forecasting methods.

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.	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)
CRSP Stock (ex-div)	Monthly stock returns from CRSP, excluding dividends	Fama and French (1993)
FF25 Size-BM	Daily Fama-French 25 portfolios: size and book-to-market	Fama and French (1993)
FX	Daily foreign exchange returns versus USD	Lettau, Maggiori, and Weber (2014)
SPX Options Portfolios	Monthly leverage-adjusted returns for SPX option portfolios (54 CJS + 18 HKM)	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	Gürkaynak, Sack, and Wright (2007)
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	Drechsler, Savov, and Schnabl (2017)
BHC Cash Liquidity	Quarterly bank holding company cash liquidity	Drechsler, Savov, and Schnabl (2017)
BHC Leverage	Quarterly bank holding company leverage ratios	Drechsler, Savov, and Schnabl (2017)
HKM Daily Factor	Intermediary risk factors: the capital risk factor (AR(1) innovation in the capital ratio) and the value-weighted intermediary investment return	He, Kelly, and Manela (2017)
HKM Monthly Factor	Same as above, but monthly	He, Kelly, and Manela (2017)
HKM All Factor	Full He, Kelly, and Manela factor file combining the released intermediary factor series	He, Kelly, and Manela (2017)
Treasury Yield Curve	Daily Nelson-Siegel-Svensson zero-coupon yields, 1-30 years	Gürkaynak, Sack, and Wright (2007)

Sources: Authors' analysis

Table 3: Data Sources by Dataset

Dataset Name	Data Sources
Returns Data	
CDS Contract	S&P Global CDS
CDS Portfolio	S&P Global CDS
Commodity	Bloomberg Terminal
Corporate Bond	Open Source Bond Asset Pricing[a]
Corporate Portfolio	Open Source Bond Asset Pricing
CRSP Stock	Center for Research in Security Prices (CRSP)
CRSP Stock (ex-div)	Center for Research in Security Prices
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	Center for Research in Security Prices
Basis Spread Data	
CDS-Bond	Center for Research in Security Prices; FINRA TRACE; S&P Global CDS and RED Entity
CIP	Bloomberg Terminal
TIPS-Treasury	Bloomberg Terminal
Treasury-SF	Bloomberg Terminal
Treasury-Swap	Bloomberg Terminal
Other Financial Data	
Bank Cash Liquidity	Call Reports, following Drechsler, Savov, and Schnabl (2017)[c]
Bank Leverage	Call Reports, following Drechsler, Savov, and Schnabl (2017)
BHC Cash Liquidity	Call Reports, following Drechsler, Savov, and Schnabl (2017)
BHC Leverage	Call Reports, following Drechsler, Savov, and Schnabl (2017)
HKM Daily Factor	He, Kelly, and Manela (2017)[d]
HKM Monthly Factor	He, Kelly, and Manela (2017)
HKM All Factor	He, Kelly, and Manela (2017)
Treasury Yield Curve	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://asaf.manela.org/data/. [e] See https://www.federalreserve.gov/data/nominal-yield-curve.htm.

Sources: Authors' analysis

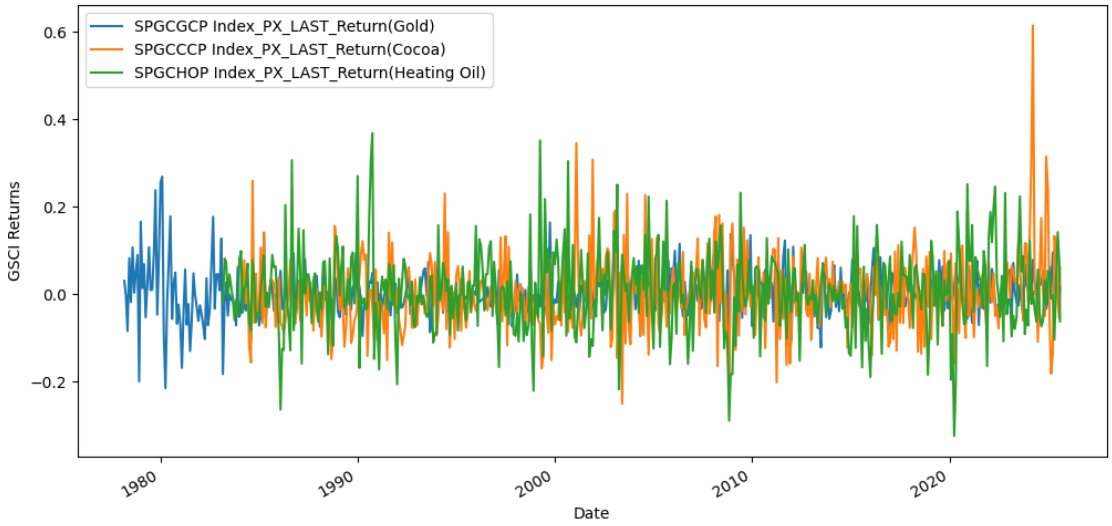
Table 4: Dataset Statistics Summary

	Frequency	Unique Entities	Min Length	Median Length	Max Length	Min Date	Max Date
Basis Spreads							
CDS-Bond	Monthly	2776	1	66	109	2002-07-31	2025-04-30
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	269	269	269	2002-08-31	2024-12-31
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	53389	1	22	272	2002-08-31	2025-03-31
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	2	516	516	516	1970-01-01	2012-12-01
HKM Daily Factor	Daily	2	4765	4765	4766	2000-01-03	2018-12-11
HKM Monthly Factor	Monthly	2	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, Compustat, Call Reports, FINRA TRACE, OptionMetrics, S&P Global, Open Source Bond Asset Pricing, He, Kelly, and Manela (2017), Authors' analysis

Commodities Commodity futures follow the protocol of Yang (2013), but we are unable to access the same Commodity Research Bureau data that the author 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 23 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 To obtain corporate bond returns, we use the corporate bond panel from the OSBAP project, 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 as the bid–ask averaged quotes in raw TRACE embed a mechanical reversal that can be misinterpreted as illiquidity and lead researchers to overstate corporate bond excess returns. With the cleaned data, we replicate and extend the ten value-weighted 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 artifact 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 HKM.

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 neutralizes maturity rollover noise introduced by the 2009 Big Bang contract change. Raw S&P Global 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, aligns observations to common month-end fixing dates, and drops quotes that violate no arbitrage bounds or sit outside the 1st to 99th percentile of the cross-sectional spread distribution. Making this transformation pipeline open source 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 datasets 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, and NASDAQ).

Foreign Exchange We provide daily returns from the individual foreign currencies against the U.S. dollar. The specified structure is based on implied returns of the U.S. dollar if converted to foreign currency then invested 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 the foundation of the approach used by He, Kelly, and Manela (2017) to construct the options portfolios used in the 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 that span 24 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 make 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 used to prepare the raw SPX options data for analysis and portfolio construction. These include technical and mathematical details on volatility estimation, kernel methods, and other relevant techniques. These details are given in the appendix and in the online code module.

Treasuries Our U.S. Treasury bond portfolios utilize the CRSP Treasury database and the same filters 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.[6] 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 only use 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, which are 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. The literature commonly attributes these failures to funding frictions or balance sheet constraints (Du, Tepper, and Verdelhan, 2018; Siriwardane, Sunderam, and Wallen, 2021). Since the Global Financial Crisis (GFC), researchers have documented several such anomalies across asset classes, and these have important implications for financial stability. Covered interest parity (CIP), once "the closest thing to a physical law in international finance," is a leading example because of numerous systemic violations since 2008 (Du, Tepper, and Verdelhan, 2018). The size of the violation is measured by the cross-currency basis, the gap between the cost of borrowing dollars directly and the cost of synthetic borrowing by lending a foreign currency and hedging the exchange rate forward. As Du, Tepper, and Verdelhan (2018) explain, when dollar funding is scarce, foreign institutions bid up the synthetic route and the basis widens, so the basis effectively prices dollar funding

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

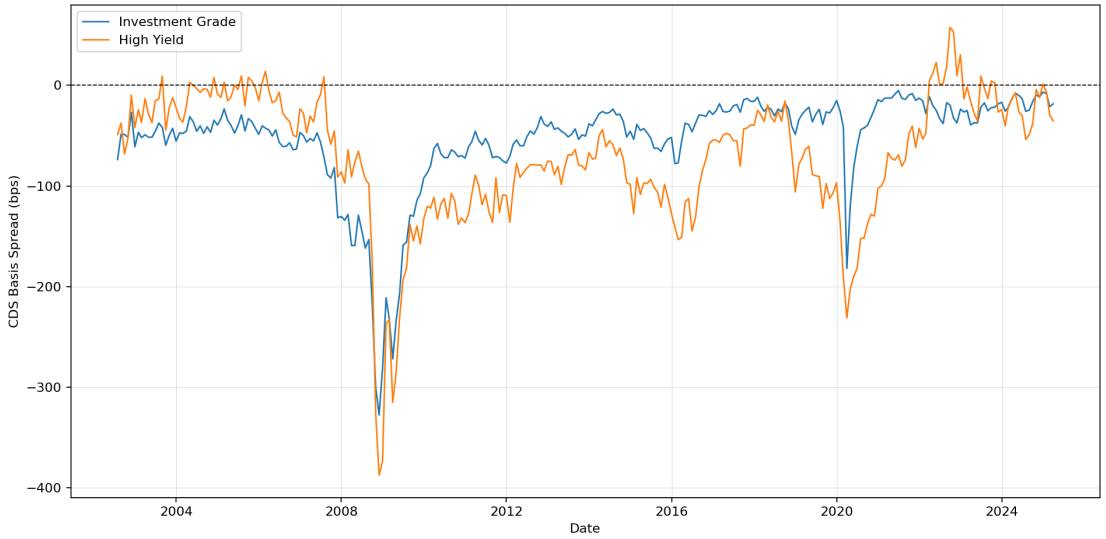
stress. There are sharp spikes during crises as demonstrated in 2008 and again in March 2020 when the basis against major currencies widened dramatically as foreign institutions scrambled for dollars (Board of Governors of the Federal Reserve System, 2020). Because the basis is such a direct gauge of funding stress, central banks monitor it. For example, the Federal Reserve describes its central-bank swap lines, which lend dollars to foreign central banks during market stress, as intended to ease strains in global dollar funding markets (Board of Governors of the Federal Reserve System, 2020).

In this paper and its associated repository, we provide code that will allow researchers to replicate several basis spreads across 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 constructed with methodologies that are well-established in literature. To validate our replication, we follow the papers recommended in Siriwardane, Sunderam, and Wallen (2021) and replicate key summary statistics from each source paper.

CDS-Bond Basis We follow Siriwardane, Sunderam, and Wallen (2021) in defining the CDS-bond basis as the difference between a maturity matched CDS par spread and the bond's floating-rate spread, and link single-name CDS quotes from S&P Global to corporate bonds on ISIN through the RED-ISIN obligation lookup table. Our bond panel uses monthly bond prices calculated from FINRA TRACE that are restricted to USD-denominated senior unsecured fixed-coupon issues with valid ISIN, one- to ten-year remaining maturity, outstanding principal of at least $100,000, no convertible structures, and an end-of-month price of no more than a 50% discount on the par value. The floating-rate spread is constructed as a bond-level Z-spread by fitting a Nelson–Siegel–Svensson Treasury curve on the CRSP daily Treasury panel (Gürkaynak, Sack, and Wright, 2007) and solving for the constant continuous compounded spread that reprices each bond's coupon and principal cash flows. The CDS par spread is interpolated by cubic spline across the 1Y, 3Y, 5Y, 7Y, and 10Y tenors. Note that our resulting basis series is monthly, whereas the daily index of Siriwardane, Sunderam, and Wallen (2021) is constructed from daily S&P Global bond prices using a set of tradability filters on ISINs. Section A.2.1 gives the full filter list, including our outlier handling. These spreads are shown in Figure 2.

Covered Interest Parity We replicate eight of the G10 currency series studied by 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), synchronize all legs to the 17:00 ET close while dropping holiday or stale quotes, and winsorize the extreme 1% of annualized 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

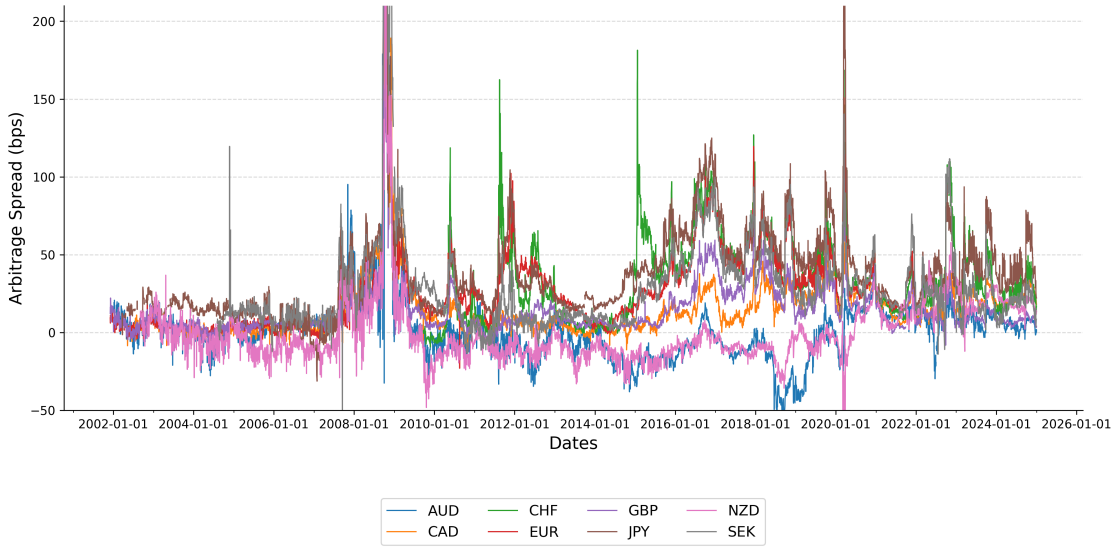


Sources: Center for Research in Security Prices, FINRA TRACE, S&P Global, Authors' analysis

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 harmonize 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 cheapest-to-deliver implied repo rates for 2-, 5-, 10-, 20-, and 30-year Treasury futures contracts from Bloomberg. For each tenor and date, we only use 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 and then 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 observations dated post-June 2004, require positive trading volume in the deferred contract, and remove outliers using a rolling 45-day median absolute deviation filter with a threshold that is 10 times the mean absolute deviation. 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) and 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.

