FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

A Framework for Evaluating Dataset Cherry-Picking Susceptibility in Time Series Forecasting

Tiago Araújo Coelho



Mestrado em Engenharia e Ciência de Dados

Supervisor: Carlos Soares

Co-supervisor: Vitor Cerqueira

External Supervisor: Luís Roque

July 23, 2026

A Framework for Evaluating Dataset Cherry-Picking Susceptibility in Time Series Forecasting

Tiago Araújo Coelho

Mestrado em Engenharia e Ciência de Dados

Approved in oral examination by the committee:

Chair: Prof. Ana Aguiar
External Examiner: Prof. Luís Matos
Supervisor: Prof. Carlos Soares
Co-Supervisor: Prof. Vítor Cerqueira

July 23, 2026

Declaration of Honor

I declare that this dissertation is of my own authorship and has not been previously presented and evaluated at this or any other institution. References to other authors (statements, ideas, thoughts), as well as the use of published works, scrupulously respect the rules of attribution and are duly indicated in the text and in the bibliographic references, in accordance with the referencing standards. The use of Artificial Intelligence is reported in the Appendix Transparency on the Use of Artificial Intelligence. I am aware that the practice of plagiarism or self-plagiarism constitutes academic offence, as established in the Code of Ethics of the University of Porto.

Student Name: Tiago Araújo Coelho
Student Number: 202105004
Course: Master in Data Science and Engineering

Signature: Assinado por: Tiago Araújo Coelho
Num. de Identificação: BI15962528
Data: 27-07-2026 18:52:47 +01:00
CHAVE MÓVEL
Location:
Date:

Abstract

Empirical evaluation plays a central role in the development and comparison of time series forecasting methods. Despite this importance, prevailing benchmarking practices often rely on ad hoc, poorly justified dataset selection procedures. This lack of rigour leads to systematic bias and "cherry-picking" effects, where performance claims are based on favourable data subsets that fail to generalise to broader populations. To address these limitations, this dissertation proposes a systematic, feature-based framework for auditing the representativeness of time series benchmarks *a priori*, enabling practitioners to detect potential bias in dataset selection without requiring empirical model comparisons.

By representing individual time series as vectors of structural characteristics, the framework embeds large collections of time series into a measurable feature space. Benchmark quality is then quantified using geometric and distributional metrics, such as Coverage, Maximum Mean Discrepancy (MMD), and Wasserstein distance, to assess a dataset's alignment with a large-scale global reference corpus.

Empirical application of this framework reveals a weak-to-moderate correlation between representativeness metrics and rank stability, with coverage and distributional distance metrics emerging as the most consistent predictors. A subsequent diagnostic experiment, in which standard structural features were replaced by rank-based features derived directly from model evaluation, produced strong correlations between representativeness and rank stability, confirming that the auditing framework itself is valid and the current bottleneck lies in the limited expressiveness of standard statistical features, which do not fully capture the inductive biases in dataset selection.

Ultimately, this work provides a tool for objective benchmark auditing. By establishing a formal link between dataset composition and model ranking stability, this dissertation paves the way for more transparent, reliable, and bias-aware empirical practices in time series forecasting.

Keywords: time series forecasting, benchmarking, dataset selection bias, dataset representativeness, feature-based auditing, rank stability.

Resumo

A avaliação empírica desempenha um papel central no desenvolvimento e comparação de métodos de previsão de séries temporais. Apesar desta importância, as práticas de benchmarking predominantes assentam frequentemente em procedimentos de selecção de conjuntos de dados ad hoc e pouco fundamentados. Esta falta de rigor conduz a enviesamentos sistemáticos e a efeitos de cherry-picking, em que as afirmações sobre desempenho baseiam-se em subconjuntos de dados favoráveis que não se generalizam a populações mais amplas. Para colmatar estas limitações, esta dissertação propõe um framework sistemático, baseado em características estruturais, para auditar a representatividade de benchmarks de séries temporais *a priori*, permitindo aos utilizadores detectar potenciais enviesamentos na selecção de dados sem necessitar de comparações empíricas entre modelos.

Ao representar séries temporais individuais como vectores de características estruturais, o framework transforma grandes colecções num espaço de características quantificável. A qualidade do benchmark é então quantificada através de métricas geométricas e distribucionais, como Coverage, Maximum Mean Discrepancy (MMD) e distância de Wasserstein, para avaliar o alinhamento de um conjunto de dados com um corpus de referência global de grande escala.

A aplicação empírica do framework revela correlações fracas a moderadas entre as métricas de representatividade e a estabilidade dos rankings, sendo as métricas de cobertura e de distância distribucional as que apresentam maior consistência preditiva. Uma experiência de diagnóstico subsequente, na qual as características estruturais padrão foram substituídas por características baseadas em rankings derivadas directamente da avaliação de modelos, produziu correlações fortes entre representatividade e estabilidade dos rankings, confirmando que o framework de auditoria é válido e que o principal obstáculo reside na expressividade limitada das características estatísticas padrão, que não capturam plenamente os enviesamentos indutivos na selecção de datasets.

