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

Jeremiah Bejarano[1] Viren Desai[2] Kausthub Keshava[2] Arsh Kumar[2] Zixiao Wang[2] Vincent Hanyang Xu[2] Yangge Xu[2]

September 22, 2025

[1]Office of Financial Research, U.S. Department of the Treasury; Financial Mathematics Program, University of Chicago

[2]Independent

Disclaimer

Any opinions expressed herein are my own and do not necessarily reflect those of the Office of Financial Research or the U.S. Department of the Treasury.

All mistakes are my own.

Overview

- Question: Can modern global forecasting methods improve financial stability monitoring beyond traditional approaches?
- Approach: Systematic benchmarking of over a dozen forecasting methods (classical to deep learning) on 25 canonical financial datasets with standardized evaluation.
- Main Result: Returns remain extremely hard to beat—the historical average dominates—yet global ML models deliver sizable MASE gains on basis spreads (CIP, Treasury swaps) and supervisory indicators (bank liquidity/leverage).
- Application: Evidence-based guidance for financial stability authorities on when to lean on classical baselines versus global ML methods across market segments.

Motivation

Financial Stability Requires Forward-Looking Monitoring

- Financial regulators have emphasized forward-looking risk monitoring to address systemic vulnerabilities
- Early warning systems and predictive analytics are critical for timely policy responses
- Effective forecasting enhances the toolkit for macroprudential surveillance and systemic risk monitoring
- Better forecasting → Spot trouble earlier → Safeguard the economy

The Benchmarking Gap in Financial Forecasting

- Time series forecasting is ubiquitous in finance, but lacks standardized evaluation frameworks
- Current problem: Time series forecasting researchers evaluate methods on arbitrarily selected datasets, making comparisons of the methods impossible
- No Free Lunch Theorem: Some benchmarks already exist, but need domain-specific evaluation to identify what works in finance (Wolpert and Macready, 1997)

Existing Benchmarks Have Limited Financial Coverage

Source	Name	Coverage of Finance Domain
McCracken (2016)	FRED-MD	US Macroeconomic Series
Hu et al. (2018)	FinTSB	Equities Returns only
Dau et al. (2019)	UCR Time Series Classification Archive	None
Bagnall et al. (2018)	UEA Multivariate Time Series Classification Archive	None
Godahewa et al. (2021)	Monash Time Series Forecasting Repository	Fred-MD is one component
Qiu et al. (2024)	TFB	NN5 Bank Cash Withdrawals, Equity (NYSE/NASDAQ), Foreign Exchange Rates

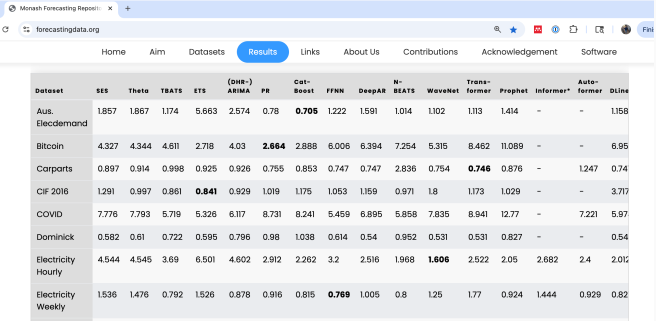
Gap: No comprehensive benchmark focused specifically on financial markets with canonical data cleaning procedures

Why This Matters: Small Gains, Big Impact

- Even 3-10% forecasting improvements can be economically significant
- In finance, small edges compound over time and across large portfolios
- Standardized benchmarks drive progress by enabling:
    - Apples-to-apples comparisons across methods
    - Identification of what works for specific market segments
    - Prevention of cherry-picking results
- Solution: Create comprehensive, literature-compliant financial forecasting benchmark repository

Monash Time Series Forecasting Repository

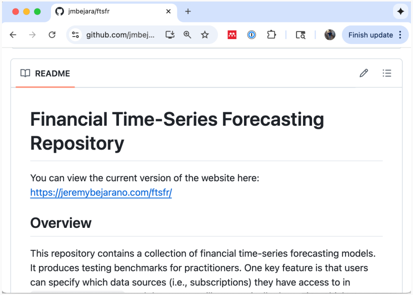
- I take inspiration from the Monash Time Series Forecasting Repository
- They provide a suite of freely distributable datasets that time series forecasting researchers can use to test global time series forecasting methods on.
- `https://forecastingdata.org/`



Dataset	SES	Theta	TBATS	ETS	(DHR-) ARIMA	PR	Cat-Boost	FFNN	DeepAR	N-BEATS	WaveNet	Trans-former	Prophet	Informer*	Auto-former	DLine
Aus. Elecdemand	1.857	1.867	1.174	5.663	2.574	0.78	**0.705**	1.222	1.591	1.014	1.102	1.113	1.414	-	-	1.158
Bitcoin	4.327	4.344	4.611	2.718	4.03	**2.664**	2.888	6.006	6.394	7.254	5.315	8.462	11.089	-	-	6.95
Carparts	0.897	0.914	0.998	0.925	0.926	0.755	0.853	0.747	0.747	2.836	0.754	**0.746**	0.876	-	1.247	0.74
CIF 2016	1.291	0.997	0.861	**0.841**	0.929	1.019	1.175	1.053	1.159	0.971	1.8	1.173	1.029	-	-	3.717
COVID	7.776	7.793	5.719	5.326	6.117	8.731	8.241	5.459	6.895	5.858	7.835	8.941	12.77	-	7.221	5.97
Dominick	0.582	0.61	0.722	0.595	0.796	0.98	1.038	0.614	0.54	0.952	0.531	0.531	0.827	-	-	0.54
Electricity Hourly	4.544	4.545	3.69	6.501	4.602	2.912	2.262	3.2	2.516	1.968	**1.606**	2.522	2.05	2.682	2.4	2.012
Electricity Weekly	1.536	1.476	0.792	1.526	0.878	0.916	0.815	**0.769**	1.005	0.8	1.25	1.77	0.924	1.444	0.929	0.82