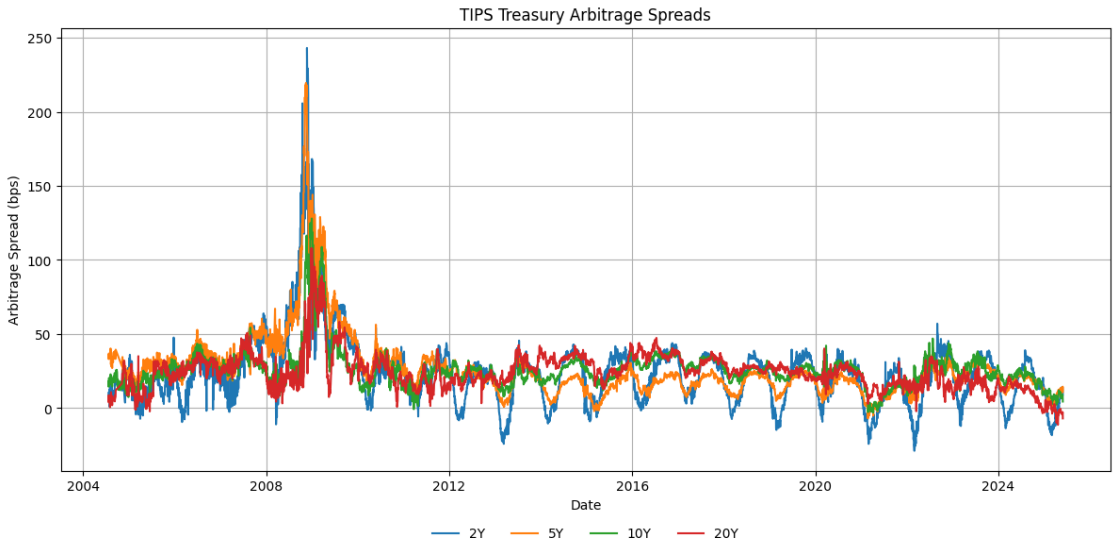
3.4 Other Financial Data

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)[7]. 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

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

Figure 4: TIPS-Treasury Basis Spreads



Sources: Bloomberg, Authors' analysis

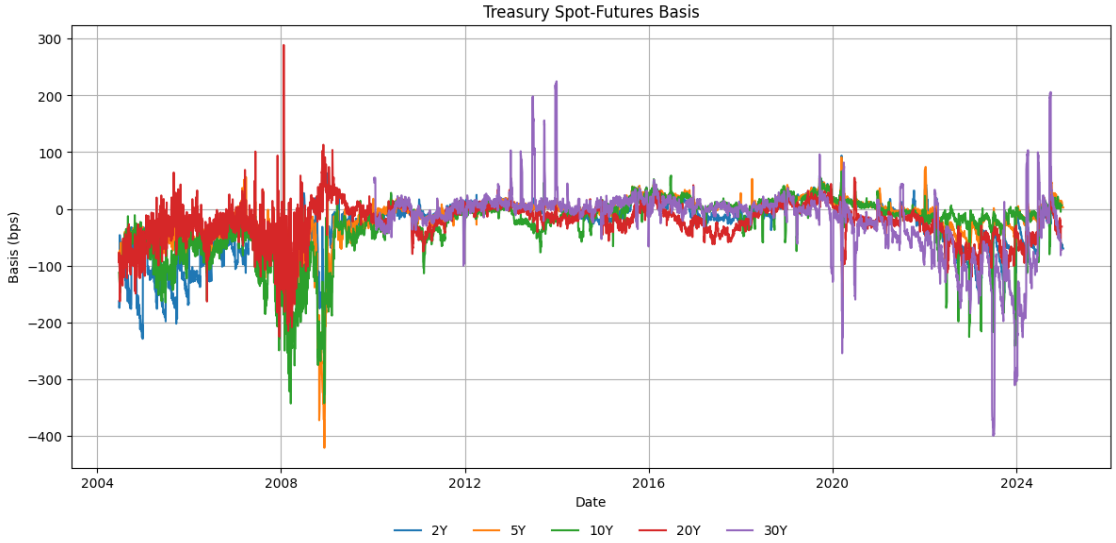
by aggregate market equity plus book debt, matching the dealer list to public holding companies in CRSP/Compustat. The intermediary capital risk factor is the AR(1) innovation in the capital ratio scaled by the lagged ratio. We obtain the published factor series directly from Asaf Manela's data website and align the monthly (and post-2000 daily) series to our panel; we 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 (excluding bills/FRNs and on-the-run issues). The six-parameter Svensson model is used post-1980 and Nelson-Siegel is used pre-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 bank indicators may serve as valuable covariates for certain panels while acting as leading indicators for others. In section 2.4, we explained that while we acknowledge the opportunity to create richer forecasts in this way, forecasting methodologies that do so are outside of the scope of this paper. Consequently, the comparative results in Table 10 remain panel specific by construction, even though the underlying data resources support

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



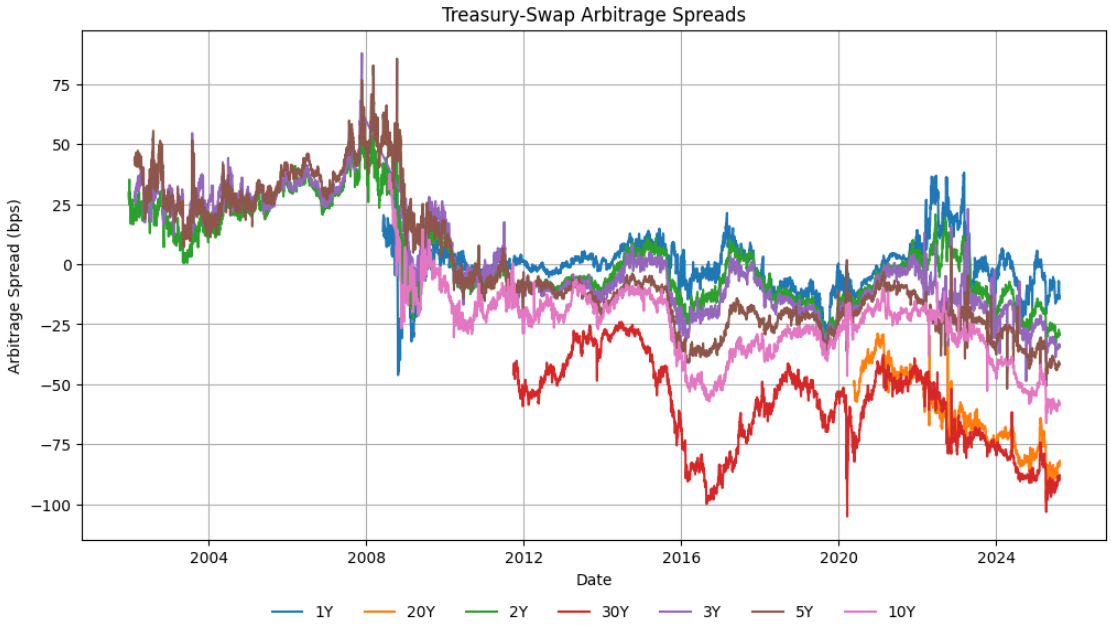
Sources: Bloomberg, Authors' analysis

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. Our 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 that is 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 an exhaustive review of all available forecasting methods, nor are they estimated to their 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`, and `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 in that every model follows the same automated selection recipe. Also, these versions mirror realistic practitioner workflows where hand tuning dozens of models across hundreds of series would not be feasible. Unless otherwise noted, references to each model below therefore correspond to its

Auto implementation.

- Historical Average (the benchmark): forecasts every future value as the training sample mean of each series. It is untuned and serves as the baseline against which R^2_{oos} is defined.
- 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 multiple 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 regularization, following Zeng et al. (2022).
- NLinear: Applies the normalization trick from Zeng et al. (2022) to handle distribution shifts while the auto wrapper tunes the deseasonalization and residual learning configuration.
- VanillaTransformer: A transformer based architecture that utilizes 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 summarizes 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 emphasizes 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 6 through 10.

Table 5: Key Properties of Baseline Forecasting Models

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

Note: 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 (https://nixtlaverse.nixtla.io/).

Sources: Authors' analysis

4.2 An Illustration: What Do These Forecasts Look Like?

Before turning to the formal evaluation, Figure 7 illustrates how differently these model families behave when extrapolating two contrasting types of data. To avoid singling out any individual series, each panel shows a cross-sectional average, and we forecast every entity in the panel and then average both the realized series and the per method forecasts across entities. Panel (a) is the Treasury spot-futures (cash-futures) basis of Figure 5, averaged across its five tenors and forecast over the twelve months of 2019 (trained through December 2018). Panel (b) is bank holding company (BHC) cash liquidity,

averaged across the BHCs that report continuously over the window and forecast over the twelve quarters of 2017–2019 (trained through 2016). Both panels show the same representative quartet with one model from each family in our roster that is the classical statistical workhorses ARIMA and Theta, the strongest hybrid linear method on basis spreads (NLinear), and the pure deep learning method with the strongest basis spread R^2_{oos} (N-BEATS) per the basis spread panel of Table 10. We hold the quartet fixed across panels for comparability and train the neural models globally over each panel's entities, consistent with our univariate-global design.

This exhibit is deliberately not an accuracy test. Whereas the benchmark evaluates each series one step ahead and rolls the origin forward (Section 4), here, we generate a single 12-step path so that the qualitative differences in shape are visible. The contrast across panels is the point. On the basis spread (Panel a) the methods fan out, and ARIMA and Theta anchor near recent levels and mean-revert while the hybrid and deep learning models trace more articulated paths that bend with the recent dynamics. On bank cash liquidity (Panel b), by contrast, the four methods cluster much more tightly around the realized average, mirroring our finding that classical and machine learning methods are roughly competitive on bank and intermediary indicators. None of this establishes which method forecasts best, which is the role of the error metrics in Section 5, but it conveys what a forecast concretely means for each family and why the appropriate default is task specific.

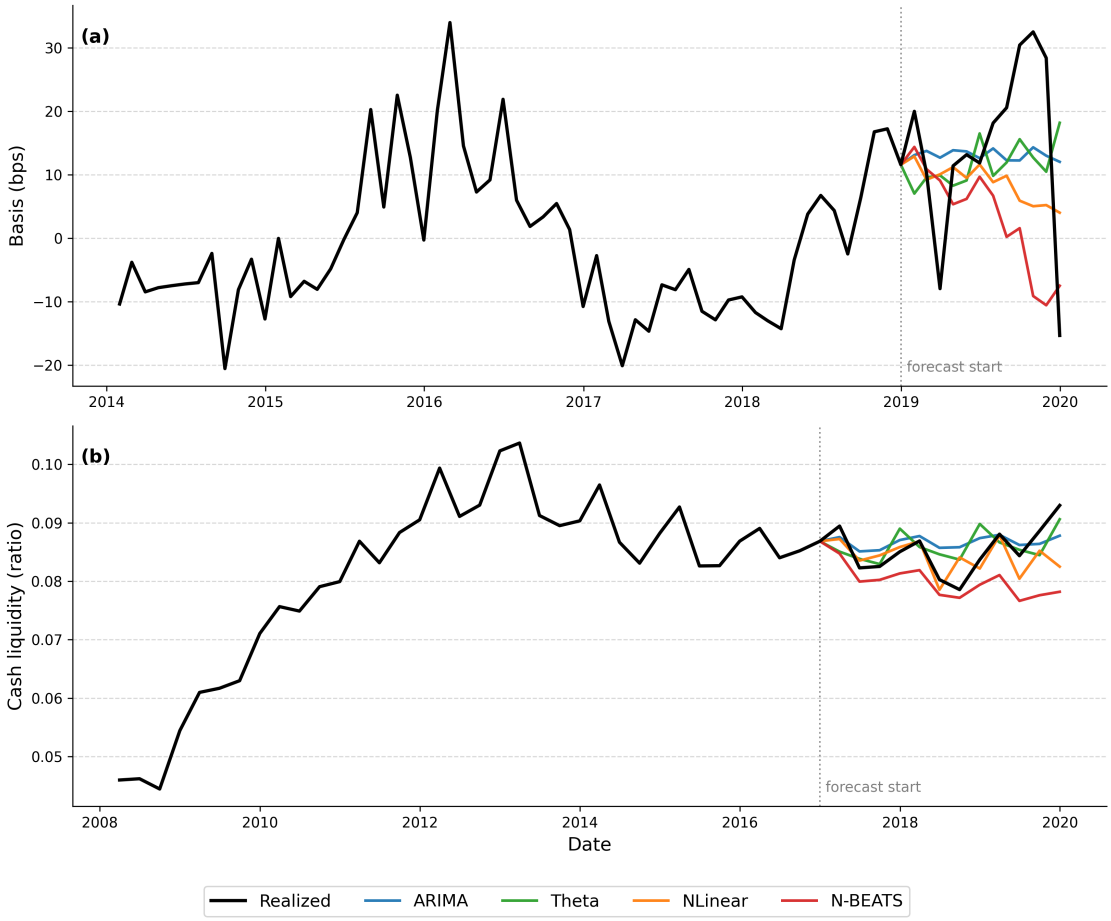
4.3 Data Preprocessing

To ensure fair model comparisons, our forecasting system applies a unified preprocessing pipeline before any model is trained. We summarize the rules below and, for each dataset, report retention statistics, median series lengths, and adaptive minimum observation thresholds in Table 15 of Appendix B.

Filtering Methodology and Fairness Guarantees. We drop any series too short to give a model a usable train/test split, and we hand the same filtered panel to every model. The length cutoff depends on the dataset in that a series must have at least as many observations as the forecast horizon plus a six-observation buffer, and never fewer than 18 observations (seasonal series must clear a higher bar so that a full seasonal cycle remains). The one exception protects small datasets. If a dataset has ten or fewer entities, we skip filtering entirely and keep all of them so that a whole category is not lost. Because the cutoff depends only on series length and not on the model, the statistical and neural estimators are always compared on identical data.

Technical Implementation. The pipeline first normalizes timestamps (e.g., aligns month-end series to true month-ends) and builds a canonical date grid, filling gaps per entity. We then split each series into train and test windows using frequency specific horizons; daily data are forecast roughly a calendar

Figure 7: Illustrative Forecasts: Basis Spread Versus Bank Cash Liquidity



Note: Each panel plots a cross-sectional average: every entity is forecast individually, and the realized series and per method forecasts are then averaged across entities. *Panel (a):* Treasury spot-futures basis averaged across its five tenors (monthly), trained through December 2018 (dotted line) and forecast over 2019. *Panel (b):* BHC cash liquidity averaged across banks reporting continuously over the window (quarterly), trained through 2016 Q4 and forecast over 2017–2019. One representative model is shown from each family: the classical statistical methods (ARIMA, Theta), the best hybrid linear method on basis spreads (NLinear), and the best pure deep learning method on basis spreads (N-BEATS). The figure illustrates the qualitative shape of each family's forecast and is not an accuracy comparison; for that, see Section 5.

Sources: Bloomberg, Call Reports, Authors' analysis

month ahead (30 calendar or 21 trading days), monthly data one month ahead, and quarterly data one quarter ahead. The horizon sets the coverage each series must meet as monthly panels need roughly 16 observations in total but only a single hold-out point while daily and business day panels must supply 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 that fails validation after imputation is discarded. The resulting train/test slices are the single source of truth for every model family.

Forecast Horizons and Cross-Validation. All models, statistical and neural alike, are evaluated with rolling origin cross-validation. The hold-out horizon matches the per-frequency target above, and the step size equals the horizon, so successive windows do not overlap. We use up to six windows, but the actual count is capped by the shortest surviving series in the dataset. For a monthly panel, this means forecasting one month ahead, then rolling the origin forward one month and forecasting again, for up to the last six month-ends; a daily panel instead forecasts the next 21 trading days, rolls forward 21 days, and repeats. A panel whose shortest series has room for only one such window therefore yields a single end-of-sample forecast, while longer panels yield six. 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 evaluation windows so long that they exhaust shorter financial histories, and it provides repeated, out-of-sample checks that guard against overfitting.

Impact and Necessity. How much a panel loses to the length filters depends on its level of aggregation. Portfolio- and index-level panels keep nearly all their series (CIP spreads, TIPS/Treasury bases, and Treasury swap spreads retain 100%, and the CDS-bond basis panel keeps 2,608 of 2,776 entities (93.9%)). Monthly portfolio returns are similarly intact with the one exception being the CDS portfolio, which retains 4 of 20 entities (20%). Disaggregated panels bear the brunt of the filters, and individual CDS contracts shrink from 6,552 to 234 series (3.6% retention), corporate bonds from 53,389 to 28,581 (53.5%), and Treasury bond securities from 2,054 to 1,912 (93.1%). Regulatory filings sit in between as 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 artifacts of uneven preprocessing.