Em última análise, este trabalho disponibiliza uma ferramenta para uso na auditoria objectiva de benchmarks. Ao estabelecer uma ligação formal entre a composição dos conjuntos de dados e a estabilidade dos rankings de modelos, esta dissertação abre caminho para práticas empíricas mais transparentes, fiáveis e conscientes do enviesamento na previsão de séries temporais.

Palavras-chave: previsão de séries temporais, benchmarking, enviesamento na seleção de dados, representatividade de conjuntos de dados, auditoria baseada em características, estabilidade de rankings.

Acknowledgements

I would like to begin by thanking my supervisors, Carlos Soares and Vítor Cerqueira, for their guidance, availability, and patience throughout this dissertation. Their feedback and discussions were invaluable in shaping both the direction and the quality of this work. I am also grateful to Luís Roque for proposing the research topic and for his continued involvement and practical perspective during the development of this project.

This work would not have been possible without the support of the people around me. To my friends, thank you for the company, the distractions when they were needed, and the encouragement when things were not going as planned. To my family, and especially my parents, thank you for your unwavering support, endless patience, and constant belief in me throughout this journey. Your encouragement has been a source of strength and motivation every step of the way.

Tiago Coelho

"It had long since come to my attention that people of accomplishment rarely sat back and let things happen to them. They went out and happened to things."

Leonardo da Vinci

Contents

1 Introduction **1**
- 1.1 Context and Motivation . . . 1
- 1.2 Research Question and Objectives . . . 2
- 1.3 Outline . . . 3

2 Literature Review **5**
- 2.1 Fundamentals of Time Series Forecasting . . . 5
- 2.2 Forecasting Methodologies . . . 7
  - 2.2.1 Statistical Time Series Methodologies . . . 7
  - 2.2.2 Machine Learning Methodologies . . . 8
  - 2.2.3 Deep Learning for Time Series Forecasting . . . 8
  - 2.2.4 Local and Global Forecasting Methodologies . . . 9
- 2.3 Evaluating Forecasting . . . 9
  - 2.3.1 Evaluation Metrics and Cross-Dataset Comparability . . . 9
  - 2.3.2 Model Rankings and Comparative Evaluation . . . 11
  - 2.3.3 Validation Protocols and Temporal Consistency . . . 12
  - 2.3.4 Benchmarking Datasets . . . 12
- 2.4 Challenges and Mitigation Strategies in Time Series Benchmarking . . . 13
  - 2.4.1 Current Mitigation Approaches . . . 13
  - 2.4.2 Methodological Limitations and Bias . . . 14
- 2.5 Feature-Based Representation of Time Series . . . 15
- 2.6 Distributional Discrepancy Measures . . . 16
- 2.7 Summary . . . 17

3 Methodology **19**
- 3.1 Introduction . . . 19
- 3.2 Feature Extraction and Representation . . . 19
- 3.3 Representativeness Auditing . . . 20
- 3.4 Rank Stability Metrics . . . 24

4 Experimental Setup and Results **27**
- 4.1 Introduction . . . 27
- 4.2 Experimental Setup . . . 27
  - 4.2.1 Reference Corpus . . . 27
  - 4.2.2 Forecasting Models and Evaluation Protocol . . . 28
  - 4.2.3 Audit and Validation Protocol . . . 30
- 4.3 Overview of Representativeness Metrics . . . 31
- 4.4 Rank Stability Results . . . 32

4.4.1 Forecasting Results and Cherry-Picking Validation 32
4.4.2 Rank Stability Across Datasets 34
4.5 Representativeness–Stability Correlation 35
4.5.1 Global Audit 35
4.5.2 Same-Frequency Audit 36
4.6 Rank-Based Feature Analysis 37
4.6.1 Results of the Rank-Based Audit 38
4.7 Discussion of Findings 41

5 Conclusions and Future Work **43**
5.1 Conclusions 43
5.2 Limitations 44
5.3 Further Work 44
5.4 Impact on Sustainable Development Goals 45

References **47**

A Reference Corpus and Experimental Setup **53**
A.1 Full Reference Corpus 54

B Supplementary Results **55**
B.1 Feature Distributions per Dataset 55
B.2 Per-Metric Dataset Representativeness Distributions 86
B.3 Representativeness–Stability Scatter Plots 91

C Transparency on the Use of AI **101**

List of Figures

3.1 High-level overview of the proposed framework. Each dataset $(d_1, d_2, \ldots, d_n)$ in the reference corpus D is mapped via feature extraction into the reference feature set D^*. A candidate dataset G undergoes the same transformation and is compared against D^* in the audit step, which quantifies its representativeness relative to the corpus. 20

3.2 Illustration of the Improved Precision metric for $k = 3$. A candidate point (cross) is counted as precise if it falls within the neighbourhood of at least one reference point (circle). 22

3.3 Illustration of the Density metric for $k = 3$. 22

3.4 Illustration of the Improved Recall metric for $k = 3$. A reference point (circle) is counted as covered if it falls within the neighbourhood of at least one candidate point (cross). 23