The Importance Of Open Source And Replicability

- Since data cannot be redistributed, we provide open source code that enables researchers to download and use the exact datasets and methods used in the paper
- Open source code packaged with replicable software environments enables
  - researchers to download and use the exact datasets and methods used in the paper
  - Community validation and improvement of the code



Financial Time-Series Forecasting Repository

You can view the current version of the website here: https://jeremybejarano.com/ftsfr/

Overview

This repository contains a collection of financial time-series forecasting models. It produces testing benchmarks for practitioners. One key feature is that users can specify which data sources (i.e., subscriptions) they have access to in

Related Literature

Related Research

- Financial Stability & Many-Predictor Methods: Principal components and factor models improve forecasts when exploiting many predictors.

  Stock & Watson (2002); OFR's Financial Stress Index; Fed's SAFE system

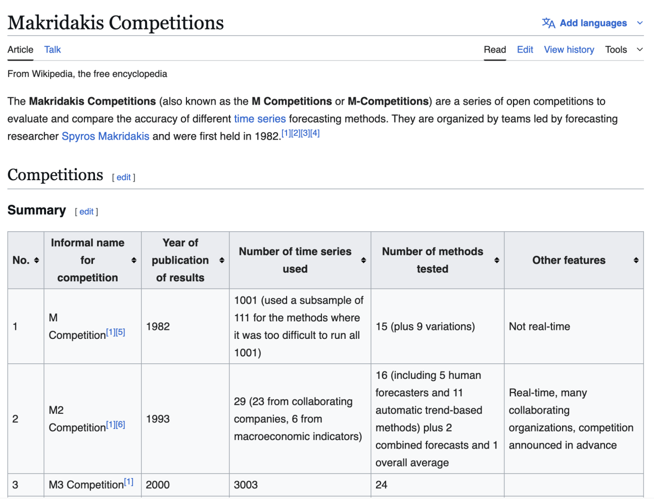
- Return Predictability: Modern asset pricing shows discount-rate variation drives return predictability across asset classes. Cross-sectional information enhances aggregate forecasts.

  Cochrane (2011); Kelly & Pruitt (2013); Kelly & Xiu (2023)

- Global Time Series Forecasting: Single models trained across many series consistently win forecasting competitions. Particularly effective for related financial series sharing economic drivers.

  M-competitions; NN3/NN5 competitions; Godahewa et al. (2021)

M-Competitions



Makridakis Competitions

From Wikipedia, the free encyclopedia

The **Makridakis Competitions** (also known as the **M Competitions** or **M-Competitions**) are a series of open competitions to evaluate and compare the accuracy of different time series forecasting methods. They are organized by teams led by forecasting researcher Spyros Makridakis and were first held in 1982.[1][2][3][4]

Competitions [edit]

Summary [edit]

No.	Informal name for competition	Year of publication of results	Number of time series used	Number of methods tested	Other features
1	M Competition[1][5]	1982	1001 (used a subsample of 111 for the methods where it was too difficult to run all 1001)	15 (plus 9 variations)	Not real-time
2	M2 Competition[1][6]	1993	29 (23 from collaborating companies, 6 from macroeconomic indicators)	16 (including 5 human forecasters and 11 automatic trend-based methods) plus 2 combined forecasts and 1 overall average	Real-time, many collaborating organizations, competition announced in advance
3	M3 Competition[1]	2000	3003	24	

Data & Contributions

Comprehensive Financial Dataset Coverage

- 25 datasets across multiple asset classes and market segments:
  - **Asset Returns:** Equities (CRSP), Corporate & Treasury bonds, CDS, Options, FX, Commodities
  - **Arbitrage Spreads:** CIP deviations, CDS-bond basis, Treasury-swap spreads, TIPS-Treasury, Treasury-Spot-Futures
  - **Other:** Bank call reports, intermediary risk factors, yield curves
- Each dataset follows canonical cleaning procedures from seminal finance papers
- Validated by replicating key results from original studies
- Scale: From 4 entity panel to panel with over 25,000 individual stocks
- Frequencies vary from daily, monthly, and quarterly, with some series starting as early as 1926

Dataset Name	Description	Citation
Returns Data		
CDS Contract	Monthly returns for individual CDS contracts	Palhares (2012)
CDS Portfolio	Aggregated into 20 CDS portfolios by tenor and credit quality	He et al. (2017)
Commodity	Monthly returns for commodity futures	Yang (2013)
Corporate Bond	Monthly returns for individual corporate bonds from TRACE	Dickerson et al. (2024)
Corporate Portfolio	Corporate bond portfolios by credit spread	Nozawa (2017)
CRSP Stock	Monthly stock returns from CRSP database	Fama & French (1993)
FF25 Size-BM	Daily Fama-French 25 portfolios: size and book-to-market	ibid.
FX	Daily foreign exchange returns vs USD	Lettau et al. (2014)
SPX Options	Monthly returns for individual SPX option contracts	Constantinides et al. (2013)
Treasury Bond	Monthly returns for individual Treasury bonds from CRSP	Gürkaynak et al. (2007)
Treasury Portfolio	Monthly returns for Treasury bond portfolios by maturity	ibid.