Forecast Stabilization. Three safeguards prevent occasional pathological fits from dominating the squared-error metrics. Every point forecast, classical and neural alike, is clipped to a per-series band of [train min − IQR, train max + IQR] computed only from observations available before the first forecast origin, so no test information enters the bound. Each neural model's final forecast averages five fits that differ only in random seed, and hyperparameters are selected on a validation window immediately preceding the test windows (twelve months for monthly panels, with analogous lengths at other frequencies), with weight decay included in the search. These choices leave accurate forecasts untouched; they bind only when a model would otherwise extrapolate far outside the range of its own training data.

4.4 Error Metrics

We evaluate forecasting performance using two complementary scale-free error metrics, Mean Absolute Scaled Error (MASE) and out-of-sample R^2 (R^2_{oos}), and we additionally report Root Mean Square Error (RMSE) to convey the magnitude of forecast errors in each series' original units. 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). MASE and R^2_{oos} are both scale-free and benchmark model performance against simple baselines, but they differ in their treatment of errors because MASE uses absolute errors (robust to outliers), while R^2_{oos} uses squared errors (emphasizing large misses). RMSE, defined as the square root of the mean squared forecast error, is reported in levels and therefore not comparable across datasets; we include it because error magnitude is directly relevant for risk management applications.

Notation. Let $\{y_{s,t}\}$ denote the target value for series s at time t, $\{\hat{y}_{s,t}\}$ the corresponding forecast, T_s the length of the test window for series s, and $\bar{y}_{\text{train},s}$ the training sample mean of series s. When a metric is computed for a single series, we suppress the s subscript.

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.

Out-of-Sample R^2 (R^2_{oos}). For each (dataset, model) cell we report the pooled, panel-wide R^2_{oos}:

$$R^2_{\text{oos}} = 1 - \frac{\sum_s \sum_{t=1}^{T_s} (y_{s,t} - \hat{y}_{s,t})^2}{\sum_s \sum_{t=1}^{T_s} (y_{s,t} - \bar{y}_{\text{train},s})^2}.$$

Squared errors are summed across all series and all test-window observations in the dataset; the benchmark is each series' own training sample mean (the Historical Average baseline). This is the standard panel definition in the asset-pricing forecasting literature (Campbell and Thompson, 2008; Gu, Kelly, and Xiu, 2020). Positive values indicate the model outperforms the historical-average benchmark on the dataset as a whole; zero or negative values indicate underperformance. We pool rather than averaging per series R^2 values because, on disaggregated panels with heterogeneous variance (e.g., CRSP equity returns, individual TRACE bonds, CDS contracts), per series ratios are sensitive to a handful of low variance series whose benchmark $\sum_t (y_{s,t} - \bar{y}_{\text{train},s})^2$ is small, which can dominate a simple mean.

5 Baseline Results and Analysis

We present forecasting results across all datasets and models in three tables. The Mean Absolute Scaled Error (MASE) provides scale-free comparison across different time series. The Root Mean Square Error (RMSE) captures the magnitude of forecasting errors in original units. The out-of-sample R^2 (R^2_{oos}) measures predictive power relative to the historical average benchmark.

Table 6 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 challenges for forecasting.

Figure 8 provides a visual representation of the MASE results from Table 6. The color-coded heatmap facilitates the identification of patterns in model performance across different dataset types, with cooler colors (blues) indicating better performance and warmer colors (reds) indicating poorer performance.

Table 7 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 8 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.

Table 6: MASE Results by Dataset and Model

	HistAvg	ARIMA	SES	Theta	DLinear	DeepAR	KAN	NBEATS	NHITS	NLinear	TiDE	Transformer
Basis Spreads												
CDS-Bond	2.23	1.22	1.65	1.19	1.16	2.06	1.17	**1.15**	1.16	1.17	1.16	1.17
CIP	0.60	0.35	0.45	0.41	0.36	0.49	0.34	**0.25**	0.30	0.36	0.36	0.40
TIPS-Treasury	1.78	0.45	0.94	**0.37**	0.48	0.70	0.44	0.47	0.48	0.45	0.49	0.38
Treasury-SF	0.95	0.99	1.15	0.89	1.10	0.96	0.85	0.90	0.77	**0.70**	0.82	1.01
Treasury-Swap	2.63	0.29	0.69	0.30	0.27	0.50	0.37	0.31	0.27	**0.25**	0.27	0.32
Returns (Portfolios)												
CDS Portfolio	0.64	0.64	0.81	0.81	0.63	0.61	0.61	0.63	0.56	0.84	0.64	**0.55**
Corporate Portfolio	**0.85**	0.90	0.87	0.88	0.99	1.07	1.12	1.07	1.02	0.97	1.02	0.88
FF25 Size-BM	1.41	**1.41**	1.45	1.45	1.45	1.46	1.45	1.44	1.46	1.46	1.45	1.44
SPX Options Portfolios	0.90	0.88	0.85	0.88	0.79	0.84	0.82	0.81	0.83	0.83	0.79	**0.77**
Treasury Portfolio	**0.52**	0.66	0.56	0.55	0.66	0.68	0.65	0.60	0.52	0.68	0.56	0.59
Returns (Disaggregated)												
CDS Contract	1.91	1.70	1.78	1.61	1.51	1.73	1.57	1.53	**1.51**	1.51	1.52	1.54
CRSP Stock	0.87	0.89	0.89	0.90	0.86	0.86	0.86	0.86	0.86	0.87	0.86	**0.86**
CRSP Stock (ex-div)	0.87	0.89	0.89	0.90	0.86	0.86	0.86	0.86	0.87	0.87	0.86	**0.86**
Commodity	0.52	0.54	0.52	0.52	0.53	0.52	0.55	0.57	**0.52**	0.53	0.54	0.53
Corporate Bond	0.62	0.61	0.60	0.58	0.60	0.62	**0.57**	0.57	0.62	0.59	0.59	0.57
FX	0.54	0.54	0.51	0.51	0.50	0.52	0.55	0.62	0.55	0.55	**0.49**	0.51
Treasuries	0.41	0.35	0.29	0.25	**0.21**	0.27	0.30	0.23	0.30	0.28	0.30	0.26
Other												
BHC Cash Liquidity	1.82	0.85	1.15	0.84	0.82	0.92	0.81	**0.79**	0.80	0.81	0.82	0.82
BHC Leverage	2.66	1.09	1.63	1.08	1.06	1.27	1.05	**1.03**	1.03	1.06	1.06	1.06
Bank Cash Liquidity	1.83	0.83	1.09	0.81	0.79	0.87	0.79	0.78	**0.77**	0.79	0.80	0.80
Bank Leverage	3.19	1.46	2.04	1.45	1.43	1.91	1.43	**1.40**	1.41	1.43	1.43	1.44
HKM All Factor	0.57	0.61	0.58	0.58	0.45	0.82	1.04	0.78	0.31	0.36	**0.28**	0.90
HKM Daily Factor	0.69	0.69	0.70	0.70	0.70	0.72	**0.67**	0.68	0.68	0.70	0.70	0.70
HKM Monthly Factor	0.43	0.42	0.43	0.43	0.47	0.56	0.62	0.52	**0.40**	0.49	0.42	0.45
Treasury Yield Curve	11.26	0.77	1.09	0.76	0.76	1.16	0.83	**0.75**	0.77	0.78	0.87	0.81

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, Compustat, Call Reports, FINRA TRACE, OptionMetrics, S&P Global, Open Source Bond Asset Pricing, He, Kelly, and Manela (2017), Authors' analysis

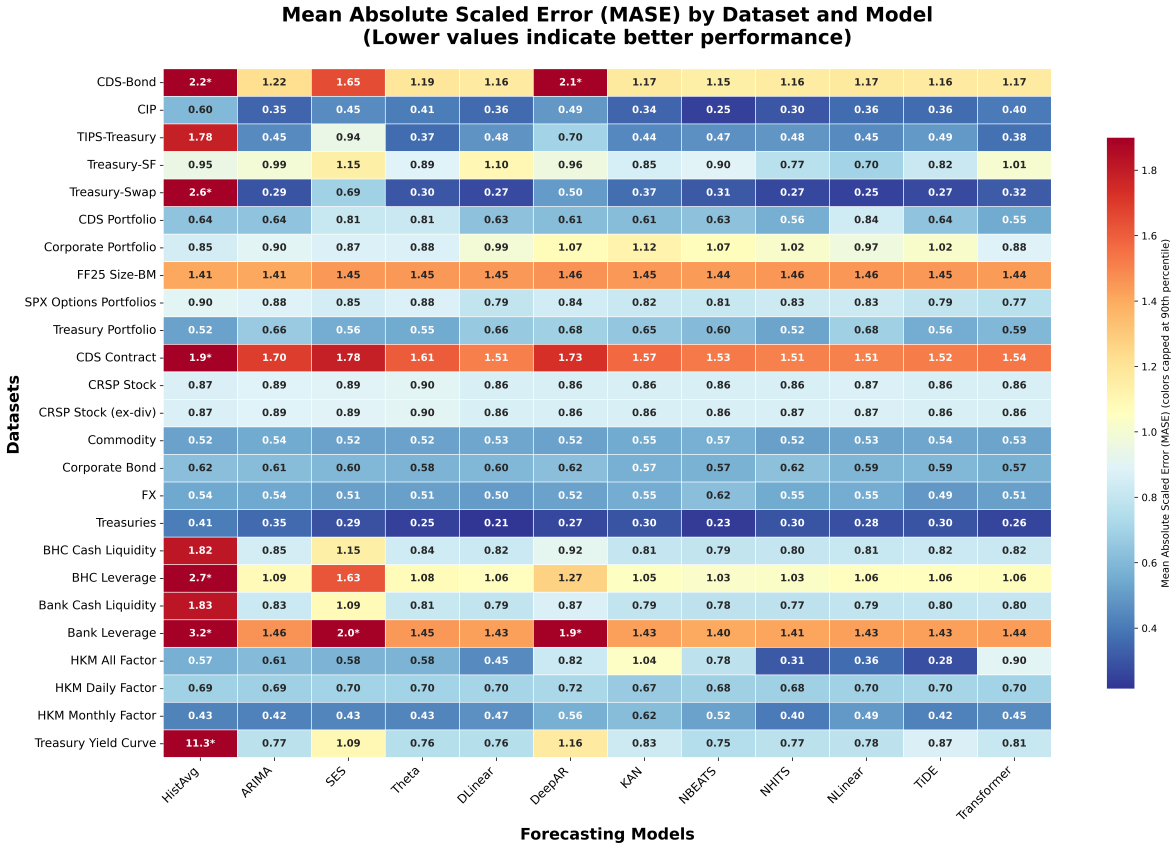
Table 7: RMSE Results by Dataset and Model

	HistAvg	ARIMA	SES	Theta	DLinear	DeepAR	KAN	NBEATS	NHITS	NLinear	TiDE	Transformer
Basis Spreads												
CDS-Bond	84.70	52.47	66.03	50.77	53.28	82.25	55.84	50.07	50.34	49.89	**49.86**	50.99
CIP	12.74	**8.08**	9.89	9.28	8.59	10.07	9.28	9.54	8.80	9.42	9.60	8.82
TIPS-Treasury	22.78	6.58	13.18	**5.47**	6.85	11.68	7.39	5.91	6.76	6.53	7.50	6.41
Treasury-SF	42.23	42.25	51.26	39.92	49.83	35.27	30.68	**28.30**	46.21	32.50	41.53	35.34
Treasury-Swap	29.19	4.26	8.08	4.11	**3.50**	6.31	6.17	3.58	4.00	3.80	3.79	4.06
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	**0.02**	0.02	0.02	0.02	0.02	0.03	0.02	0.03	0.02	0.02	0.02	0.02
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	0.04	0.04	0.04	0.04	0.04	**0.03**	0.04	0.04	0.04	0.03	0.04	0.03
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	**0.21**	0.21	0.21	0.21	0.21	0.21	0.21
CRSP Stock (ex-div)	0.21	0.21	0.21	0.21	0.21	**0.21**	0.21	0.21	0.21	0.21	0.21	0.21
Commodity	**0.05**	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.05	0.05	0.05
Corporate Bond	0.02	0.02	0.02	0.02	0.02	**0.02**	0.02	0.02	0.02	0.02	0.02	0.02
FX	0.00	0.00	0.00	0.00	0.00	**0.00**	0.00	0.01	0.00	0.00	0.00	0.00
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	0.02	0.02	0.02	**0.02**	0.02	0.02	0.02
BHC Leverage	5.34	3.85	4.43	3.82	3.80	4.89	3.95	**3.78**	3.79	3.78	3.79	3.88
Bank Cash Liquidity	0.05	0.03	0.03	0.03	0.03	0.03	0.03	**0.02**	0.02	0.02	0.03	0.03
Bank Leverage	9.47	7.52	8.29	7.48	7.62	8.77	7.57	**7.43**	7.43	7.44	7.63	7.46
HKM All Factor	0.04	0.05	0.05	0.05	0.04	0.08	0.06	0.08	0.06	0.04	**0.04**	0.04
HKM Daily Factor	0.01	0.01	0.01	0.01	0.01	0.01	0.01	**0.01**	0.01	0.01	0.01	0.01
HKM Monthly Factor	0.04	0.03	0.04	0.04	0.03	0.04	**0.03**	0.05	0.04	0.04	0.04	0.03
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, Compustat, Call Reports, FINRA TRACE, OptionMetrics, S&P Global, Open Source Bond Asset Pricing, He, Kelly, and Manela (2017), Authors' analysis

Figure 8: MASE Results Heatmap by Dataset and Model



Note: Cooler colors (blue) indicate better performance (lower error), while warmer colors (red) indicate poorer performance (higher error). Color scaling is capped at the 90th percentile to highlight differences among the majority of values; extreme outliers are marked with asterisks (*) and show actual values.