3.5 Illustration of the Coverage metric for $k = 3$. 23

4.1 Distribution of the six representativeness metrics across all 26 datasets, for the global and same-frequency audit variants. Each box shows the interquartile range across datasets; circles denote outliers. Higher precision, density, recall, and coverage indicate greater representativeness; lower MMD and Wasserstein distances indicate greater similarity to the reference distribution. 31

4.2 Average rank of each model per dataset, derived from per-series MASE scores. Lower rank (greener) indicates better performance. The colour variation across rows confirms that model rankings are not consistent across datasets. 33

4.3 Best-performing model per dataset (left) and number of datasets where each model ranks first (right). NHITS and NBEATS together dominate, yet disagree on which datasets each performs best — evidence of dataset-dependent ranking behaviour. 34

4.4 Distribution of the six rank stability metrics across all 26 datasets, for the global and same-frequency audit variants. For correlation-based metrics, higher values indicate greater stability; for distance-based metrics, lower values indicate greater stability. 35

4.5 Spearman correlation between representativeness metrics and rank stability metrics at $k = 10$, global audit. Positive values indicate that higher representativeness is associated with greater stability; negative values indicate the opposite. For distance-based representativeness metrics (MMD, Wasserstein), higher values indicate lower representativeness, so positive correlations with distance-based stability metrics are consistent with the hypothesis. 36

4.6 Spearman correlation between representativeness metrics and rank stability metrics at $k = 10$, same-frequency audit. 37

4.7 Spearman correlation between representativeness and stability metrics using rank-based features (Global Audit, $k = 10$). . . . 39
4.8 Scatter plots detailing the strong monotonic relationship between MMD distance and rank stability metrics in the rank-based feature space. . . . 40
4.9 Scatter plots demonstrating the relationship between Wasserstein distance and rank stability metrics. . . . 40
4.10 Scatter plots demonstrating the negative correlation between Coverage and ranking displacement metrics. . . . 41

B.1 Feature distributions for M4 Daily. . . . 56
B.2 Feature distributions for M4 Hourly. . . . 57
B.3 Feature distributions for M4 Monthly. . . . 58
B.4 Feature distributions for M4 Quarterly. . . . 59
B.5 Feature distributions for M4 Weekly. . . . 60
B.6 Feature distributions for M4 Yearly. . . . 61
B.7 Feature distributions for CIF 2016 (Monthly). . . . 62
B.8 Feature distributions for CIF 2016 (Quarterly). . . . 63
B.9 Feature distributions for Electricity (Hourly). . . . 64
B.10 Feature distributions for Electricity (Daily resampled). . . . 65
B.11 Feature distributions for Electricity (Weekly). . . . 66
B.12 Feature distributions for FRED-MD (Monthly). . . . 67
B.13 Feature distributions for FRED-MD (Quarterly). . . . 68
B.14 Feature distributions for Hospital. . . . 69
B.15 Feature distributions for M1 (Monthly). . . . 70
B.16 Feature distributions for M1 (Quarterly). . . . 71
B.17 Feature distributions for M1 (Yearly). . . . 72
B.18 Feature distributions for M3 (Monthly). . . . 73
B.19 Feature distributions for M3 (Quarterly). . . . 74
B.20 Feature distributions for M3 (Yearly). . . . 75
B.21 Feature distributions for Tourism (Monthly). . . . 76
B.22 Feature distributions for Tourism (Quarterly). . . . 77
B.23 Feature distributions for Tourism (Yearly). . . . 78
B.24 Feature distributions for Traffic (Hourly). . . . 79
B.26 Feature distributions for Traffic (Weekly resampled). . . . 80
B.25 Feature distributions for Traffic (Daily resampled). . . . 81
B.27 Feature distributions for Weather (Daily). . . . 82
B.28 Feature distributions for Weather (Monthly resampled). . . . 83
B.29 Feature distributions for Weather (Weekly resampled). . . . 84
B.30 Feature distributions for NN5. . . . 85
B.31 Precision scores per dataset at $k = 10$, ordered by frequency group. Dark bars = global audit; light bars = same-frequency audit. . . . 86
B.32 Density scores per dataset at $k = 10$, ordered by frequency group. . . . 87
B.33 Recall scores per dataset at $k = 10$, ordered by frequency group. . . . 88
B.34 Coverage scores per dataset at $k = 10$, ordered by frequency group. . . . 89
B.35 MMD distance per dataset at $k = 10$, ordered by frequency group. . . . 90
B.36 Wasserstein distance per dataset at $k = 10$, ordered by frequency group. . . . 91
B.37 Precision vs rank stability metrics — global audit, $k = 10$. . . . 92
B.38 Density vs rank stability metrics — global audit, $k = 10$. . . . 92
B.39 Recall vs rank stability metrics — global audit, $k = 10$. . . . 93