Dataset Name	Description	Citation
	Basis Spread Data	
CDS-Bond	Monthly CDS-bond basis spreads	Siriwardane et al. (2021)
CIP	Monthly covered interest parity deviations	Du et al. (2018)
TIPS-Treasury	Monthly TIPS-Treasury basis spreads	Fleckenstein et al. (2014)
Treasury-SF	Monthly Treasury-SF arbitrage spreads	Fleckenstein et al. (2020)
Treasury-Swap	Monthly Treasury-Swap arbitrage spreads	Siriwardane et al. (2021)
	Other Financial Data	
Bank Cash Liquidity	Quarterly cash liquidity from call report data	Drechsler et al. (2017)
Bank Leverage	Quarterly leverage ratios from call report data	ibid.
BHC Cash Liquidity	Quarterly bank holding company cash liquidity	ibid.
BHC Leverage	Quarterly bank holding company leverage ratios	ibid.
HKM Daily Factor	Intermediary risk factors, including capital ratio, capital risk factor	He et al. (2017)
Treasury Yield Curve	Daily Nelson-Siegel-Svensson zero-coupon yields, 1-30 years	Gürkaynak et al. (2007)

Dataset Selection: Financial Stability & Intermediary Asset Pricing

- Theoretical foundation: He et al. (2017) intermediary asset pricing framework
- Key insight: Retail investors access only stocks; intermediaries access exotic assets (CDS, options, bonds, commodities)
- Intermediary budget constraints create risk factors that drive pricing across asset classes
  - Risk factor impact stronger for assets only intermediaries can trade
  - These assets crucial for financial stability monitoring
- Basis spreads gauge stress in financial system
  - Critical early warning indicators for intermediary distress
  - Predicting basis spread changes essential for financial stability
- Return predictability fundamental to understanding economic risk across all asset classes

Dataset Selection: Methodological Rigor

- Canonical data cleaning: He et al. (2017) identified seminal papers in each asset class
  - Obtained exact data methodologies from original authors
  - No reinvention of cleaning procedures
  - Established best practices for each asset class
- Similar approach for basis spreads: Siriwardane et al. (2021) references canonical methods
  - CDS-bond basis, Treasury-swap spreads, CIP deviations, etc.
  - Each follows established academic procedures
- Additional financial stability data:
  - HKM factors: leverage ratios, capital constraints
  - Bank regulatory data: liquidity, leverage metrics
  - Yield curves: term structure risk factors
- Result: Comprehensive coverage of assets critical for intermediaries and financial stability

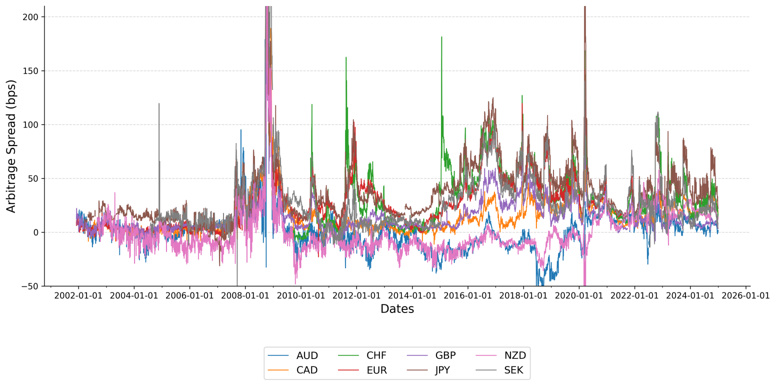
	Frequency	Unique Entities	Min Length	Median Length	Max Length	Min Date	Max Date
Basis Spreads							
CDS-Bond	Monthly	3402	1	16	169	2002-09-30	2022-09-30
CIP	Monthly	8	3997	5732	6030	2001-12-04	2025-02-28
TIPS-Treasury	Monthly	4	5126	5162	5197	2004-07-21	2025-05-30
Treasury-SF	Monthly	5	3783	5185	5192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

Key Innovation: Literature-Compliant Data Curation

- Fully reproducible data pipeline: automated download, cleaning, formatting
- Available on GitHub: `github.com/jmbejara/ftsfr`
- Handles data licensing constraints:
  - Provides scripts rather than redistributing proprietary data
  - Works with institutional subscriptions (WRDS, Bloomberg)
  - Enables exact dataset reproduction across researchers
- Both aggregated portfolios (for replication) and disaggregated data (for global forecasting)
- Virtual environments ensure perfect reproducibility
- Each dataset uses exact methodology from canonical finance papers:
- Code to replicate these datasets was not previously publicly available. This paper makes all of this code freely available on GitHub. The repository contains over 170,000 lines of code at last count.