Sources: Bloomberg, Board of Governors of the Federal Reserve System, Center for Research in Security Prices, Compustat, Call Reports, FINRA TRACE, OptionMetrics, S&P Global, Open Source Bond Asset Pricing, He, Kelly, and Manela (2017), Authors' analysis

Table 8: 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	0.57	0.34	0.59	0.61	-0.02	0.56	0.63	0.63	**0.63**	0.63	0.62
CIP	0.00	**0.45**	0.33	0.39	0.45	0.30	0.21	0.23	0.41	0.30	0.35	0.41
TIPS-Treasury	0.00	0.92	0.65	**0.94**	0.91	0.69	0.89	0.93	0.91	0.92	0.89	0.92
Treasury-SF	0.00	-0.63	-0.73	-0.29	-0.43	0.16	0.46	**0.48**	-0.63	0.17	-0.30	0.02
Treasury-Swap	0.00	0.98	0.93	0.98	**0.99**	0.96	0.96	0.99	0.98	0.98	0.98	0.98
Returns (Portfolios)												
CDS Portfolio	**0.00**	-0.01	-0.49	-0.50	-0.54	-0.13	-0.62	-5.40	-0.55	-4.54	-0.53	-0.39
Corporate Portfolio	**0.00**	-0.08	-0.08	-0.09	-0.64	-1.61	-0.34	-1.12	-0.62	-0.32	-0.57	-0.01
FF25 Size-BM	0.00	**0.00**	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	-0.02	-0.02	-0.01	-0.01
SPX Options Portfolios	0.00	-0.03	-0.03	-0.06	0.01	**0.25**	-0.08	-0.08	-0.07	0.19	-0.25	0.06
Treasury Portfolio	0.00	-0.59	-0.13	-0.13	-0.56	**0.03**	-0.48	-0.63	-0.59	-0.23	-0.74	-0.67
Returns (Disaggregated)												
CDS Contract	0.00	-0.02	-0.04	-0.02	0.02	-0.02	**0.04**	0.03	0.03	0.02	0.01	0.01
CRSP Stock	0.00	-0.04	-0.02	-0.05	0.02	0.03	0.03	0.03	0.02	0.02	0.03	**0.03**
CRSP Stock (ex-div)	0.00	-0.04	-0.02	-0.05	0.02	0.03	0.03	0.03	0.03	0.02	0.02	**0.03**
Commodity	**0.00**	-0.07	-0.10	-0.10	-0.04	-0.06	-0.07	-0.17	-0.15	-0.06	-0.14	-0.09
Corporate Bond	0.00	-0.09	-0.05	-0.08	-0.01	0.00	0.01	0.00	**0.02**	-0.03	0.01	0.01
FX	0.00	0.01	0.01	0.01	0.01	**0.06**	-0.14	-0.81	-0.15	-0.64	-0.13	-0.04
Treasuries	0.00	-0.02	0.07	0.06	-0.60	**0.10**	-0.54	-0.38	-0.39	-0.42	-0.54	-0.26
Other												
BHC Cash Liquidity	0.00	0.74	0.55	0.74	0.77	0.73	0.77	0.78	**0.78**	0.77	0.77	0.75
BHC Leverage	0.00	0.03	0.01	0.03	0.04	-0.01	0.04	0.04	0.04	**0.04**	0.04	0.03
Bank Cash Liquidity	0.00	0.68	0.49	0.68	0.70	0.63	0.70	**0.72**	0.72	0.71	0.71	0.67
Bank Leverage	0.00	-0.02	-0.03	-0.02	0.01	0.00	0.01	0.01	0.01	0.01	0.01	**0.01**
HKM All Factor	0.00	-0.24	-0.16	-0.16	0.09	-2.46	-1.05	-3.03	-1.36	0.21	**0.22**	0.04
HKM Daily Factor	0.00	0.02	0.04	0.03	0.03	-0.01	0.08	**0.08**	0.04	0.04	0.03	0.04
HKM Monthly Factor	0.00	0.03	-0.04	-0.04	0.10	-0.44	**0.25**	-0.96	-0.06	-0.45	-0.49	0.09
Treasury Yield Curve	0.00	0.99	0.99	0.99	0.99	0.99	0.99	**0.99**	0.99	0.99	0.99	0.99

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. Bolded values highlight the best performing model for each dataset. – indicates missing results.

Sources: Bloomberg, Board of Governors of the Federal Reserve System, Center for Research in Security Prices, Compustat, Call Reports, FINRA TRACE, OptionMetrics, S&P Global, Open Source Bond Asset Pricing, He, Kelly, and Manela (2017), Authors' analysis

Figure 9 visualizes the out-of-sample R^2 results, highlighting which datasets and models deliver the 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 9 reports median and mean statistics for MASE and out-of-sample R^2. Because MASE is the workhorse metric in the forecasting literature and is used to anchor the discussion, `DLinear` delivers the lowest median MASE (0.755), with `NHITS` (0.766), `ARIMA` (0.772), `NBEATS` (0.775), and `NLinear` (0.778) within a few hundredths. The auto-deep models comfortably beat the Historical Average (0.873) and simple exponential smoothing (0.874), though the classical `ARIMA` and `Theta` (0.810) remain firmly in the leading pack. When we pivot to R^2_{oos}, the magnitudes shrink dramatically because *every* model in the suite, including the classical baselines, posts a median R^2_{oos} within 0.04 of zero. Mean R^2_{oos} is modestly positive for most of the suite (`Transformer` 0.171, `Theta` 0.155, `ARIMA` 0.142, `DLinear` 0.118), with `NBEATS` (-0.264) the one clear laggard. This is an artifact of a single small portfolio panel on which the level of a near-constant series is misjudged. The takeaway from the pivot is that an aggregate R^2_{oos} near zero is the rule rather than the exception in this univariate benchmark; the historical-average benchmark itself is already the right forecaster for the typical panel and gains on MASE come from sharpening the level forecast rather than from extracting time series predictability that would translate into a large positive R^2_{oos}.

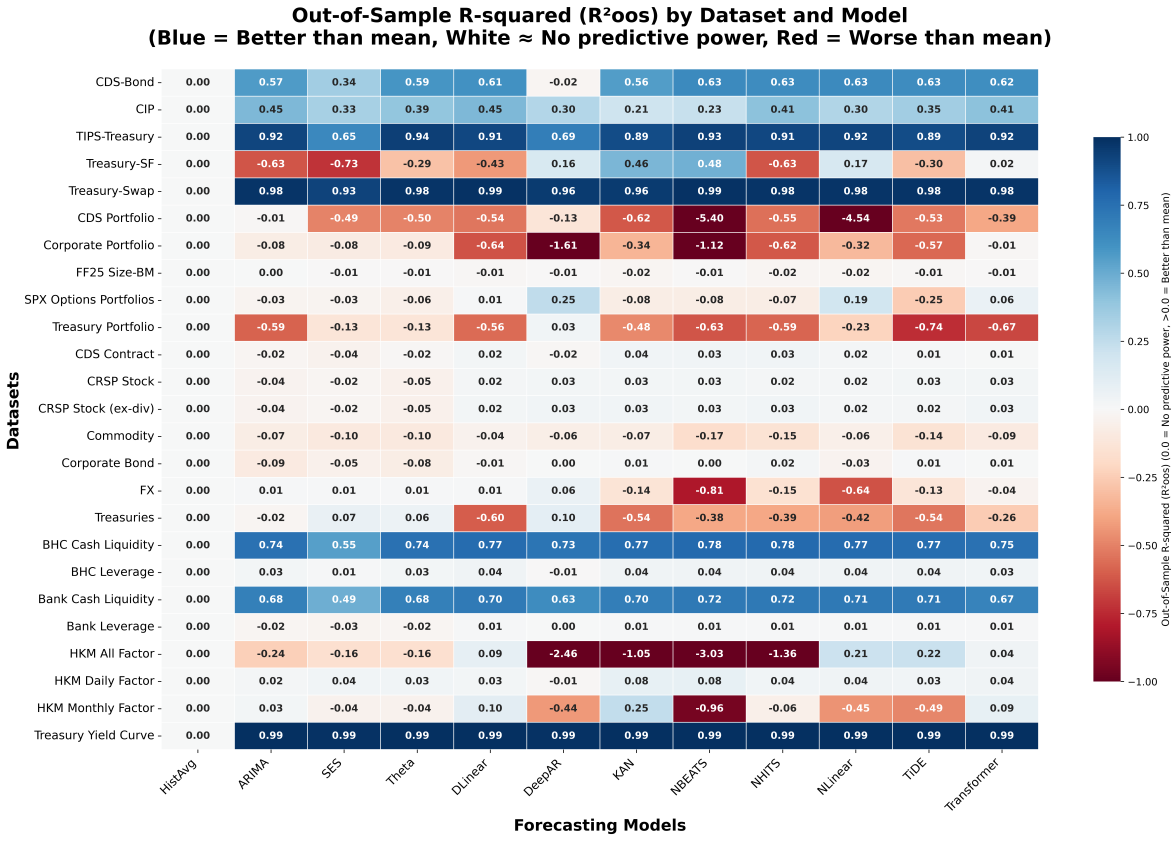


Figure 9: Out-of-Sample R^2 (R^2_{oos}) Heatmap by Dataset and Model

Note: Warmer colors indicate larger improvements over the Historical 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, Compustat, Call Reports, FINRA TRACE, OptionMetrics, S&P Global, Open Source Bond Asset Pricing, He, Kelly, and Manela (2017), Authors' analysis

Table 9: Overall Model Performance Summary: MASE and R² Across All Datasets

	N	Med MASE	Mean MASE	Med R²	Mean R²
HistAvg	25	0.873	1.628	0.000	0.000
ARIMA	25	0.772	0.802	-0.009	0.142
SES	25	0.874	0.945	-0.023	0.099
Theta	25	0.810	0.786	-0.015	0.155
DLinear	25	**0.755**	0.778	0.024	0.118
DeepAR	25	0.844	0.920	0.031	0.009
KAN	25	0.807	0.813	0.031	0.107
NBEATS	25	0.775	0.784	0.025	-0.264
NHITS	25	0.766	**0.751**	0.017	0.042
NLinear	25	0.778	0.773	0.023	-0.027
TiDE	25	0.792	0.764	0.012	0.080
Transformer	25	0.802	0.784	**0.032**	**0.171**

Note: Summary statistics across all datasets for each model. MASE shows absolute performance (lower is better), and R² 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, Compustat, Call Reports, FINRA TRACE, OptionMetrics, S&P Global, Open Source Bond Asset Pricing, He, Kelly, and Manela (2017), Authors' analysis

Table 10 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.

Table 10: Model Performance by Dataset Category

Category	Model	N	Med MASE	Mean MASE	Med R^2	Mean R^2
Basis Spreads	ARIMA	5	0.446	0.660	0.569	0.458
	DLinear	5	0.478	0.674	0.609	0.504
	DeepAR	5	0.696	0.942	0.295	0.417
	HistAvg	5	1.784	1.638	0.000	0.000
	KAN	5	0.444	0.637	0.559	0.614
	NBEATS	5	0.469	0.614	0.633	**0.653**
	NHITS	5	0.480	0.595	0.631	0.462
	NLinear	5	0.452	**0.587**	**0.633**	0.600
	SES	5	0.939	0.977	0.344	0.304
	Theta	5	0.411	0.632	0.590	0.522
	TiDE	5	0.494	0.621	0.633	0.512
	Transformer	5	**0.400**	0.655	0.617	0.590
Returns	ARIMA	12	0.773	0.836	-0.032	-0.081
	DLinear	12	0.726	0.800	-0.008	-0.193
	DeepAR	12	0.764	0.838	**0.017**	-0.108
	HistAvg	12	0.746	0.839	0.000	**0.000**
	KAN	12	0.736	0.826	-0.075	-0.182
	NBEATS	12	0.722	0.816	-0.125	-0.709
	NHITS	12	0.727	0.801	-0.110	-0.203
	NLinear	12	0.835	0.831	-0.046	-0.501
	SES	12	0.832	0.836	-0.034	-0.074
	Theta	12	0.845	0.820	-0.054	-0.083
	TiDE	12	0.716	0.801	-0.136	-0.237
	Transformer	12	**0.679**	**0.779**	-0.008	-0.109
Other	ARIMA	8	0.799	0.840	0.028	0.279
	DLinear	8	0.775	0.810	0.097	**0.343**
	DeepAR	8	0.893	1.029	-0.004	-0.070
	HistAvg	8	1.823	2.806	0.000	0.000
	KAN	8	0.820	0.904	**0.163**	0.223
	NBEATS	8	0.776	0.843	0.062	-0.170
	NHITS	8	**0.772**	**0.774**	0.041	0.145
	NLinear	8	0.786	0.803	0.128	0.292
	SES	8	1.090	1.089	0.024	0.231
	Theta	8	0.784	0.830	0.031	0.282
	TiDE	8	0.809	0.798	0.134	0.286
	Transformer	8	0.815	0.873	0.065	0.328

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

Sources: Bloomberg, Board of Governors of the Federal Reserve System, Center for Research in Security Prices, Compustat, Call Reports, FINRA TRACE, OptionMetrics, S&P Global, Open Source Bond Asset Pricing, He, Kelly, and Manela (2017), Authors' analysis

Error-metric choice is especially consequential for the returns category. Out-of-sample R^2 is the most informative measure here because the Historical Average is already a difficult benchmark to beat for asset returns. The table shows that nearly every model evaluated here, including the auto deep networks, posts R^2_{oos} values clustered around zero for equity, bond, FX, and option returns. We emphasize that this finding pertains to the class of models that we evaluate, to include univariate global forecasts without exogenous regressors (see Section 2.4). Within that class, simply forecasting the historical mean is nearly optimal, an outcome that mirrors decades of empirical asset-pricing research and reinforces the difficulty in extracting persistent return predictability without conditioning information. Extending the benchmark to methods that admit exogenous regressors (ARIMAX, DeepAR with covariates, Temporal Fusion Transformers) or to multivariate panel methods (VARs, state-space

systems) could, in principle, close part of this gap; we leave that comparison for follow-up work.

Basis spreads tell a different story. On median MASE the `Transformer` leads (0.400), with the classical `Theta` (0.411), `KAN` (0.444), `ARIMA` (0.446), and `NLinear` (0.452) clustered closely behind. The deep architectures excel here 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., the quarter-end funding pressure documented by Du, Tepper, and Verdelhan, 2018) and medium-term mean reversion, and the auto wrappers concentrate the search on those horizons. The separation appears on R^2_{oos}: the deep and hybrid models (NLinear, NBEATS, TiDE, NHITS) all post median R^2_{oos} of about 0.63, ahead of the best classical contenders (`Theta` at 0.59, `ARIMA` at 0.57), and `NBEATS` posts the strongest mean (0.65). The modern architectures thus deliver large improvements over the Historical Average and simple exponential smoothing baselines and a consistent, if not overwhelming, edge over classical methods on these spreads.

The Other category, which includes bank regulatory filings, volatility indicators, and macro-financial composites, is more nuanced. Median MASE is tightly bunched as `NHITS` (0.772), `DLinear` (0.775), `NBEATS` (0.776), and the classical `Theta` (0.784) sit within roughly a hundredth of one another. On R^2_{oos} the hybrid and lighter deep models hold a modest edge—medians of 0.10–0.16 for `KAN`, `TiDE`, `NLinear`, and `DLinear` versus 0.03 for `Theta` and `ARIMA`—and mean R^2_{oos} is solidly positive for most of the suite (`DLinear` 0.34, `Transformer` 0.33, `NLinear` 0.29), with `DeepAR` and `NBEATS` the laggards. These datasets are typically longer and smoother than the high-volatility returns series but lack the strong seasonal spikes of basis spreads, so parsimonious trend models and the lighter decomposition-based architectures both fit naturally, and the gap between classical and modern methods is correspondingly small.

Taken together, the cross-metric evidence confirms that performance is conditional. MASE-based rankings highlight the strength of the decomposition-based architectures (DLinear, NHITS, NBEATS) on error scales commonly used by practitioners, while out-of-sample R^2 shows most methods beating the historical mean on average across the benchmark, with essentially all of that advantage earned on basis spreads and the smoother bank and intermediary indicators rather than on returns. Analysts selecting a forecast should therefore align the evaluation metric with the economic decision. R^2_{oos} is appropriate for return forecasting, where the historical mean is a credible trading benchmark. Scaled absolute errors are appropriate for arbitrage spreads where seasonal accuracy matters more than mean-squared scorekeeping and for bank and intermediary indicators where parsimonious level forecasts capture most of the persistent dynamics.

Tables 9–10 rank individual models, but a practitioner often asks a coarser question: As a class, do deep learning architectures actually beat classical statistical methods, or do the recent hybrid

decomposition-based models bridge the gap? Table 11 aggregates the same metrics into four groups of models, using the taxonomy from Table 5 and separating out the Historical Average baseline because it is the benchmark against which R^2_{oos} is defined as *Historical Average*, *classical statistical* (ARIMA, SES, Theta), *hybrid* (DLinear, NLinear), and *deep learning* (DeepAR, N-BEATS, N-HiTS, Vanilla Transformer, TiDE, KAN). Panel A pools across all datasets; Panel B repeats the comparison within each dataset category.