B.40 Coverage vs rank stability metrics — global audit, $k = 10$. 93
B.41 MMD distance vs rank stability metrics — global audit, $k = 10$. 94
B.42 Wasserstein distance vs rank stability metrics — global audit, $k = 10$. 94
B.43 Precision vs rank stability metrics — same-frequency audit, $k = 10$. 95
B.44 Density vs rank stability metrics — same-frequency audit, $k = 10$. 95
B.45 Recall vs rank stability metrics — same-frequency audit, $k = 10$. 96
B.46 Coverage vs rank stability metrics — same-frequency audit, $k = 10$. 96
B.47 MMD distance vs rank stability metrics — same-frequency audit, $k = 10$. 97
B.48 Wasserstein distance vs rank stability metrics — same-frequency audit, $k = 10$. . 97
B.49 Precision vs rank stability metrics — rank-based global audit, $k = 10$. The distribution is perfectly flat at 1.0, reflecting complete containment within the global combinatorial space. 98
B.50 Density vs rank stability metrics — rank-based global audit, $k = 10$. 98
B.51 Recall vs rank stability metrics — rank-based global audit, $k = 10$. Similar to Precision, Recall saturates near 1.0. 99

List of Tables

4.1 Datasets used in the forecasting and representativeness experiments (filtered corpus). 29

A.1 Full reference corpus used for representativeness auditing. 54

Abbreviations

TSF	Time Series Forecasting
SOTA	State of the Art
ML	Machine Learning
DL	Deep Learning
ARIMA	AutoRegressive Integrated Moving Average
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
TFB	Towards Comprehensive and Fair Benchmarking
TS	Time Series
pyMDMA	Python Multimodal Data Metrics for Auditing
MASE	Mean Absolute Scaled Error
MAE	Mean Absolute Error
MMD	Maximum Mean Discrepancy
STL	Seasonal and Trend decomposition using Loess
ACF	Autocorrelation Function
PACF	Partial Autocorrelation Function
KPSS	Kwiatkowski–Phillips–Schmidt–Shin
ARCH	Autoregressive Conditional Heteroscedasticity
GARCH	Generalised Autoregressive Conditional Heteroscedasticity
NHITS	Neural Hierarchical Interpolation for Time Series
NBEATS	Neural Basis Expansion Analysis for Time Series
TiDE	Time-series Dense Encoder
MLP	Multilayer Perceptron
NLinear	Normalised Linear
DLinear	Decomposition Linear
SDG	Sustainable Development Goal
FFORMA	Feature-based Forecast Model Averaging
M1	First Makridakis Competition
M3	Third Makridakis Competition
M4	Fourth Makridakis Competition
FRED-MD	Federal Reserve Economic Data – Monthly Database
NN5	Neural Network 5 Competition
CIF	Computational Intelligence in Forecasting

Chapter 1

Introduction

1.1 Context and Motivation

Time series forecasting (TSF) is essential to the success of various modern applications, including finance (Kim, 2003; Sezer et al., 2020), energy systems (Martínez-Álvarez et al., 2011; Deb et al., 2017), healthcare (Chimmula and Zhang, 2020; Kaushik et al., 2020), and industrial processes (Kashpruk et al., 2023). Driven by the critical need for accurate predictions in strategic decision-making (Hyndman and Athanasopoulos, 2021), considerable effort has been devoted to improving forecasting accuracy, leading to the development of methods ranging from classical statistical models to complex deep learning architectures (Lim and Zohren, 2021; Hochreiter and Schmidhuber, 1997; Box and Pierce, 1970). To evaluate these advancements reliably, large public repositories such as the M-competitions (Makridakis and Hibon, 2000; Makridakis et al., 2020, 2022), the UCR archive (Dau et al., 2019), and the Monash Forecasting Archive (Godahewa et al., 2021) have been established. By providing standardised, reproducible testing environments, these repositories serve as indispensable benchmarks and are widely regarded as neutral proxies for real-world forecasting scenarios.

Despite the availability of these comprehensive resources, recent work has exposed critical flaws in the conduct of empirical evaluations in TSF (Roque et al., 2025; Hewamalage et al., 2023). Driven by factors such as computational constraints, current benchmarking practices often entail methodological compromises. Three recurring issues can be broadly identified:

1. Inconsistent evaluation settings across studies, such as the choice of error measures and validation protocols (Hewamalage et al., 2023);

2. Benchmark datasets whose statistical properties fail to reflect the data encountered in practical forecasting applications (Koh et al., 2021);

3. Conclusions drawn from a few datasets, or from isolated subsets extracted from larger archives (Roque et al., 2025).

This dissertation focuses on the third issue, which has received comparatively less methodological attention. The work that does exist approaches it retrospectively, identifying and quantifying the consequences of biased dataset selection after evaluation (Roque et al., 2025). However, no principled methodology has yet been proposed to guide benchmark construction and prevent this bias from arising in the first place. This dissertation addresses that gap.

1.2 Research Question and Objectives

The overarching objective of this dissertation is to mitigate dataset selection bias in time-series forecasting by establishing a systematic framework for assessing the representativeness of benchmark datasets prior to model evaluation. Rather than focusing on forecasting accuracy itself, the aim is to improve the reliability, robustness, and reproducibility of empirical conclusions in forecasting research.