Replications: CIP Arbitrage

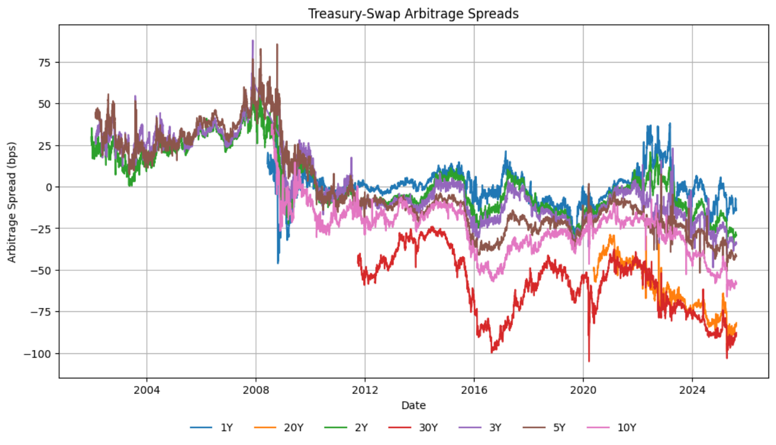
Figure 1CIP Arbitrage spreads



Sources: Bloomberg, Authors' creation

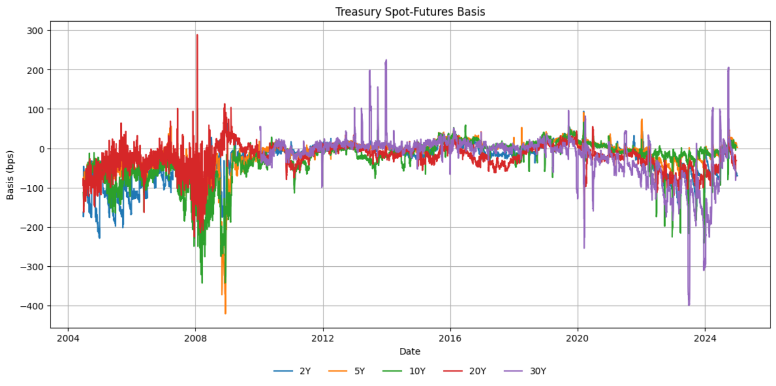
Replications: Treasury Swap Arbitrage

Figure 2 Treasury Swap Arbitrage spreads



Sources: Bloomberg, Authors' creation

Replications: Treasury Spot-Futures Basis



Successfully replicates Treasury spot-futures basis patterns from literature

Methodology

Forecasting Methods: 24 Models Across Three Categories

- Classical Statistical Methods:
  - Naive benchmarks plus Theta, SES (simple exponential smoothing family), and ARIMA
  - Computationally efficient, interpretable, strong baselines
- Machine Learning:
  - Linear models, gradient boosting (e.g., CatBoost)
  - Balance between complexity and interpretability
- Deep Learning (Global Methods):
  - NBEATS, NHITS, VanillaTransformer, TiDE, KAN, DLinear, NLinear
  - Train single model across all time series in dataset with automated hyperparameter search
  - Learn cross-sectional patterns while guarding against overfitting via shared folds

Model Deep Dive: Classical & Machine Learning

- Classical Statistical Methods
  - **Auto wrappers:** Theta, ARIMA, and SES reuse the rolling CV folds to tune parameters and avoid overfitting
  - **Theta Models:** Decompose series by modifying local curvature into short and long-term components, then extrapolate separately
  - Simple but effective, often outperforming sophisticated methods on short, unrelated, or noisy series
  - **ARIMA Models:** Auto-Regressive Integrated Moving Average models for stationary time series; mixed performance but generally outperform simple baselines
  - **Simple Exponential Smoothing (SES):** Straightforward baseline with exponential weighting of past observations
- Machine Learning Models
  - **Linear Models:** Effective baseline but limited for non-linear dependencies between sequences and covariates
  - **Gradient Boosting (CatBoost):** Ensemble of decision trees designed to prevent prediction shift and target leakage
  - Global ML models particularly effective for intermittent datasets where traditional univariate models struggle

Model Deep Dive: Deep Learning (Global Methods)

- **Global Auto Models** (NBEATS, NHITS, TiDE, KAN, VanillaTransformer, DLinear, NLinear) train a single model across all time series and tune hyperparameters with Bayesian search on shared validation folds
- Transformer-based Models
  - **Transformer:** Self-attention mechanisms without recurrence; effectiveness for long-term forecasting debated
  - Studies show simple linear models can outperform Transformers on standard benchmarks
- MLP-based Models
  - **DLinear & NLinear:** Simple one-layer models; DLinear decomposes trend/seasonal, NLinear normalizes inputs
  - **TiDE:** MLP encoder-decoder, 5-10x faster than Transformers, handles covariates and non-linear dependencies
  - **N-BEATS:** Deep architecture with residual links; can be interpretable via trend/seasonality decomposition
  - **N-HiTS:** Extension of N-BEATS for long horizons using hierarchical interpolation and multi-rate sampling

Evaluation Framework

- Consistent data preprocessing: All models receive identical filtered datasets
  - Removes series too short for reliable forecasting
- Fair comparison methodology:
  - Adaptive filtering based on median entity length
  - Seasonality-aware forecast horizons
  - Protection for small datasets ($\leq$10 entities)
- Two complementary error metrics:
  - MASE: Standard in time series forecasting (robust to outliers)
  - Out-of-sample R^2: Standard in finance (emphasizes large errors)
- Focus on baseline performance with minimal hyperparameter tuning

Error Metrics: Mathematical Definitions

- Mean Absolute Scaled Error (MASE):

$$\mathrm{MASE} = \frac{\frac{1}{T}\sum_{t=1}^{T} |y_t - \widehat{y}_t|}{\dfrac{1}{N-s}\displaystyle\sum_{t=s+1}^{N} |y_t - y_{t-s}|}$$

  - **Why chosen:** Standard accuracy measure in time series forecasting (Hyndman & Koehler 2006)
  - Scale-free comparison across different financial series
  - Robust to outliers (uses absolute errors)
  - Values < 1 indicate improvement over naive benchmark

- Out-of-Sample R^2:

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

  - **Why chosen:** Standard for evaluating predictive models in finance (Campbell & Thompson 2008)
  - Measures percentage reduction in MSE vs historical average
  - Emphasizes large forecast errors (quadratic loss)
  - Positive values indicate outperformance of simple benchmark

Results

Forecasting Design and Metrics

- **Horizon:** One month ahead for monthly data; 21 trading days (business) or 30 calendar days (daily); one quarter for quarterly series
- **Rolling windows:** up to 6 expanding cross-validation windows, limited by the shortest surviving series and reused by the Auto models to keep hyperparameters honest
- **Auto model selection:** Nixtla's Auto wrappers (NBEATS, NHITS, NLinear, Theta, ARIMA, DeepAR) run Bayesian searches on these folds, helping avoid overfitting while keeping the workflow reproducible
- **MASE:** Performance relative to a seasonal naïve baseline
  - Values $< 1.0 =$ better than naïve
  - Values $> 1.0 =$ worse than naïve
- **Out-of-sample R^2:** Percentage reduction in MSE vs historical mean
  - Positive values = model outperforms benchmark
  - Negative values = model underperforms benchmark
- Following slides report detailed tables and heatmaps for both metrics