Table 11: Model Performance by Forecasting Method Type

Panel A: Overall (all datasets)

			MASE		R^2	
Model Type	# Models	N	Med	Mean	Med	Mean
Historic Average	1	25	0.873	1.628	0.000	0.000
Classical Statistical	3	75	0.814	0.844	-0.015	**0.132**
Hybrid	2	50	**0.766**	**0.776**	0.024	0.045
Deep Learning	6	150	0.784	0.803	**0.025**	0.024

Panel B: By Dataset Category

				MASE		R^2	
Category	Model Type	# Models	N	Med	Mean	Med	Mean
Basis Spreads	Historic Average	1	5	1.784	1.638	0.000	0.000
	Classical Statistical	3	15	0.691	0.756	0.569	0.428
	Hybrid	2	10	**0.465**	**0.630**	0.621	**0.552**
	Deep Learning	6	30	0.492	0.677	**0.624**	0.541
Returns	Historic Average	1	12	0.746	0.839	**0.000**	**0.000**
	Classical Statistical	3	36	0.832	0.831	-0.038	-0.079
	Hybrid	2	24	0.812	0.816	-0.025	-0.347
	Deep Learning	6	72	**0.726**	**0.810**	-0.029	-0.258
Other	Historic Average	1	8	1.823	2.806	0.000	0.000
	Classical Statistical	3	24	0.818	0.920	0.030	0.264
	Hybrid	2	16	**0.786**	**0.806**	**0.097**	**0.318**
	Deep Learning	6	48	0.805	0.870	0.044	0.124

Note: Each row aggregates every (model × dataset) observation whose model belongs to the indicated method type. # Models counts distinct models in the type and N counts (model × dataset) observations. The Historical Average benchmark is reported in its own row because it is the benchmark against which R^2_{oos} is defined. Lower MASE indicates better performance; higher R^2_{oos} indicates better performance. Best values within each panel (or, in Panel B, within each dataset category) are bolded.

Sources: Bloomberg, Board of Governors of the Federal Reserve System, Center for Research in Security Prices, Compustat, Call Reports, FINRA TRACE, OptionMetrics, S&P Global, Open Source Bond Asset Pricing, He, Kelly, and Manela (2017), Authors' analysis

The aggregated view sharpens the qualitative picture from the model-level tables. On the MASE scale the three method families finish much closer than the individual rankings might suggest that hybrid wins median MASE (0.766) with deep learning (0.784) and classical (0.814) within a few hundredths, and all three comfortably ahead of the benchmark (0.873). On out-of-sample R^2, every family posts a *positive* overall mean (classical 0.132, hybrid 0.045, deep learning 0.024) with medians within 0.03 of zero; no family systematically loses to the historical average once forecasts are stabilized as described in Section 4. Panel B shows where that aggregate advantage is earned. In the basis spread category, hybrid and deep learning models exploit the seasonal structure of funding-related spreads and beat the classical aggregate on median MASE (0.465 and 0.492 versus 0.691), with mean R^2_{oos} of 0.55 and 0.54 (versus 0.43). On the Returns category every family clusters near the benchmark on MASE, with deep learning posting the best median (0.726), while median R^2_{oos} sits at roughly -0.03

for all three families. Relative to the historical mean, returns remain essentially unforecastable. The remaining negative *mean* R^2_{oos} values on returns (classical -0.08, deep learning -0.26, hybrid -0.35) trace to a handful of small, low-volatility portfolio panels on which an in-range level bias is heavily punished by the squared-error metric; the median, which is robust to those cases, is the more informative summary. The Other category (bank regulatory filings and intermediary indicators) modestly favors the hybrid family on both median MASE (0.786) and R^2_{oos} (median 0.097, mean 0.318), with classical methods close behind on the mean (0.264) and deep learning trailing (0.124). The takeaway is not that one family dominates uniformly. Rather, the value add of moving from a classical method to a deep or hybrid architecture is concentrated in segments with persistent seasonal or cross-sectional structure (basis spreads), essentially absent in segments where the historical mean is already an oracle (returns), and modest in segments where parsimony pays off (bank and intermediary indicators).

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. 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 do not consistently outperform simpler alternatives. The newer KAN architecture appears in only a subset of datasets and is not consistently ranked. Notably, linear deep models like DLinear and NLinear achieve competitive results on many datasets despite their simplicity.

DeepAR exhibits relatively weak out-of-sample results across several datasets relative to the other deep learning architectures (Table 9). All neural models in this benchmark use Nixtla's Auto implementations, which automate hyperparameter selection through Bayesian optimization and related search strategies.[8] The Auto wrappers (AutoDeepAR, AutoNBEATS, AutoNHITS, and others) 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. These searches do not guarantee optimal performance for every individual dataset.

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. Achieving the best forecast for any given dataset may require domain-specific expertise and manual tuning beyond a generic hyperparameter search. The benchmark's value lies not in claiming to provide optimal

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

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, which fall outside the scope of establishing the benchmark itself.

6 Sensitivity to Cleaning Method

Section 5 held one cleaning convention fixed per dataset and asked which forecasting method wins. The motivating premise of FTSFR, however, is that this ranking is entangled with the cleaning recipe itself. We now operationalize that claim. We start with a case study on corporate bonds where a well-documented cleaning convention has a quantified effect on cross-sectional sorts (Section 6.1). We then show that the same pattern recurs across three other asset classes (Section 6.2) that are options portfolios under different filtering depths, Fama-French 25 portfolios under NYSE versus all-CRSP breakpoints, and Treasury bond portfolios under the Gurkaynak-Sack-Wright (2007) filters versus a permissive universe. In every demonstration, the underlying data come from the same source, and the realized return series is held fixed. Only the cleaning convention applied during construction differs. Yet the forecasting leaderboard reshuffles materially, often by enough to flip the practitioner's choice of best model.

6.1 A TRACE Case Study

Dickerson, Robotti, and Rossetti (2024) documents that the bid–ask averaged TRACE prices used to construct the credit spread signal contain market microstructure noise (MMN) that mechanically distorts cross-sectional sorts. The Open Source Bond Asset Pricing panel exposes the relevant variation directly: each bond-month carries both an MMN-biased credit spread (the conventional construction) and an MMN-corrected credit spread (the construction recommended by Dickerson, Robotti, and Rossetti, 2024). The realized monthly return is held fixed across the two conventions, so any downstream difference comes from cleaning the signal alone.

Before turning to forecasting, we confirm that the MMN correction is a meaningful, literature-endorsed cleaning choice rather than an arbitrary perturbation. We hold the credit spread sort fixed and switch only the realized return convention from the standard month-end measure (R^{end}, whose start of period price reuses the same bid–ask averaged quote that defined the signal, so microstructure noise enters both) to the OSBAP month-begin measure (R^{bgn}, which separates the signal and return-start prices by a full bid–ask cycle and breaks the mechanical link). Table 12 shows the apparent credit spread premium falling from 5.66% per year ($t = 1.48$) to 4.18% per year ($t = 1.15$), a 26% drop

that moves it from marginally significant to insignificant. The reduction is modest here only because the OSBAP main panel is already heavily MMN-cleaned; on rawer signals the same bias is far larger, with Dickerson, Robotti, and Rossetti (2024) documenting reductions exceeding 90% on short-term reversal.

Table 12: Replicating Dickerson, Robotti, and Rossetti (2024): Decile-10-Minus-Decile-1 Long–Short by Realized Return Convention

Realised-return convention	$\overline{\text{D10}-\text{D1}}$ (%/mo)	t-stat	Annualised (%)	Sharpe
R^{end} (contaminated)	0.472	(1.48)	5.66	0.31
R^{bgn} (implementable)	0.348	(1.15)	4.18	0.24
Δ (implementable − contaminated)			-1.48	

Note: Each row reports the value-weighted long–short return (decile 10 minus decile 1) from sorting corporate bonds into ten value-weighted credit spread deciles each month using the OSBAP credit spread signal, then computing the realized return under the indicated convention. Both rows use identical bond universes, sort dates, decile assignments, and weighting variables. The only difference is the realized return convention. R^{end} uses month-end TRACE prices for both the signal at t and the return-start at t, so any residual market-microstructure noise enters both; R^{bgn} uses month-begin TRACE prices for the return-start at $t+1$, separating signal-date and return-start prices by a full bid–ask cycle. t-statistics are not Newey–West corrected (i.i.d. standard errors) and are reported only for relative comparison. Sample: 269 months, 2002–2024.

Sources: Open Source Bond Asset Pricing (OSBAP 2025 main panel), Authors' analysis

We now turn to the forecasting question. On the same OSBAP panel we build two parallel panels of value-weighted decile portfolios sorted on each cleaning convention of the signal itself, holding the universe and the realized return series identical. The cross-decile rank correlation between the two signals is 0.98, and the resulting portfolio return time series are 98–99% correlated within decile, yet, the long–short spread (decile 10 minus decile 1) differs by roughly a factor of two (2.8% per year under the MMN-biased signal versus 5.6% per year under the MMN-corrected signal). The implication is that cleaning the signal does not merely reproduce the same portfolio with cosmetic differences but rather changes which bonds occupy each decile in a way that reshuffles the cross-section enough to alter both the apparent premium and the relative ranking of forecasting models. On these two near identical panels, we run the full twelve-model suite used elsewhere in this paper under the same rolling origin cross-validation protocol. Table 13 reports the results.

The four classical baselines (Historical Average, ARIMA, SES, Theta) are indifferent to the cleaning choice ($|\Delta R^2_{\text{oos}}| \leq 0.005$), and five of the eight deep models are nearly as stable (DLinear, TiDE, and Transformer move by less than 0.01; NBEATS and DeepAR by less than 0.08). The remaining three swing by an order of magnitude more. NHITS moves by +0.34 (from −0.32 to +0.01), NLinear by +0.29 (from −0.14 to +0.15), and KAN by −0.28 (from +0.11 to −0.17). Note the signs: Cleaning the signal does not push every deep model in the same direction. Some architectures gain from the corrected signal while others lose. This is the entanglement between the forecasting method and the cleaning method made concrete on a single pipeline.

The reshuffling at the top of the leaderboard depends on the metric. By out-of-sample R^2 the

Table 13: Forecasting Sensitivity to Signal Cleaning: Credit Spread Decile Portfolios

	R^2_{oos}			MASE		
Model	MMN-biased	MMN-corrected	Δ	MMN-biased	MMN-corrected	Δ
HistAvg	0.000	0.000	+0.000	2.168	2.159	-0.009
ARIMA	-0.102	-0.106	-0.004	2.340	2.359	+0.019
SES	0.192	0.197	+0.004	2.005	1.999	-0.005
Theta	0.195	0.198	+0.004	2.002	1.997	-0.005
DLinear	0.123	0.120	-0.004	2.050	2.032	-0.018
DeepAR	**0.269**	**0.347**	+0.078	1.880	1.937	+0.057
KAN	0.109	-0.167	-0.277	**1.767**	2.027	+0.260
NBEATS	-0.258	-0.315	-0.057	1.926	2.450	+0.524
NHITS	-0.321	0.014	+0.335	1.997	1.927	-0.069
NLinear	-0.136	0.151	+0.287	1.958	**1.915**	-0.043
TiDE	0.195	0.188	-0.007	1.957	1.967	+0.010
Transformer	0.170	0.165	-0.006	1.900	2.193	+0.294

Note: Out-of-sample R^2 (R^2_{oos}) and Mean Absolute Scaled Error (MASE) for the twelve forecasting models reported elsewhere in the paper, applied to two value-weighted decile-portfolio panels of corporate bonds. The MMN-biased panel sorts bonds by a credit spread constructed from MMN-contaminated TRACE prices. The MMN-corrected panel applies the correction of Dickerson, Robotti, and Rossetti (2024). The realized return, universe, and cross-validation protocol are held fixed; only the signal used for decile assignment differs. Δ columns report corrected minus biased. Bolded values highlight the best performing model within each level column (R^2_{oos} higher is better; MASE lower is better).

Sources: Open Source Bond Asset Pricing, Authors' analysis

distributional DeepAR model holds the top spot on both panels (0.269 biased, 0.347 corrected), but by MASE, the winner flips. KAN posts the lowest MASE on the MMN-biased panel (1.767) yet falls below the historical average baseline once the signal is cleaned ($R^2_{\text{oos}} = -0.167$), while NLinear takes the corrected panel's lowest MASE (1.915) and moves from negative to solidly positive R^2_{oos}. NHITS makes the reverse journey from the bottom, climbing from the worst R^2_{oos} in the biased ranking (-0.32) to modestly positive on the corrected ranking. A practitioner who selects the best forecaster for credit spread decile portfolios on scaled absolute error will therefore choose a different architecture purely as a function of which cleaning convention they happened to apply. Restricting attention to the simpler family does not help since the four classical baselines are stable across conventions, but none clears $R^2_{\text{oos}} = +0.20$ on either panel, so a researcher wanting the best available forecast in this setting will be drawn toward models that are more cleaning sensitive. There are two qualifications. First, the magnitude and direction of each deep model's swing is partly an interaction with the Auto hyperparameter search; small differences in the bonds that occupy each decile change the cross-validation surface enough that the Auto wrapper selects materially different configurations. Second, the table evaluates only a single forecasting setup (univariate, no exogenous regressors). The directional lesson is robust regardless; ex-ante prediction of which method will be cleaning sensitive is unreliable, holding the data fixed is a prerequisite for comparing methods, and reproducible pipelines like FTSFR are what make that comparison possible.

6.2 Robustness Across Asset Classes

The TRACE case study uses corporate bonds because that is where the cleaning convention has been most precisely quantified in the literature. The same mechanism, however, is not specific to bonds. To check that the lesson is not an artifact of the MMN setting, we replicate the exercise on three additional asset classes where well-known, defensible cleaning conventions admit a natural pairwise comparison in SPX index options, Fama-French 25 portfolios, and U.S. Treasury bond portfolios. In each case, we hold the data source, the universe before filtering, and the realized return series fixed, and vary only one cleaning convention. We then run the same twelve-model suite and rolling origin protocol used in Section 5.

The contrasts span a range of cleaning intensity, which we summarize by how correlated the two variants' within bucket return series remain. The SPX options contrast is the widest (average within bucket correlation 0.68), and we apply the three-level filter stack of Constantinides et al. (2013) either in full or only at Level 1 so the Level-1-only panel retains the extreme-implied-volatility, off-grid-moneyness, and put-call-parity-violating quotes that the canonical recipe drops. The Fama-French 25 contrast sits in the middle (correlation about 0.90), and we switch only the size and book-to-market breakpoints between NYSE only (the canonical convention) and all-CRSP, which loads the small-cap corners with microcaps. The Treasury contrast is the closest to the MMN setup (correlation above 0.99), and we switch only the universe filter between a permissive set and the Gürkaynak, Sack, and Wright (2007) strict filters, which drop callables and recent issues. Table 14 stacks the three panels under the same column layout used in Table 13.