Concretely, this work pursues three aims: (i) providing a principled basis for assessing benchmark representativeness in time series forecasting; (ii) enabling the detection of biased dataset selections prior to experimental evaluation; and (iii) supporting the construction of balanced, diverse benchmarking subsets that more faithfully reflect a target domain or population.

To achieve these objectives, the proposed approach relies on a feature-based characterisation of time series datasets (Fulcher and Jones, 2017; Lubba et al., 2019; Barandas et al., 2020; Hyndman et al., 2019), which allows benchmark collections to be analysed and compared as distributions in a common representation space. Within this space, statistical distance measures procedures are used to assess representativeness and identify potential selection bias (Façoco et al., 2025). The effectiveness of the proposed framework is then evaluated empirically by analysing its impact on benchmark stability, model ranking consistency, and the reproducibility of experimental results.

Ultimately, these aims and methodological steps are designed to address the central research question guiding this dissertation:

- **RQ:** How to leverage time series feature extraction, and distributional analysis to quantify the representativeness of a candidate dataset relative to the population of time series it is intended to proxy?

This question is decomposed into three more granular sub-questions:

- **RQ1:** To what extent can structural time series features characterise the composition of a benchmark dataset relative to a broader reference population?
- **RQ2:** How effectively can feature-based distributional metrics quantify and detect lack of representativeness within benchmark datasets?
- **RQ3:** To what extent can the proposed representativeness metrics act as an *a priori* indicator of model rank instability caused by dataset selection bias?

1.3 Outline

In addition to the present chapter, the remainder of this dissertation is organised as follows:

- **Chapter 2: Literature Review** reviews the relevant literature and key concepts foundational to this research. It covers time series forecasting methodologies, current empirical evaluation practices, and feature-based dataset representation. Furthermore, it examines existing work on benchmarking bias and dataset auditing, establishing the context and identifying the gaps this dissertation addresses.
- **Chapter 3: Methodology** formalises the research problem and presents the proposed framework, a tool-agnostic methodology for quantifying the representativeness of a benchmark dataset with respect to a reference corpus. It describes the framework's conceptual components: the mapping of time series into a structural feature space, the representativeness metrics used to compare feature distributions, and the rank stability measures used to characterise model ranking behaviour.
- **Chapter 4: Experimental Setup and Results** describes the empirical instantiation of the framework — including the reference corpus, the forecasting models, and the audit protocol — and presents the results of the dataset audits. It reports the correlation analysis between representativeness scores and model rank stability, contrasting the global audit with the temporally isolated same-frequency audit. It further details a rank-based diagnostic experiment that isolates the framework's geometric effectiveness from the limitations of the statistical feature space.
- **Chapter 5: Conclusion and Future Work** summarises the main contributions and the answers to the research questions. It discusses the limitations of the proposed approach and outlines directions for future research, such as integrating meta-learning landmarkers.

Chapter 2

Literature Review

This chapter provides an overview of the theories and methods for evaluating empirical results in time series forecasting. It focuses on benchmarking practices and dataset selection, summarising the existing literature on TSF methodologies, evaluation metrics, and statistical tools for dataset-level analysis. By highlighting the limitations of existing benchmarking, particularly concerns about representativeness and ranking variability, this review establishes the motivation for the methodology presented in the following chapter.

The rest of this chapter is organised as follows. Section 2.1 introduces the basics of time series data and reviews fundamental concepts, such as the structural characteristics of temporal data. Section 2.2 reviews the evolution of TSF methodologies, from classical statistical methods through machine learning and deep learning architectures, including the distinction between local and global training strategies. Section 2.3 examines the various metrics and validation protocols used to evaluate forecasting performance, the benchmark datasets currently driving empirical research, and the role of model rankings in comparative evaluation. Section 2.4 analyses the challenges in current benchmarking practices, focusing on dataset selection bias and the limitations of existing mitigation strategies. Section 2.5 presents feature-based representations, which map time series into a vector space to enable auditing and selection. Section 2.6 reviews distributional discrepancy measures used to quantify how far a benchmark deviates from a reference distribution. Finally, Section 2.7 concludes the chapter by summarising the main findings and synthesising identified research gaps, providing a clear justification for the methodology developed.

2.1 Fundamentals of Time Series Forecasting

A time series is a sequence of data points arranged in chronological order. It is usually represented as $Y = \{y_1, y_2, \ldots, y_T\}$ for a univariate case, where only one variable is measured over time. A multivariate time series extends this concept by analysing multiple variables simultaneously. These points can be recorded with different frequencies, such as daily and monthly (Hyndman and Athanasopoulos, 2021).