MASE Results by Dataset and Model

Lower values = better performance; Values < 1 beat seasonal naive baseline

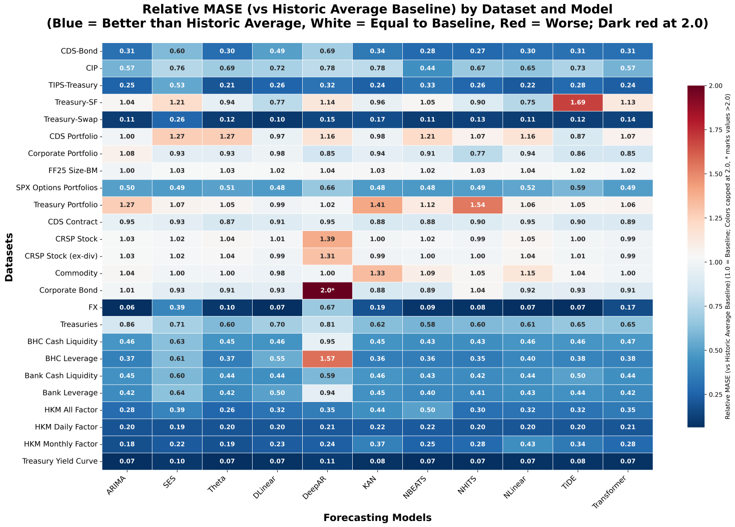
	HistAvg	ARIMA	SES	Theta	DLinear	DeepAR	KAN	NBEATS	NHITS	NLinear	TiDE	Transformer
Basis Spreads												
CDS-Bond	4.67	1.43	2.81	1.39	2.31	3.23	1.57	1.29	**1.27**	1.40	1.45	1.42
CIP	0.60	0.34	0.45	0.41	0.43	0.47	0.46	**0.26**	0.40	0.39	0.44	0.34
TIPS-Treasury	1.78	0.45	0.94	**0.37**	0.45	0.57	0.43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

Relative MASE Results by Dataset and Model

Performance relative to Historic Average: $< 1 =$ *better*, $> 1 =$ *worse*

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

Relative MASE Heatmap: Performance vs Historic Average



Relative MASE (vs Historic Average Baseline) by Dataset and Model
(Blue = Better than Historic Average, White = Equal to Baseline, Red = Worse; Dark red at 2.0)

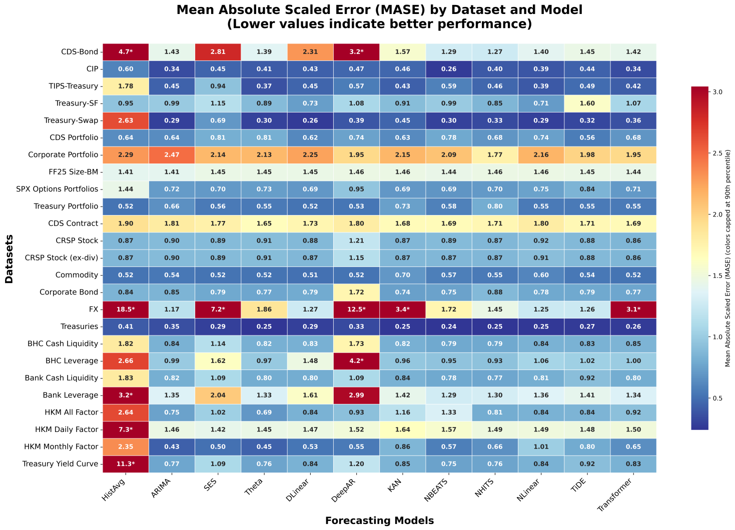
Datasets	ARIMA	SES	Theta	DLinear	DeepAR	KAN	NBEATS	NHITS	NLinear	TiDE	Transformer
CDS-Bond	0.31	0.60	0.30	0.49	0.69	0.34	0.28	0.27	0.30	0.31	0.31
CIP	0.57	0.76	0.69	0.72	0.78	0.78	0.44	0.67	0.65	0.73	0.57
TIPS-Treasury	0.25	0.53	0.21	0.26	0.32	0.24	0.33	0.26	0.22	0.28	0.24
Treasury-SF	1.04	1.21	0.94	0.77	1.14	0.96	1.05	0.90	0.75	1.69	1.13
Treasury-Swap	0.11	0.26	0.12	0.10	0.15	0.17	0.11	0.13	0.11	0.12	0.14
CDS Portfolio	1.00	1.27	1.27	0.97	1.16	0.98	1.21	1.07	1.16	0.87	1.07
Corporate Portfolio	1.08	0.93	0.93	0.98	0.85	0.94	0.91	0.77	0.94	0.86	0.85
FF25 Size-BM	1.00	1.03	1.03	1.02	1.04	1.03	1.02	1.03	1.04	1.02	1.02
SPX Options Portfolios	0.50	0.49	0.51	0.48	0.66	0.48	0.48	0.49	0.52	0.59	0.49
Treasury Portfolio	1.27	1.07	1.05	0.99	1.02	1.41	1.12	1.54	1.06	1.05	1.06
CDS Contract	0.95	0.93	0.87	0.91	0.95	0.88	0.88	0.90	0.95	0.90	0.89
CRSP Stock	1.03	1.02	1.04	1.01	1.39	1.00	1.02	0.99	1.05	1.00	0.99
CRSP Stock (ex-div)	1.03	1.02	1.04	0.99	1.31	0.99	1.00	1.00	1.04	1.01	0.99
Commodity	1.04	1.00	1.00	0.98	1.00	1.33	1.09	1.05	1.15	1.04	1.00
Corporate Bond	1.01	0.93	0.91	0.93	2.0*	0.88	0.89	1.04	0.92	0.93	0.91
FX	0.06	0.39	0.10	0.07	0.67	0.19	0.09	0.08	0.07	0.07	0.17
Treasuries	0.86	0.71	0.60	0.70	0.81	0.62	0.58	0.60	0.61	0.65	0.65
BHC Cash Liquidity	0.46	0.63	0.45	0.46	0.95	0.45	0.43	0.43	0.46	0.46	0.47
BHC Leverage	0.37	0.61	0.37	0.55	1.57	0.36	0.36	0.35	0.40	0.38	0.38
Bank Cash Liquidity	0.45	0.60	0.44	0.44	0.59	0.46	0.43	0.42	0.44	0.50	0.44
Bank Leverage	0.42	0.64	0.42	0.50	0.94	0.45	0.40	0.41	0.43	0.44	0.42
HKM All Factor	0.28	0.39	0.26	0.32	0.35	0.44	0.50	0.30	0.32	0.32	0.35
HKM Daily Factor	0.20	0.19	0.20	0.20	0.21	0.22	0.22	0.20	0.20	0.20	0.21
HKM Monthly Factor	0.18	0.22	0.19	0.23	0.24	0.37	0.25	0.28	0.43	0.34	0.28
Treasury Yield Curve	0.07	0.10	0.07	0.07	0.11	0.08	0.07	0.07	0.07	0.08	0.07