The Options panel (Panel A) is the widest contrast and reads mostly as a level shift with a reordered upper tier. Moving from the permissive L1 panel to the full L3 stack raises MASE for every model, the benchmark included, and lowers R^2_{oos} for nine of the twelve models; only DeepAR (+0.10) and NBEATS (+0.02) gain. `TiDE` posts the best R^2_{oos} on both variants (0.28 and 0.21), but the best-MASE model flips from DLinear on L1 to TiDE on L3, and three architectures cross zero R^2_{oos} (ARIMA, DeepAR, Transformer). The heavier L3 cleaning is also the defensible convention rather than a stylistic preference, and its filters drop deep out of the money and far from maturity options with bid–ask spreads exceeding 50% of price and near-zero volume, quotes no market maker would fill at scale, so the L1 panel admits returns no investor could realize even in principle.

The FF25 panel (Panel B) sits in the middle. Every R^2_{oos} is negative on both variants, consistent with the broader equity return result in Section 5, so there is no leaderboard winner to flip, but the deep models still swing materially and in mixed directions (DeepAR by −0.31, DLinear by +0.17, NHITS by +0.16, Transformer by −0.13, KAN by −0.12) while the classical baselines move by at most 0.10 (SES and Theta by just 0.01). The magnitudes are large enough that a paper claiming a

Table 14: Forecasting Sensitivity to Cleaning Method: Robustness Across Asset Classes

	R^2_{oos}			MASE		
Panel A: Options (CJS 54)						
Model	L1 filters	L3 filters	Δ	L1 filters	L3 filters	Δ
HistAvg	0.000	0.000	+0.000	0.674	0.865	+0.190
ARIMA	0.108	-0.000	-0.109	0.638	0.838	+0.200
SES	0.152	0.017	-0.136	0.613	0.829	+0.216
Theta	0.155	0.008	-0.147	0.607	0.834	+0.226
DLinear	0.264	0.130	-0.134	**0.533**	0.748	+0.215
DeepAR	-0.008	0.093	+0.101	0.543	0.771	+0.227
KAN	0.134	0.061	-0.073	0.592	0.818	+0.226
NBEATS	0.054	0.072	+0.018	0.556	0.760	+0.204
NHITS	0.172	0.132	-0.040	0.643	0.750	+0.108
NLinear	0.233	0.207	-0.027	0.606	0.828	+0.222
TiDE	**0.282**	**0.214**	-0.068	0.548	**0.737**	+0.189
Transformer	0.068	-0.047	-0.115	0.572	0.748	+0.175
Panel B: FF25 (Size x BM)						
Model	NYSE breaks	CRSP breaks	Δ	NYSE breaks	CRSP breaks	Δ
HistAvg	**0.000**	**0.000**	+0.000	**0.879**	**0.897**	+0.018
ARIMA	-0.190	-0.095	+0.095	0.972	0.967	-0.006
SES	-0.106	-0.092	+0.014	0.929	0.953	+0.024
Theta	-0.107	-0.092	+0.014	0.929	0.953	+0.024
DLinear	-0.354	-0.180	+0.174	1.003	1.006	+0.003
DeepAR	-0.139	-0.447	-0.308	0.929	0.907	-0.022
KAN	-0.266	-0.389	-0.123	1.120	1.014	-0.106
NBEATS	-0.373	-0.311	+0.062	1.010	0.986	-0.024
NHITS	-0.529	-0.365	+0.164	0.984	1.038	+0.054
NLinear	-0.049	-0.074	-0.025	0.916	0.984	+0.068
TiDE	-0.204	-0.133	+0.071	0.923	0.926	+0.003
Transformer	-0.036	-0.165	-0.130	0.963	0.995	+0.032
Panel C: Treasury portfolios						
Model	Permissive	GSW strict	Δ	Permissive	GSW strict	Δ
HistAvg	0.000	0.000	+0.000	0.554	0.524	-0.030
ARIMA	0.105	0.119	+0.014	**0.407**	**0.412**	+0.005
SES	0.074	0.082	+0.008	0.440	0.433	-0.007
Theta	0.080	0.088	+0.008	0.417	0.418	+0.001
DLinear	-0.037	0.061	+0.098	0.467	0.456	-0.011
DeepAR	-0.141	-0.363	-0.223	0.469	0.419	-0.050
KAN	0.066	-0.016	-0.082	0.446	0.473	+0.026
NBEATS	-0.034	**0.132**	+0.167	0.510	0.492	-0.018
NHITS	-0.244	-0.025	+0.219	0.485	0.464	-0.021
NLinear	-0.003	0.082	+0.084	0.436	0.432	-0.005
TiDE	0.071	0.071	-0.000	0.464	0.434	-0.030
Transformer	**0.111**	0.072	-0.038	0.435	0.439	+0.005

Note: Out-of-sample R^2 (R^2_{oos}) and Mean Absolute Scaled Error (MASE) for the twelve forecasting models used elsewhere in the paper, applied to three pairs of panels. Panel A: SPX CJS 54-portfolio returns built from options data filtered with Level-1 filters only versus the full Level-1+Level-2+Level-3 stack of Constantinides et al. (2013). Panel B: 5×5 size × book-to-market portfolios with NYSE-only breakpoints (Fama and French 1993 convention) versus all-CRSP breakpoints. Panel C: maturity-bucketed Treasury bond portfolios with a permissive universe (callables and on-the-run issues kept) versus the Gürkaynak, Sack, and Wright (2007) strict universe (callables dropped, on-the-run and first off-the-run dropped for post-1980 securities). Within each panel, the realized return series, holding-period definition, and cross-validation protocol are held fixed; only the cleaning convention differs. Δ columns report the right-hand variant minus the left-hand variant. Bolded values mark the best-performing model within each level column (R^2_{oos} higher is better; MASE lower is better).

Sources: Center for Research in Security Prices, Compustat, OptionMetrics, Authors' analysis

particular deep model works on FF25 could see its claim reversed by switching breakpoint conventions.

The Treasury panel (Panel C) is the closest analogue to the MMN setting, and the two universes are 99%+ correlated within bucket, yet the ranking still reorders. The classical baselines barely move ($|\Delta R^2_{\text{oos}}| \leq 0.014$), with ARIMA second under both filter choices, while the top spot flips between deep architectures in that Transformer leads the permissive universe (0.111) but NBEATS leads the strict one (0.132). Also, the large swings live entirely among the deep models: DeepAR falls by 0.22 while NHITS (+0.22), NBEATS (+0.17), DLinear (+0.10), and NLinear (+0.08) move the other way, with four architectures crossing zero. As in the TRACE case, essentially all the ranking volatility lives among the deep models, with classical baselines essentially stable.

Two qualifications carry over. First, the magnitude of each deep model's swing is partly an interaction with the Auto hyperparameter search. Small universe changes shift the cross-validation surface enough that the Auto wrapper selects materially different configurations. Second, every panel evaluates the same univariate global setup without exogenous regressors. The robustness exercise does not turn on which cleaning convention is correct in any given case, and the literature settles that question, often differently, in each one. It argues, rather, that any comparison of forecasting methods that does not hold the cleaning procedure constant is at risk of confounding the intended comparison. The Options and Treasury panels add a second warning: Cleaning can flip not only the relative ranking of models but the sign of several architectures' R^2_{oos}, so even a coarse summary such as which model produces positive out-of-sample R^2 is cleaning dependent. Reproducible, version-controlled pipelines like FTSFR are what allow these comparisons to be made on fixed data, and they are what allow a reader to verify which side of each contrast they care about.

7 Conclusions and Future Work

This paper introduces the Financial Time Series Forecasting Repository (FTSFR), an open-source benchmark for financial time series forecasting. By holding the data constant across methods (through reproducible cleaning of every series), FTSFR isolates method skill from cleaning choices and turns the question about which forecaster wins into one with reproducible answers.

Our work makes three primary contributions. First, we deliver an open-source pipeline that downloads, cleans, and formats each dataset according to the canonical paper for that asset class, and we validate every replication against published figures or summary statistics. Because financial data are subject to licensing restrictions that prevent redistribution, we ship code rather than data; any researcher with the appropriate WRDS or Bloomberg subscriptions can reproduce every series exactly. This addresses the under-appreciated source of cross-study disagreement that motivated the

paper: Two researchers nominally using the same financial series often produce materially different time series, contaminating any subsequent comparison of forecasting methods.

Second, the benchmark covers seven major asset classes (with both aggregated portfolio and disaggregated security level data), five basis spread datasets following Siriwardane, Sunderam, and Wallen (2021), and bank and intermediary indicators including bank Call Reports (Drechsler, Savov, and Schnabl, 2017) and the intermediary capital factors of He, Kelly, and Manela (2017).

Third, we establish reproducible baselines using twelve forecasting models ranging from the Historical Average benchmark and classical statistical methods (ARIMA, Simple Exponential Smoothing, Theta) to hybrid linear architectures (DLinear, NLinear) to deep learning models (DeepAR, N-BEATS, N-HiTS, Vanilla Transformer, TiDE, KAN), all evaluated under one preprocessing pipeline and one rolling origin cross-validation protocol. These baselines turn the benchmark into something researchers can build against rather than alongside.

Consistent with the M-competitions' lesson that no single method dominates universally, the empirical results show that model choice must respect the data generating process, and that apparent disagreements across prior cross-study comparisons can be partly resolved once the data are held fixed. 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 across equities, corporate bonds, U.S. Treasuries, FX, options, and CDS. We caveat this finding since it pertains to the class of models the benchmark evaluates, namely, univariate global forecasts without exogenous regressors (see Section 2.4), and is consistent with decades of empirical asset-pricing evidence on how thin the exploitable signal is in that class. Extending the benchmark to methods that admit exogenous regressors and to other model classes (multivariate panels, ARIMAX, VARX, DeepAR with covariates, Temporal Fusion Transformers) is a natural follow-up and may sharpen what is forecastable in returns. The remaining two segments produce more nuanced rankings. On basis spreads, the deep and hybrid architectures `NLinear`, `NBEATS`, `TiDE`, and `NHITS` deliver materially more forecasting power than the historical-average baseline (median R^2_{oos} of about 0.63, ahead of the best classical contenders at 0.57–0.59), exploiting persistent seasonal funding-quarter structure that classical univariate methods do not encode. On bank and intermediary indicators, by contrast, the family rankings are close as hybrid linear methods hold a modest edge on both median and mean R^2_{oos}, with the classical baselines just behind and the heavier deep architectures trailing. The simple classical methods (ARIMA, Theta) therefore remain a defensible default for bank balance sheet projection, even though their typical performance is no better than the hybrid competitors. The asset pricing, risk management, and financial stability use cases for these forecasts therefore call for different default models, with the per task rankings in Section 5 guiding the choice.

Beyond the per segment rankings, the sensitivity exercise in Section 6.2 confirms the motivat-

ing thesis of the paper. Across four pairwise cleaning contrasts (TRACE-MMN signal cleaning, Constantinides-Jackwerth-Savov option-filter depth, NYSE-versus-CRSP breakpoints on Fama-French 25, and the Gürkaynak-Sack-Wright Treasury filter), the leaderboard reshuffles materially when only the cleaning convention changes. Classical baselines are stable across conventions, while deep models can swing by enough to flip the sign of R^2_{oos} or move from top to below the historical-average baseline. The empirical takeaway is that the model that works best is a function of the cleaning recipe, and that reproducible, version-controlled pipelines are a precondition for any rigorous answer. We hope FTSFR provides a common benchmark against which future methodological progress in financial time series forecasting can be measured.

Acknowledgments

We would like to thank the following individuals. With their permission, we have adapted and used pieces of code in this repository from 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, Philippe Mueller, and Cesare Robotti. 2023. "Priced risk in corporate bonds." *Journal of Financial Economics* 150 (2):103707.

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." .

Gu, Shihao, Bryan Kelly, and Dacheng Xiu. 2020. "Empirical Asset Pricing via Machine Learning." *The Review of Financial Studies* 33 (5):2223–2273.

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.

Hewamalage, Hansika, Klaus Ackermann, and Christoph Bergmeir. 2023. "Forecast evaluation for data scientists: common pitfalls and best practices." *Data Mining and Knowledge Discovery* 37 (2):788–832.

Hu, Yifan, Yuante Li, Peiyuan Liu, Yuxia Zhu, Naiqi Li, Tao Dai, Dawei Cheng, Changjun Jiang, and Shu-tao Xia. 2025. "FinTSB: A Comprehensive and Practical Benchmark for Financial Time Series Forecasting." *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. 2018a. "Statistical and Machine Learning forecasting methods: Concerns and ways forward." *PLoS ONE* 13 (3):e0194889.

———. 2018b. "The M4 Competition: Results, findings, conclusion and way forward." *International Journal of Forecasting* 34 (4):802–808.

———. 2020. "The M4 Competition: 100,000 time series and 61 forecasting methods." *International Journal of Forecasting* 36 (1):54–74.

———. 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 HKM, 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 10) and the average three month rolling standard deviation for a given date (Figure 11). 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 are more diverse, leading to lower average returns and lower volatility.

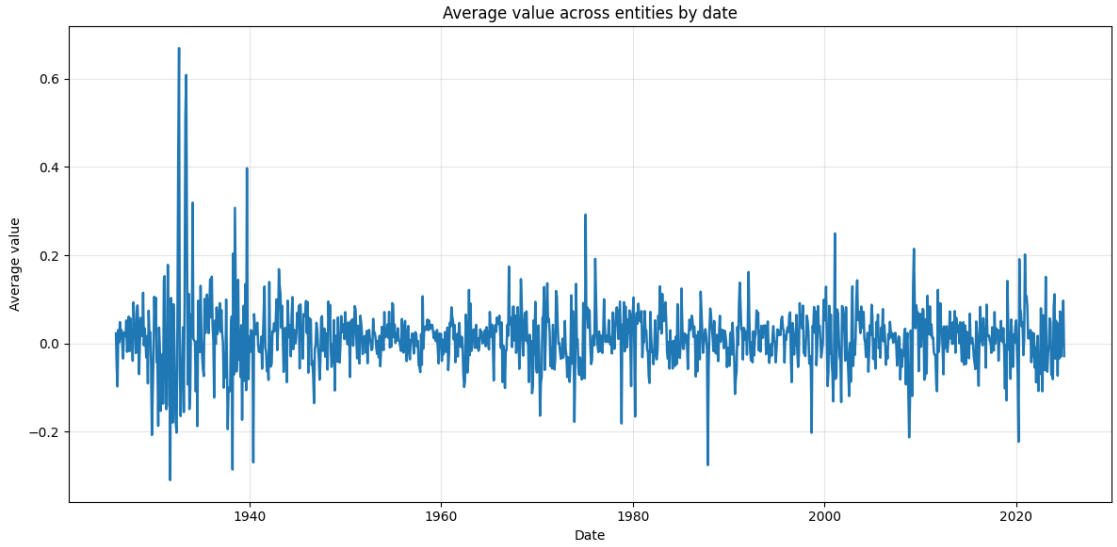


Figure 10: Average Returns of All Stocks

Sources: Center for Research in Security Prices, Authors' analysis

A.1.2 U.S. Treasuries

The U.S. 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.