Time Series Forecasting involves estimating future values of a stochastic process based on its past realisations (Hyndman and Athanasopoulos, 2021). Formally, let $\{y_t\}_{t=1}^{T}$ denote an observed univariate time series. Given an information set $\mathscr{I}_T = \{y_1, \ldots, y_T\}$ available at time T, the task is to produce predictions $\{\hat{y}_{T+h}\}_{h=1}^{H}$ over a forecast horizon H through a forecasting function $f(\cdot)$ such that:

$$\hat{y}_{T+h} = f(\mathscr{I}_T), \quad h = 1, \ldots, H \tag{2.1}$$

where $f(\cdot)$ is designed to capture the temporal dependence and structural patterns present in the data.

In the machine learning paradigm, forecasting is treated as a supervised regression task. Rather than modelling the temporal process directly, the series is unrolled into a tabular structure through a procedure known as time-delay embedding (Bontempi et al., 2013). Concretely, a window of length L (the lookback or input size) slides across the series: at each position t, the L preceding observations form a predictor vector $\mathbf{x}_t = [y_{t-1}, y_{t-2}, \ldots, y_{t-L}]^\top \in \mathbb{R}^L$, paired with the value y_t to be predicted. Collecting these pairs yields a training set $\mathscr{S} = \{(\mathbf{x}_t, y_t)\}_{t=L+1}^{T}$ suitable to any standard regression learner. The lookback length L thus acts as a modelling lever, governing how much historical context the learner can exploit. When the data comprises not a single series but a collection of related series, this embedding step is applied to each series independently, and the resulting supervised pairs are pooled into a single corpus. Denoting the per-series training sets as $\mathscr{S}^{(1)}, \ldots, \mathscr{S}^{(K)}$ for K series, the global learner is fitted on their union $\mathscr{S}_{\text{glob}} = \bigcup_{k=1}^{K} \mathscr{S}^{(k)}$, allowing a single parameter set to absorb regularities shared across the collection (Montero-Manso and Hyndman, 2021).

Unlike traditional machine learning, which typically assumes observations are independent and identically distributed, time series data violate this assumption by construction (Hyndman and Athanasopoulos, 2021; Box et al., 2015). Observations at any given time t are correlated with past values, and these temporal dependencies pose specific challenges for model evaluation — requiring validation strategies such as temporal split approaches to prevent information leakage (Bergmeir and Benítez, 2012; Bergmeir et al., 2018).

A common analytical method represents a univariate time series as a combination of hidden structural components. Decomposing a series into these components makes it possible to model each source of variation separately, improving both interpretability and forecasting accuracy. These main components are usually described as follows (Adhikari and Agrawal, 2013; Hyndman and Athanasopoulos, 2021):

- **Trend–Cycle Component:** Represents the long-term movement of the series, capturing persistent increases or decreases as well as non-periodic cyclical behaviour.
- **Seasonal Component:** Describes systematic and calendar-related patterns that repeat at fixed and known frequencies, such as daily, weekly, or annual cycles.
- **Remainder Component:** Encompasses irregular and unpredictable variations that are not explained by trend or seasonality, including noise and external disturbances.

Beyond these standard components, time series often exhibit additional complexities, such as non-stationarity, structural breaks, regime changes, and missing or corrupted observations (Box et al., 2015; Hyndman and Athanasopoulos, 2021). Such irregularities can significantly impact both model behaviour and the validity of empirical performance comparisons.

The presence and interaction of these structural properties differ across applications and datasets. This leads to notable differences within large forecasting collections (Godahewa et al., 2021; Makridakis et al., 2020). Such structural heterogeneity means that empirical conclusions are inherently conditioned on the datasets chosen for evaluation, a dependency that is rarely made explicit in benchmarking studies. Characterising the distribution of structural properties across benchmark collections should therefore be a prerequisite for reliable and generalisable empirical evaluation.

2.2 Forecasting Methodologies

Forecasting methodologies have evolved substantially over the past decades, driven by advances in statistical theory, computational capacity, and the growing availability of large-scale time series data. This section traces that evolution across four paradigms: classical statistical models, machine learning approaches, deep learning architectures, and the distinction between local and global training strategies.

2.2.1 Statistical Time Series Methodologies

Early forecasting methodologies were predominantly grounded in classical statistical analysis, with ARIMA and Exponential Smoothing forming the foundational base of the field (Box and Pierce, 1970; Box et al., 2015; Hyndman et al., 2008). ARIMA models combine autoregression, differencing, and moving averages to capture linear dependencies and trends in stationary sequences (Box and Pierce, 1970; Box et al., 2015). Exponential smoothing methods, such as Holt-Winters, model trend and seasonality through weighted averages of past observations (Hyndman et al., 2008). These approaches are characterised by explicit probabilistic formulations that are interpretable and well understood theoretically (Box et al., 2015; Hyndman et al., 2008), and are typically applied independently to each individual series (Montero-Manso and Hyndman, 2021).

Despite their strengths, classical methods carry important limitations. ARIMA requires the series to be stationary and depends on careful manual selection of differencing and lag parameters (Box and Pierce, 1970; Box et al., 2015). Exponential smoothing methods can struggle with complex or multiple seasonal patterns (Hyndman et al., 2008). More broadly, the assumption of linearity restricts their ability to capture non-linear dynamics present in many real-world series (Box et al., 2015; Adhikari and Agrawal, 2013). Empirical evaluation under this paradigm was conducted on relatively small, carefully curated datasets (Makridakis and Hibon, 2000) — a practice that persists in some benchmarking traditions today.