Blue = outperforms Historic Average (< 1), Red = underperforms baseline (> 1)

Relative MASE Analysis: Key Patterns

- Auto deep models lead on MASE: NBEATS (median) and NHITS (mean) post the lowest scaled errors, beating DAR, Historic Average, and simple exponential smoothing by comfortable margins
- Basis spreads love global decompositions: NLinear, NHITS, and NBEATS capture seasonal funding stress cycles and medium-run mean reversion
- Classics still matter: Theta and ARIMA remain among the best performers on smoother supervisory series
- Returns remain tough: MASE gains are modest where the Historic Average is already a strong benchmark (equities, FX, options)

Key Finding: No Universal Best Model



Mean Absolute Scaled Error (MASE) by Dataset and Model
(Lower values indicate better performance)

Datasets	HistAvg	ARIMA	SES	Theta	DLinear	DeepAR	KAN	NBEATS	NHITS	NLinear	TiDE	Transformer
CDS-Bond	4.7*	1.43	2.81	1.39	2.31	3.2*	1.57	1.29	1.27	1.40	1.45	1.42
CIP	0.60	0.34	0.45	0.41	0.43	0.47	0.46	0.26	0.40	0.39	0.44	0.34
TIPS-Treasury	1.78	0.45	0.94	0.37	0.45	0.57	0.43	0.59	0.46	0.39	0.49	0.42
Treasury-SF	0.95	0.99	1.15	0.89	0.73	1.08	0.91	0.99	0.85	0.71	1.60	1.07
Treasury-Swap	2.63	0.29	0.69	0.30	0.26	0.39	0.45	0.30	0.33	0.29	0.32	0.36
CDS Portfolio	0.64	0.64	0.81	0.81	0.62	0.74	0.63	0.78	0.68	0.74	0.56	0.68
Corporate Portfolio	2.29	2.47	2.14	2.13	2.25	1.95	2.15	2.09	1.77	2.16	1.98	1.95
FF25 Size-BM	1.41	1.41	1.45	1.45	1.45	1.46	1.46	1.44	1.46	1.46	1.45	1.44
SPX Options Portfolios	1.44	0.72	0.70	0.73	0.69	0.95	0.69	0.69	0.70	0.75	0.84	0.71
Treasury Portfolio	0.52	0.66	0.56	0.55	0.52	0.53	0.73	0.58	0.80	0.55	0.55	0.55
CDS Contract	1.90	1.81	1.77	1.65	1.73	1.80	1.68	1.69	1.71	1.80	1.71	1.69
CRSP Stock	0.87	0.90	0.89	0.91	0.88	1.21	0.87	0.89	0.87	0.92	0.88	0.86
CRSP Stock (ex-div)	0.87	0.90	0.89	0.91	0.87	1.15	0.87	0.87	0.87	0.91	0.88	0.86
Commodity	0.52	0.54	0.52	0.52	0.51	0.52	0.70	0.57	0.55	0.60	0.54	0.52
Corporate Bond	0.84	0.85	0.79	0.77	0.79	1.72	0.74	0.75	0.88	0.78	0.79	0.77
FX	18.5*	1.17	7.2*	1.86	1.27	12.5*	3.4*	1.72	1.45	1.25	1.26	3.1*
Treasuries	0.41	0.35	0.29	0.25	0.29	0.33	0.25	0.24	0.25	0.25	0.27	0.26
BHC Cash Liquidity	1.82	0.84	1.14	0.82	0.83	1.73	0.82	0.79	0.79	0.84	0.83	0.85
BHC Leverage	2.66	0.99	1.62	0.97	1.48	4.2*	0.96	0.95	0.93	1.06	1.02	1.00
Bank Cash Liquidity	1.83	0.82	1.09	0.80	0.80	1.09	0.84	0.78	0.77	0.81	0.92	0.80
Bank Leverage	3.2*	1.35	2.04	1.33	1.61	2.99	1.42	1.29	1.30	1.36	1.41	1.34
HKM All Factor	2.64	0.75	1.02	0.69	0.84	0.93	1.16	1.33	0.81	0.84	0.84	0.92
HKM Daily Factor	7.3*	1.46	1.42	1.45	1.47	1.52	1.64	1.57	1.49	1.49	1.48	1.50
HKM Monthly Factor	2.35	0.43	0.50	0.45	0.53	0.55	0.86	0.57	0.66	1.01	0.80	0.65
Treasury Yield Curve	11.3*	0.77	1.09	0.76	0.84	1.20	0.85	0.75	0.76	0.84	0.92	0.83