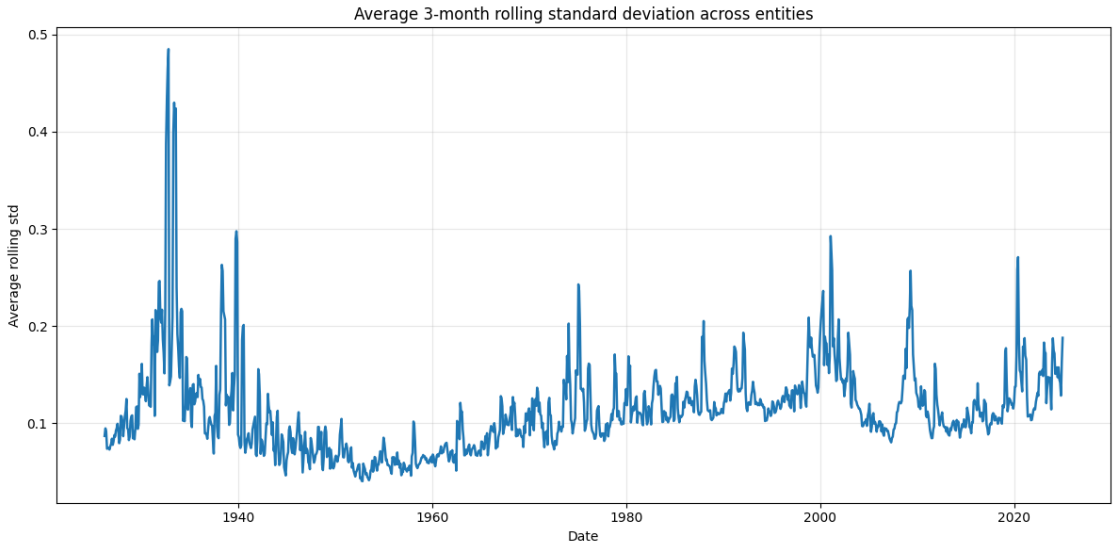


Figure 11: Average Three Month Rolling Standard Deviation

Sources: Center for Research in Security Prices, Authors' analysis

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.

Second, our maturity-sorted return portfolios focus on bonds with remaining maturities between roughly six months and five years, grouped into six-month buckets. This is a construction choice for the FTSFR portfolios rather than a restriction in the Gürkaynak-Sack-Wright curve itself, which is fit across maturities out to thirty years. The short-to-intermediate segment represents the most actively traded Treasury securities, where price discovery is most efficient and the longest maturities can suffer from illiquidity and fragmented trading.

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 12.

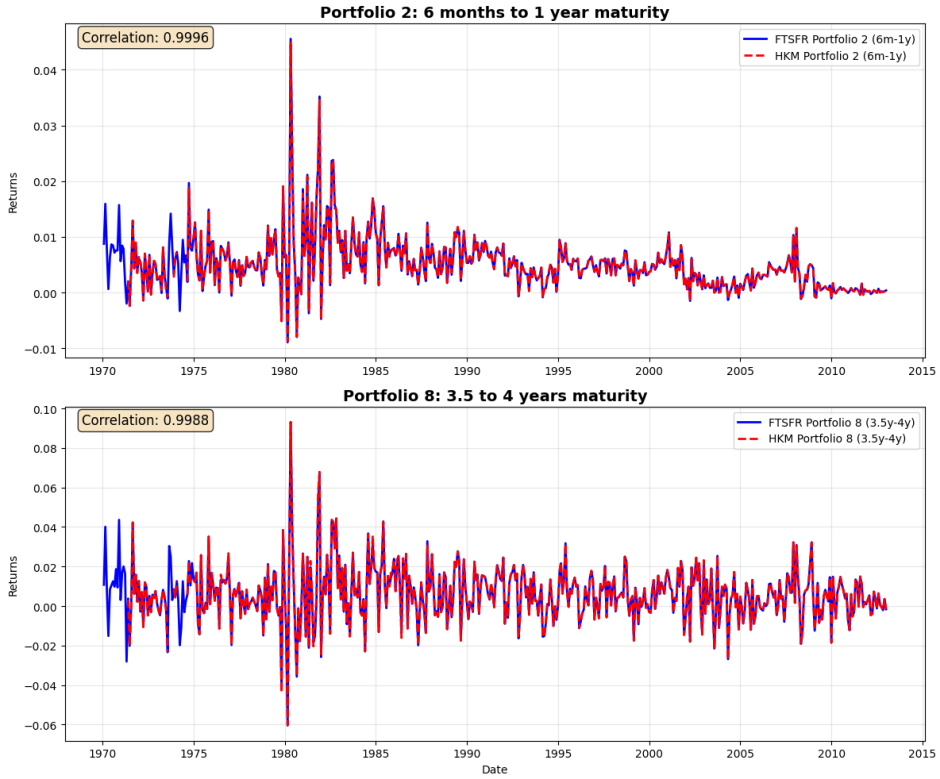


Figure 12: Comparison of He, Kelly, and Manela (2017) Treasury Bond Returns Dataset with FTSFR Dataset

Sources: Center for Research in Security Prices, He, Kelly, and Manela (2017), Authors' analysis

A.1.3 Corporate Bonds

The corporate bond returns data come from the OSBAP corporate bond panel, built from the TRACE (Trade Reporting and Compliance Engine) transaction tape, and follow the methodology outlined in 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, amounts 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 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 13, 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 (27 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 (24 years of data). As shown in Figure 14, 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

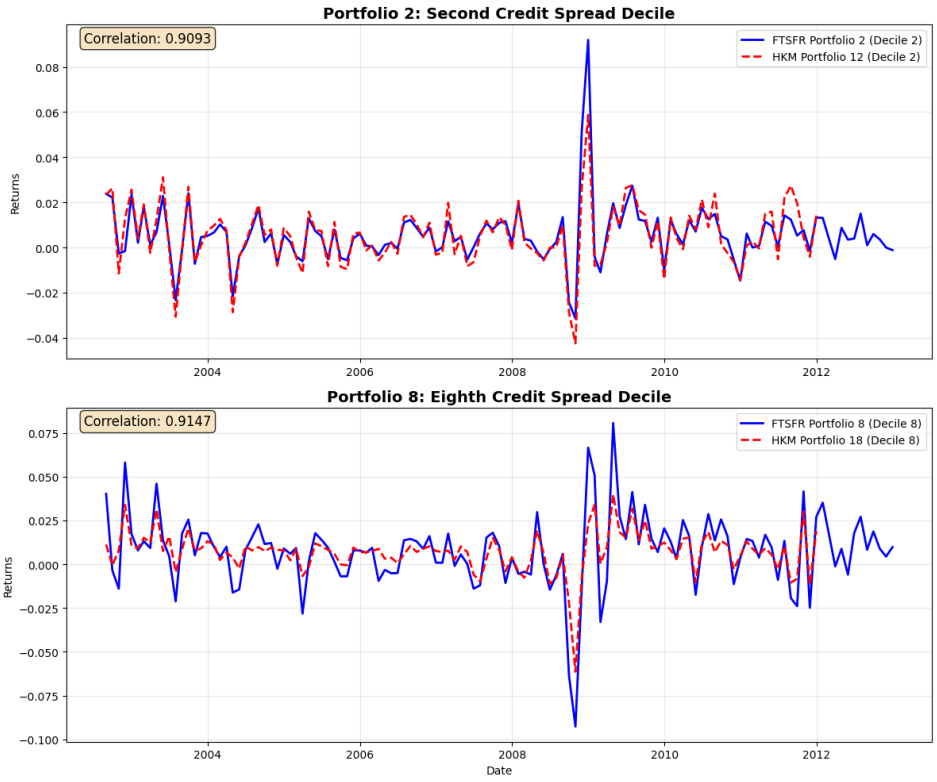


Figure 13: Comparison of He, Kelly, and Manela (2017) Corporate Bond Returns Dataset with FTSFR Dataset

Sources: Open Source Bond Asset Pricing, He, Kelly, and Manela (2017), Authors' analysis

outlined in Constantinides et al. (2013) and (2) Transforming these returns series into the 18 portfolios outlined in He, Kelly, and Manela (2017).

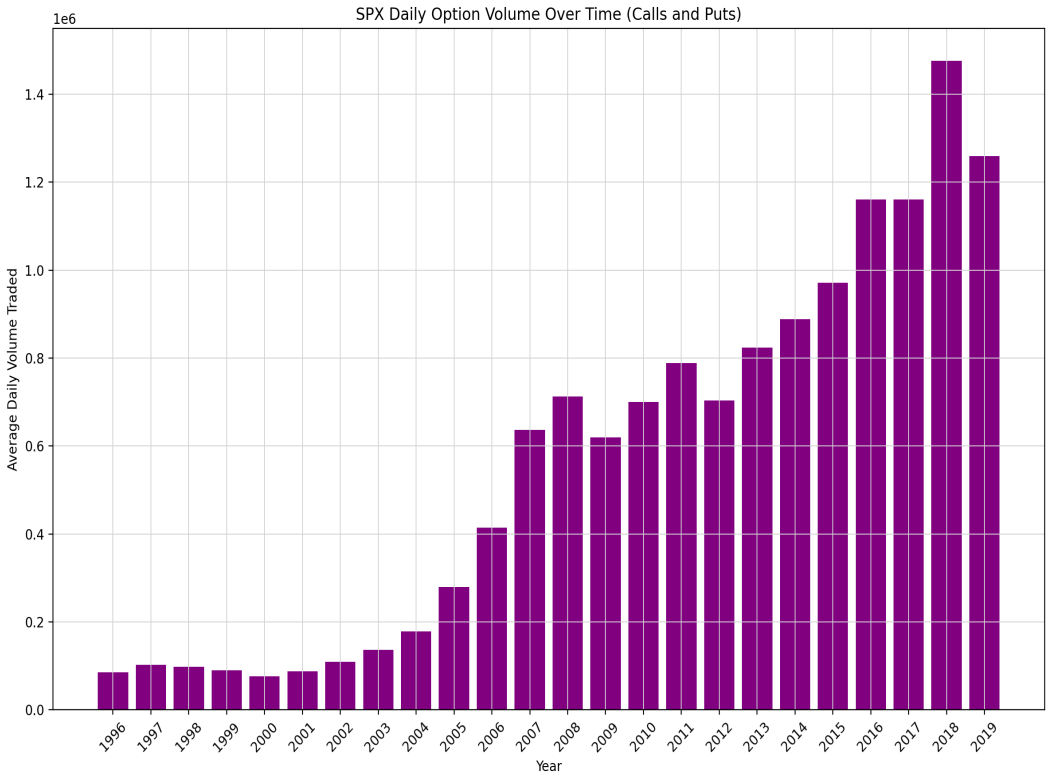
Options Data Series Access The final FTSFR data series comprise the monthly leverage-adjusted returns for call and put portfolios for both CJS 2013 (54 portfolios) and HKM 2017 (18 portfolios) for the period from Jan 1996–Dec 2019. Each portfolio is identified by a unique string of the form:

`{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.

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,

Figure 14: Growth in the SPX Options Market



Sources: OptionMetrics, Authors' analysis

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 USD-based foreign currency returns.

We are replicating a process where we convert our USD into the foreign currency i at end of day

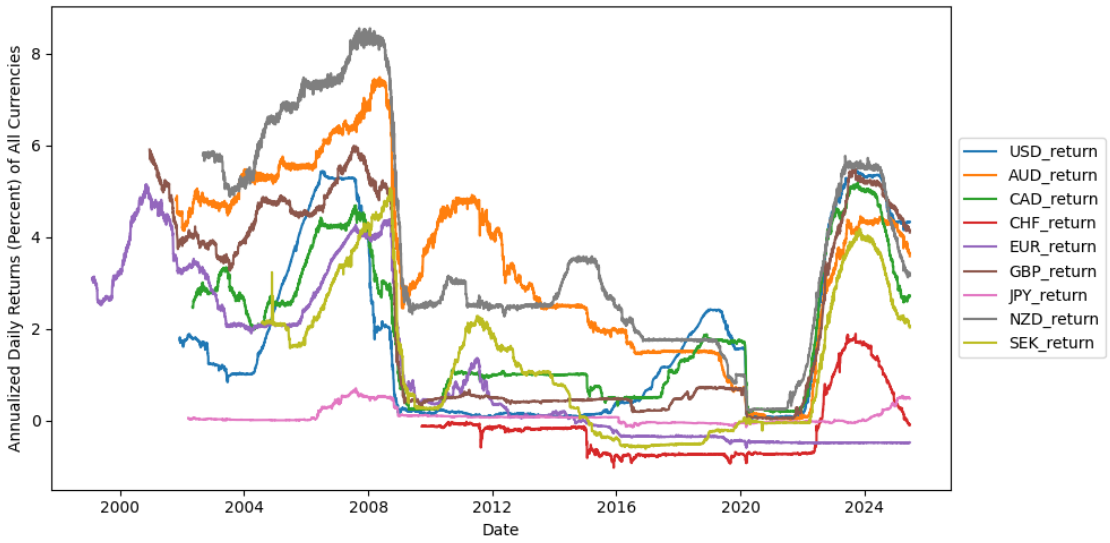


Figure 15: Foreign Currency Returns

Sources: Bloomberg, Authors' analysis

$t-1$, 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)

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 is straightforward. We retrieve the designated Bloomberg excess return indices, align all series to a monthly frequency by selecting the last observation in each calendar

month, compute simple percentage returns, harmonize naming conventions, and drop missing values to ensure that each commodity return series is complete and comparable across the sample period. Finally, we compare these series with those published by HKM. HKM 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 HKM. 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 16 reports the pairwise correlations between our Bloomberg GSCI-based returns and the corresponding series in HKM. 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 underlying 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 HKM. Furthermore, HKM uses a different data source 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 17 highlights this heterogeneity visually, with clusters of high-correlation matches alongside a handful of clear mismatches.

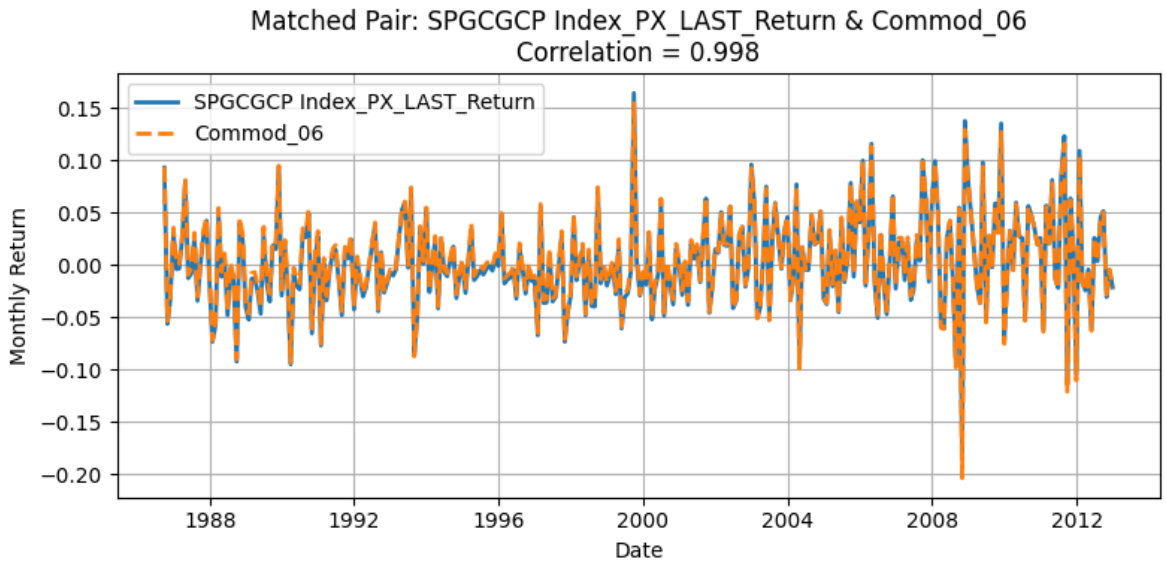


Figure 16: Pairwise Correlations between GSCI Commodity Returns and Yang (2013) Commodity Returns