2.2.2 Machine Learning Methodologies

The supervised learning framing introduced in the previous section enabled the application of a broad class of machine learning algorithms to forecasting. Ensemble methods such as Random Forests and Gradient Boosting became widely adopted due to their ability to capture non-linear relationships without requiring explicit model specification (Breiman, 2001; Chen and Guestrin, 2016). These approaches relax the linearity assumptions of classical methods and can naturally incorporate exogenous features, calendar effects, and cross-series information within the tabular input structure, with tree-based implementations in particular demonstrating strong empirical performance (Januschowski et al., 2022).

That said, their performance is intrinsically linked to the quality of the feature representation, with feature engineering as the primary lever practitioners use to improve results in practice (Januschowski et al., 2022). This dependence on feature engineering introduced a new source of experimental variability, making fair cross-study comparisons difficult when preprocessing choices are not standardised. Compounding this, the lack of consensus on performance estimation procedures further undermines the reliability of cross-study comparisons (Cerqueira et al., 2020).

2.2.3 Deep Learning for Time Series Forecasting

Deep learning introduced a fundamentally different approach to forecasting: rather than specifying an explicit model structure, these methods learn temporal representations directly from data through flexible function approximation (Lim and Zohren, 2021). The earliest deep learning architectures applied to forecasting were Recurrent Neural Networks (RNNs) (Elman, 1990) and their variants, notably Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), which were designed to capture long-range dependencies in sequential data through gated memory mechanisms. While effective, recurrent architectures are prone to vanishing gradients and are computationally expensive to train on long sequences (Bengio et al., 1994).

Transformer-based architectures, originally developed for natural language processing (Vaswani et al., 2017), were subsequently adapted to time series forecasting and demonstrated strong performance across a range of benchmarks (Wen et al., 2023). Their self-attention mechanism allows each position in the input sequence to attend directly to every other position, enabling the capture of complex long-range dependencies without the sequential bottleneck of recurrent models. However, this comes at the cost of quadratic computational and memory complexity with respect to sequence length, motivating efficient variants such as the Informer (Zhou et al., 2021).

Despite the apparent dominance of Transformer-based models on forecasting leaderboards, a critical study by Zeng et al. demonstrated that a simple one-layer linear model could outperform state-of-the-art Transformer architectures across multiple standard benchmarks (Zeng et al., 2023). This finding raised fundamental questions about whether reported improvements reflected genuine modelling advances or were artefacts of experimental design — a concern documented across the forecasting literature (Hewamalage et al., 2023) and directly relevant to the benchmarking problem addressed in this dissertation. It also renewed interest in simpler, computationally efficient

architectures, including MLP-based models such as NHITS (Challu et al., 2023) and TiDE (Das et al., 2023).

Deep learning models support diverse forecasting setups — point, interval, and probabilistic — and can incorporate covariates, calendar effects, and cross-series information. Their capacity to scale across large, heterogeneous collections of time series makes them particularly relevant in global forecasting settings (Salinas et al., 2020).

2.2.4 Local and Global Forecasting Methodologies

Forecasting approaches can also be distinguished according to the scope over which models are trained (Januschowski et al., 2020). In local forecasting methodologies, a separate model is fitted independently to each individual time series, allowing model parameters to be specialised to the dynamics of a single sequence. This strategy is commonly used in traditional statistical forecasting and in settings where series exhibit highly idiosyncratic behaviour.

Global forecasting methodologies, in contrast, train a single model simultaneously across a collection of multiple time series (Januschowski et al., 2020). By sharing parameters across series, global models can exploit common temporal patterns that would be invisible to a local model operating on a single sequence. Crucially, empirical evidence has established that global models frequently outperform local counterparts even across highly heterogeneous collections (Montero-Manso and Hyndman, 2021; Godahewa et al., 2021). The performance advantage is attributed to cross-series generalisation, whereby the model learns to identify and transfer structural representations across different subsets of the data space.

This distinction between local and global training is directly relevant to benchmarking. Because global models possess the capacity to learn complex, domain-specific representations, a dataset collection used to evaluate them must be sufficiently diverse to reflect the true variance of real-world scenarios. A benchmark composed of structurally homogeneous series may inadvertently reward global models that overfit to a narrow set of patterns. Evaluating models on such restricted subsets inflates their apparent generalisation capability, yielding biased performance estimates compared to those obtained in a truly representative evaluation setting (Roque et al., 2025).