Forecasting Models
Mean Absolute Scaled Error (MASE) (colors capped at 90th percentile)
3.0
2.5
2.0
1.5
1.0
0.5

Performance relative to Naive baseline (blue = better, red = worse)

Main Results

- Metric choice changes leaders: MASE crowns NBEATS/NHITS, while R^2_{oos} favours Theta and ARIMA
- Returns behave as expected: Out-of-sample R^2 is essentially zero, signalling that the Historic Average remains a formidable benchmark for asset returns
- Basis spreads reward global decompositions: NLinear and NHITS exploit seasonal funding stress patterns and outperform DAR, HA, and SES by large margins
- Supervisory/other panels tilt classical: Theta (with NHITS and ARIMA close behind) offers the best trade-off between scaled errors and positive R^2

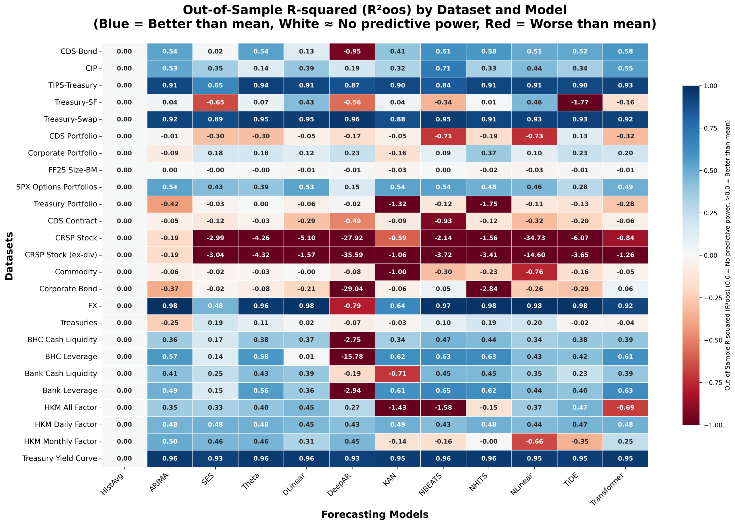
Economic Significance of Small Improvements

- Typical improvements: 3-10% better than naive baselines
- Why this matters in finance:
  - Small edges compound over time
  - Large portfolios amplify benefits
  - Institutional scale makes modest gains valuable
- Model insights by type:
  - Traditional methods (Theta, ARIMA): Still competitive
  - Deep learning: Mixed results, dataset-dependent
  - Linear deep models (DLinear/NLinear): Surprisingly effective
  - Complex architectures may not align with financial time series characteristics

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

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

Out-of-Sample R^2 Heatmap: Predictive Performance



Out-of-Sample R-squared (R²oos) by Dataset and Model
(Blue = Better than mean, White ≈ No predictive power, Red = Worse than mean)

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

Forecasting Models

Out-of-Sample R-squared (R²oos) (0.0 = No predictive power, >0.0 = Better than mean)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Positive values = better than historical mean, Negative = worse

Overall Model Performance Summary: MASE, Relative MASE, and $R^2_{\mathbf{oos}}$ Across All Datasets

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

Model Performance by Dataset Category

Basis Spreads vs Returns vs Other

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

OOS R^2 Analysis: Predictability Insights

- Overall challenge: Beating the historical mean remains difficult—R^2_{oos} hovers near zero for most models and datasets
- Classical leaders: Theta and ARIMA top the R^2 tables, reflecting their strength at reducing mean-squared error relative to the historic average
- Returns tell a familiar story: Equity, FX, option, and bond returns deliver $R^2_{\text{oos}} \approx 0$, implying the historical average is an almost unbeatable forecast—consistent with efficient-markets intuition
- Where positive R^2 emerges: Basis spreads, bank ratios, and yield curves contain structural signals (seasonality, regulation, term-structure dynamics) that Theta/NHITS can harness

Comparing Metrics: MASE vs OOS R^2

- MASE insights:
  - Absolute-error scaling highlights seasonal accuracy—crucial for basis spreads and supervisory indicators with strong cyclical structure
  - NBEATS/NHITS/NLinear dominate on this view, far ahead of DAR, HA, and SES
  - Preferred when the goal is operational monitoring or sizing seasonal funding pressure
- OOS R^2 insights:
  - Compares forecasts to the Historic Average—the right benchmark for return forecasting
  - Theta and ARIMA lead the R^2 table, but values hover near zero for returns, underscoring how hard it is to beat the mean
  - Best suited to investment and risk questions where the economic payoff hinges on improving mean-squared performance
- Divergence reveals: The same model can win on MASE yet show $R^2 \approx 0$ on returns—metric choice must match the question
- For financial stability: Use MASE-style scaling for basis spread surveillance and R^2_{oos} for asset-return risk assessments

Practical Guidance for Financial Stability Monitoring

- Basis spreads: Lean on NLinear/NHITS/NBEATS—seasonal decompositions and global training generate the largest error reductions
- Asset returns: Historic Average already performs well; use R^2_{oos} to document that most models (even auto deep nets) struggle to add value
- Supervisory & other panels: Theta with support from NHITS/ARIMA balances low MASE and positive R^2
- Global vs local: Automated global models prevent overfitting across thousands of series while keeping tuning reproducible—match the metric to the policy question