Sources: Bloomberg, He, Kelly, and Manela (2017), Authors' analysis

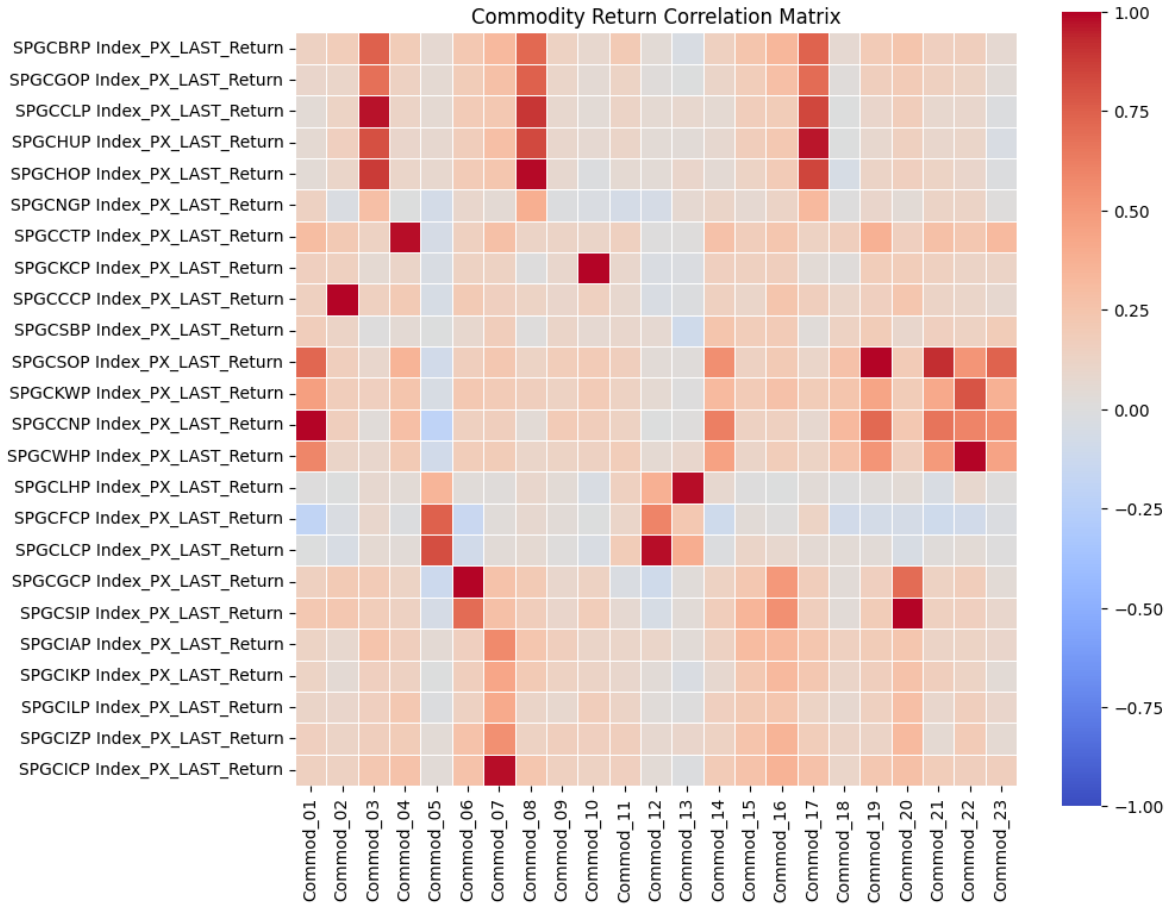


Figure 17: Heatmap of Pairwise Correlations between GSCI Commodity Returns and Yang (2013) Commodity Returns

Sources: Bloomberg, He, Kelly, and Manela (2017), Authors' analysis

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 rollover 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 to 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.

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 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 floating-rate spread is conventionally proxied by the bond's Z-spread, and the CDS par spread is interpolated to the bond maturity by cubic spline; we refer the reader to Siriwardane, Sunderam, and Wallen (2021) for the full construction.

Our implementation draws bond observations from FINRA TRACE, following the filters used in the monthly cleaning procedure of Dickerson, Robotti, and Rossetti (2024), and CDS quotes from S&P Global via WRDS, restricting to USD-denominated senior unsecured contracts on U.S. firms at the 1-, 3-, 5-, 7-, and 10-year tenors and joining the two sides on ISIN through the RED-ISIN obligation lookup table. The bond filters follow the source paper: senior unsecured USD debt with a valid ISIN, fixed coupon, one- to ten-year remaining maturity, outstanding principal of at least $100,000, no convertible structures, and end-of-month price at least fifty cents on the dollar. The floating-rate spread is constructed as a true bond-level Z-spread, and for each (date, bond), we fit a Nelson–Siegel–Svensson Treasury curve on the CRSP daily Treasury panel following Gürkaynak, Sack, and Wright (2007), fall back to the Federal Reserve's published NSS parameters on dates when the CRSP fit is pathological, and then solve numerically for the constant continuous-compounded spread that reprices the bond's coupon and principal cash flows off the fitted curve. The CDS par spread is interpolated by cubic spline across the 1Y, 3Y, 5Y, 7Y, and 10Y tenors and matched to the bond's days-to-maturity; we require at least two distinct CDS tenors per firm-date and discard firm-dates that fail this requirement rather than extrapolating. Finally, we drop Z-spread fits that hit the numerical root-finder's bracket boundary, drop bond-level bases beyond ±1000 bps, and require at least five bonds per rating bucket and date when averaging into the IG and HY indices.

Frequency difference versus Siriwardane, Sunderam, and Wallen (2021). The most consequential implementation difference from the source paper is the output frequency. The cleaned bond returns panel is a monthly end-of-month panel, with one observation per bond per month indexed on the end-of-month price date, so our basis series is constructed at month-end and plotted in Figure 18 as a monthly time series. Siriwardane, Sunderam, and Wallen (2021), in contrast, plot a daily series by reading bond prices directly from the S&P Global bond pricing tables and using the WRDS Bond Return dataset only as a tradability filter on ISIN (in a given month they only include bonds whose ISINs are included in the WRDS Bond Return dataset). Reproducing the daily frequency would require adding a daily S&P Global bond pricing source alongside the Dickerson, Robotti, and Rossetti (2024) filter; the rest of the construction is unchanged.

Outlier handling. Three pieces of outlier handling are worth flagging explicitly because they affect the visible level of the IG and HY series. (i) On a handful of quote dates the CRSP NSS curve fit converges to a pathological local optimum (we observed, e.g., a 25% fitted 1Y spot rate on 2010-

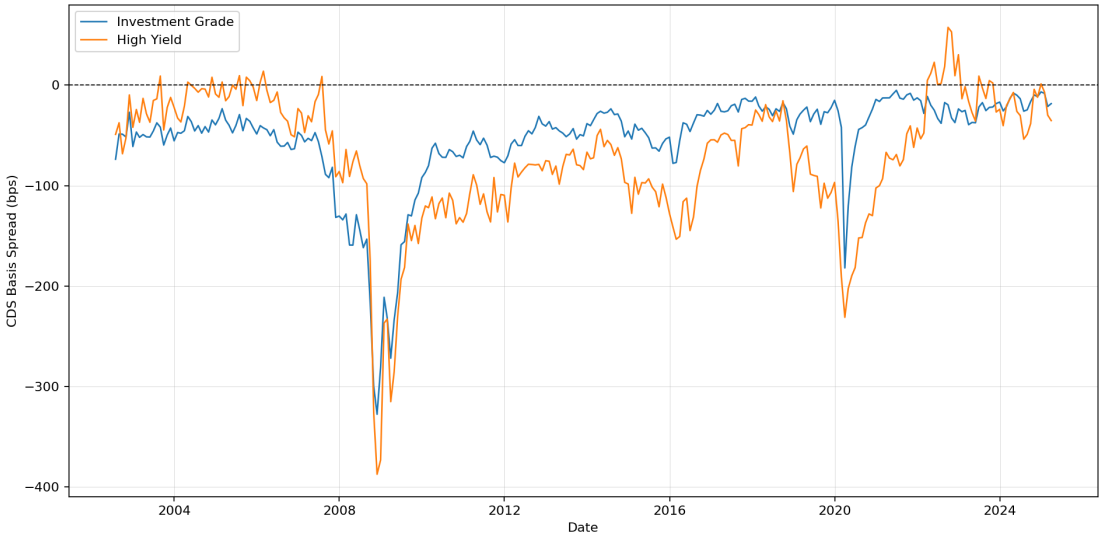


Figure 18: CDS Arbitrage Spreads, High Yield and Investment Grade

Sources: Center for Research in Security Prices, FINRA TRACE, S&P Global, Authors' analysis

06-30 and 2011-10-31). We detect this by comparing the fitted 1Y, 5Y, and 10Y spot rates to the Federal Reserve's published SVENY series for the same date and, when the absolute discrepancy exceeds 50 bps at any of those points, substitute the Fed's published NSS parameters for that date. Without this fallback, the Z-spreads on those days are essentially noise and produce month-end spikes of 800–1000 bps in the rating-bucket means. (ii) We discard individual Z-spread fits that hit the numerical root-finder's ±20% bracket, which are failed fits rather than realized spreads. (iii) We discard bond-level bases whose magnitude exceeds 1,000 bps; these are typically driven by cubic-spline extrapolation of the CDS par spread to bonds whose maturity sits at the edge of the [1Y, 10Y] CDS tenor grid. After these filters, we still require at least five surviving bonds in a given (rating, date) cell before reporting a rating-bucket average; this last screen suppresses the single-bond spikes that otherwise dominate the early high-yield sub-sample where the CDS-covered HY universe is sparse.

The CDS-Bond basis is plotted in Figure 18.

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. Du, Tepper, and Verdelhan (2018) link the large crisis-period deviations to a severe dollar funding shortage in the presence of limits to arbitrage, and attribute the persistence of post-crisis deviations to the balance sheet constraints that banking regulation places on intermediaries.

Before the GFC, deviations from CIP were typically small in magnitude and short-lived, consis-

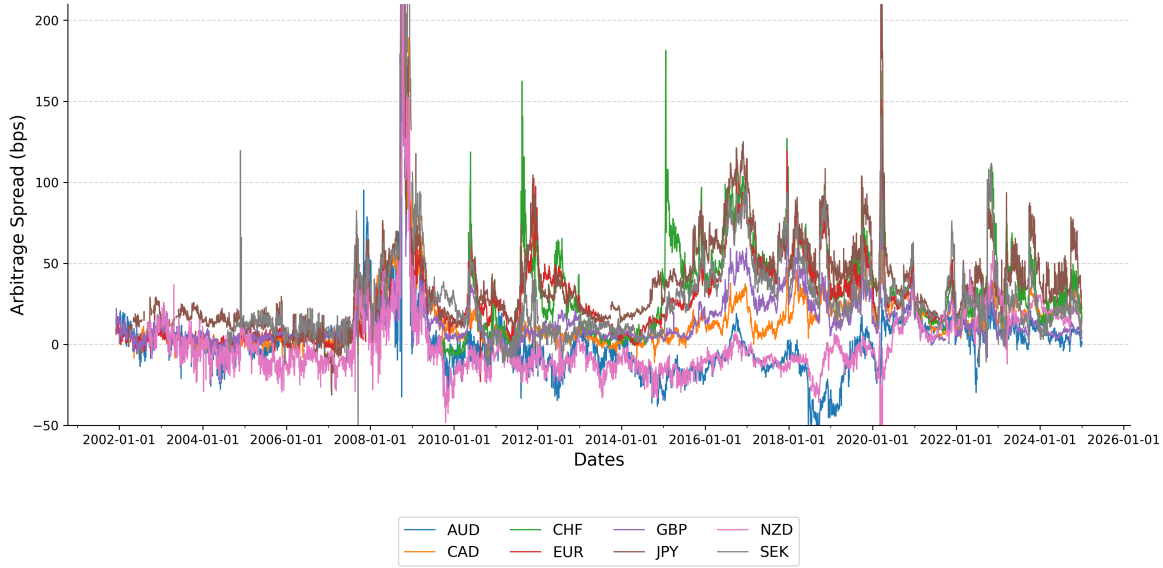


Figure 19: Comparison of CIP Arbitrage Spreads

Sources: Bloomberg, Authors' analysis

tent with arbitrage activity keeping the parity tight (see the literature reviewed in Du, Tepper, and Verdelhan, 2018).

Our analysis examines CIP deviations across eight of the ten G10 currencies (AUD, CAD, CHF, EUR, GBP, JPY, NZD, and SEK; we omit NOK and DKK) against the USD, using daily Bloomberg spot rates, three-month forward points, and three-month overnight index swap (OIS) rates from January 2010 onward; OIS serves as the risk-free benchmark, matching the convention of Siriwardane, Sunderam, and Wallen (2021) so that CIP deviations are directly comparable to the other basis spreads in our benchmark. Two construction details are easy to get wrong and worth naming. Forward points are scaled per 10,000 for every currency except JPY, which by market convention is scaled per 100, and the four currencies conventionally quoted as USD per foreign currency (EUR, GBP, AUD, NZD) must be inverted to a foreign per USD convention before any CIP residual is computed. Failure to do either produces sign errors that are large enough to swamp the actual deviations. We then remove outliers with a rolling 45-day median absolute deviation filter at a threshold of ten mean absolute deviations (set empirically), flagging values to missing rather than dropping them so the calendar of business days is preserved across currencies.

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 19, 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; Fleckenstein, Longstaff, and Lustig (2014) attribute the persistence of deviations from this equivalence to slow-moving capital and partial segmentation in the ownership of TIPS and Treasury bonds. 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 synthetic nominal yield in basis points as $10{,}000 \times (\exp(r + \log(1 + \pi)) - 1)$, where r is the continuously compounded TIPS 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. Availability of inflation swap data 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 with substantial positive values during 2008-2009 (often exceeding 100 basis points), narrowing spreads 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 are consistent with the finding that Treasury bonds are consistently expensive relative to TIPS during stress periods, in line 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 mean absolute deviation. 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 the 2008–2009 and March 2020 stress episodes, when the futures-implied rate exceeded OIS by substantial margins, a pattern that Fleckenstein and Longstaff (2020) attribute to the balance sheet costs facing financial intermediaries.

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, i.e., 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, and 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. The literature attributes this persistence to balance sheet costs, supplementary leverage ratio requirements, and margin constraints that limit dealer arbitrage capacity (Jermann, 2020; Du, Hébert, and Li, 2023; Hanson, Malkhozov, and Venter, 2023). 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.

B Forecasting Preprocessing Pipeline

Table 15 reports the impact of the unified preprocessing pipeline described in Section 4 on each dataset in the benchmark. For every dataset we list the original and retained entity counts, the implied retention rate, median series lengths before and after filtering, and the adaptive minimum observation threshold imposed by the length screen. The filtering rules are uniform across statistical and neural estimators so that every model sees the same vetted panel.

Table 15: 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	2776	2608	93.9%	66	72	2002-07-31 – 2025-04-30
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%	269	269	2002-08-31 – 2024-12-31
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	53389	28581	53.5%	22	53	2002-08-31 – 2025-03-31
FX	Monthly	9	8	88.9%	5991	275	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	2	2	100.0%	516	516	1970-01-01 – 2012-12-01
HKM Daily Factor	Daily	2	2	100.0%	4765	6917	2000-01-03 – 2018-12-11
HKM Monthly Factor	Monthly	2	2	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, Compustat, Call Reports, FINRA TRACE, OptionMetrics, S&P Global, Open Source Bond Asset Pricing, He, Kelly, and Manela (2017), Authors' analysis