Conclusions

Contributions & Future Work

- Main Contributions:
  - First comprehensive financial time series forecasting benchmark
  - 25 datasets with canonical academic cleaning procedures
  - Baseline evidence: returns stay hard to forecast (historical average wins) while global ML models deliver large MASE gains for basis spreads (CIP, Treasury swaps) and supervisory metrics (bank liquidity, leverage)
  - Open-source implementation enabling reproducible research
- For practitioners: Use classical baselines for returns, but deploy global ML models when monitoring funding stress and liquidity indicators
- Future work:
  - Expand coverage to international markets
  - Multi-step and long-horizon forecasting
  - Integration with stress testing frameworks
  - Real-time updating and monitoring systems

Repository & Resources

Paper & Code:

`github.com/jmbejara/ftsfr`

Questions?

Contact: jbejarano@uchicago.edu

Appendix

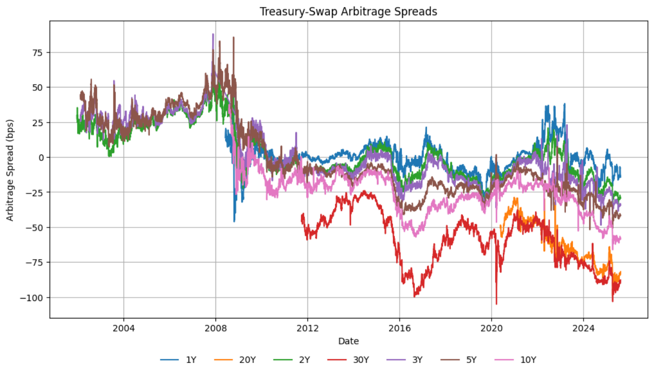
Dataset Statistics Summary

- Scale varies dramatically across datasets (after filtering):
  - Commodities: 23 series (GSCI-based)
  - CDS Contracts: 234 individual contracts retained (out of 6,552)
  - CRSP Stock: 25,095 individual equities after filtering (93.8% retention)
  - Corporate Bonds: 16,719 individual bonds (71.2% retention)
- Time coverage: Most datasets span 10-25 years
- Frequency: Daily (FX, yield curves) to quarterly (bank data); horizons set to one calendar or trading month (daily) and one period (monthly/quarterly)
- Two levels of aggregation:
  - Portfolio-level (10-50 series) for replication
  - Security-level (100s-1000s) for global forecasting
  - Some daily panels drop out if 21-day hold-outs cannot be formed (e.g., FF25 daily portfolios)

Technical Implementation Details

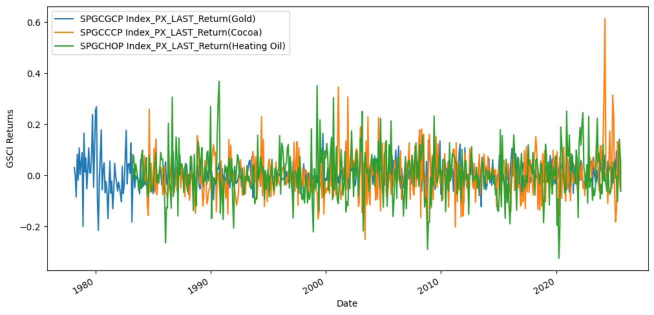
- Forecasting packages:
  - StatsForecast (Nixtla): Classical methods
  - NeuralForecast (Nixtla): Deep learning models
  - Standardized implementations ensure reproducibility
- Data preprocessing:
  - Forward-fill for missing values (no look-ahead bias)
  - Length-based filtering with dataset-specific thresholds (minimum test coverage now matches horizon)
  - Frequency-specific horizons: 30-day (calendar), 21-day (trading), 1-month, 1-quarter; up to six rolling windows per dataset
- Evaluation:
  - Rolling-origin cross-validation aligned across statistical and neural models
  - Two complementary error metrics for robust comparison
  - Minimal hyperparameter tuning for baseline results

Validation: Treasury-Swap Arbitrage Spreads



Replicates persistently negative swap spreads documented in literature

Validation: GSCI Commodity Returns



24 GSCI commodity futures following canonical Yang (2013) methodology

Institutional Quality Data Sources

- All datasets use institutional-grade data providers
- Industry-standard sources for financial research
- Strict adherence to canonical data cleaning procedures
- Next slide shows detailed data source mapping

Data Sources by Dataset

Dataset Name	Data Sources
CDS Contract	S&P Global CDS (formerly Markit)
Corporate Bond	WRDS TRACE, following Open Source Bond Asset Pricing
CRSP Stock	Center for Research in Security Prices (CRSP)
SPX Options	OptionMetrics IvyDB
Treasury Bond	Center for Research in Security Prices
CIP	Bloomberg Terminal
Bank Cash Liquidity	WRDS Bank Regulatory Call Reports
HKM Daily Factor	CRSP and Compustat, following He et al. (2017)
Treasury Yield Curve	Board of Governors of the Federal Reserve System

Replication & Validation Results

- Successfully replicated key results from canonical papers:
  - Corporate bond credit spread patterns (Nozawa 2017)
  - Options volatility risk premiums (Constantinides et al. 2013)
  - CDS basis spread dynamics (Palhares 2012)
  - Treasury yield curve characteristics (Gürkaynak et al. 2007)
- Data quality validation:
  - Cross-checked against published summary statistics
  - Verified time series properties match literature
  - Ensured no look-ahead bias in data construction
- Open source code enables:
  - Community validation and improvement
  - Extension to new datasets and methods
  - Elimination of researcher degrees of freedom