2.3 Evaluating Forecasting

2.3.1 Evaluation Metrics and Cross-Dataset Comparability

The evaluation of forecasting models depends on quantitative measures that summarise the differences between observed values and their corresponding forecasts. In time series forecasting, these measures help compare performance across different forecast horizons, series, and datasets. They are important for large-scale benchmarking studies (Hyndman and Koehler, 2006). Different types of metrics capture various aspects of forecast error and reflect different assumptions about scale, symmetry, and comparability across datasets. The most commonly used forecasting metrics in the literature can be broadly categorised as follows:

- **Absolute error metrics** quantify the magnitude of forecast errors in the same scale as the observed data. A commonly used example is the Mean Absolute Error (MAE),

$$\text{MAE} = \frac{1}{H}\sum_{h=1}^{H}|y_{T+h} - \hat{y}_{T+h}|,$$

  while the Root Mean Squared Error (RMSE),

$$\text{RMSE} = \sqrt{\frac{1}{H}\sum_{h=1}^{H}(y_{T+h} - \hat{y}_{T+h})^2},$$

  places greater emphasis on larger deviations. Although intuitive within individual datasets, both metrics are inherently scale-dependent, limiting their suitability for aggregation across heterogeneous collections.

- **Percentage error metrics** express forecast errors relative to the magnitude of the observed values. A classical example is the Mean Absolute Percentage Error (MAPE),

$$\text{MAPE} = \frac{100}{H}\sum_{h=1}^{H}\left|\frac{y_{T+h} - \hat{y}_{T+h}}{y_{T+h}}\right|,$$

  which aims to provide a scale-free comparison. However, percentage-based metrics are sensitive to small or zero-valued observations and may exhibit asymmetric error penalisation.

- **Relative error metrics** evaluate forecast performance with respect to a reference or benchmark model. A common way to show error ratios is to compare a candidate model to a baseline forecast, framing evaluation in comparative rather than absolute terms.

$$\text{RelErr} = \frac{\sum_{h=1}^{H}|y_{T+h} - \hat{y}_{T+h}|}{\sum_{h=1}^{H}|y_{T+h} - \hat{y}^{\text{ref}}_{T+h}|},$$

- **Scaled error metrics** normalise forecast errors using statistics derived from the training data or from a benchmark forecast. An example is the Mean Absolute Scaled Error (MASE),

$$\text{MASE} = \frac{\frac{1}{H}\sum_{h=1}^{H}|y_{T+h} - \hat{y}_{T+h}|}{\frac{1}{T-m}\sum_{t=m+1}^{T}|y_t - y_{t-m}|},$$

  where m denotes the seasonal period, this formulation enables scale-independent comparison across time series with differing magnitudes and units.

The choice of metric is therefore not merely a technical convention — it encodes assumptions about what constitutes good forecast performance and directly shapes how models are ranked in comparative studies (Hyndman and Koehler, 2006; Hewamalage et al., 2023).

Beyond metric selection, the validity of evaluation results depends on how forecast performance is estimated. The temporal dependence structure of time series data violates the independence assumption underlying standard k-fold cross-validation, necessitating evaluation protocols that respect chronological ordering (Bergmeir and Benítez, 2012; Bergmeir et al., 2018). Together, metric choice and evaluation protocol determine the reliability of empirical comparisons, and both are sensitive to the statistical composition of the datasets on which they are applied (Roque et al., 2025; Hewamalage et al., 2023).

2.3.2 Model Rankings and Comparative Evaluation

Beyond individual error metrics, empirical studies in forecasting frequently summarise model performance through rankings — ordinal comparisons that indicate which methods perform best across a set of datasets or series. Rankings offer an intuitive summary of relative model quality and are widely used in large-scale benchmarking studies and forecasting competitions to identify state-of-the-art methods (Makridakis et al., 2020; Godahewa et al., 2021).

Rankings are typically derived by aggregating per-series or per-dataset error scores, either through average rank, win-count statistics, or the aggregation of scale-independent metrics (such as MASE). To ensure that observed differences in rankings are not merely due to random variance, rigorous evaluations couple these ordinal summaries with statistical significance testing, such as the Friedman test or the Diebold-Mariano test (Hewamalage et al., 2023). Furthermore, when analysing the stability of a model's performance across different evaluation conditions, rank-based correlation measures are employed. Spearman's rank correlation (Zar, 2005) and Kendall's τ (Kendall, 1938) are standard choices, as they quantify the degree of agreement between two orderings without assuming a linear relationship between the underlying scores.

However, model rankings are profoundly sensitive to a range of methodological choices that extend beyond the metric or statistical test used. The selection of datasets, the forecast horizon, the choice of baseline models, and the validation protocol can all substantially alter the relative ordering of methods (Roque et al., 2025; Hewamalage et al., 2023). A model that ranks first on one subset of datasets may rank last on another, even when both subsets are drawn from the exact same archive (Roque et al., 2025). This instability suggests that reported rankings often reflect the specific datasets and conditions under which they were produced just as much as they reflect genuine algorithmic superiority.

This observation is central to the present dissertation. When benchmark composition systematically influences model rankings, the representativeness of the benchmark — the degree to which it reflects the target population of time series — becomes a strict prerequisite for drawing reliable conclusions from any comparative evaluation. Establishing principled tools to assess this representativeness is therefore not merely a methodological refinement, but a fundamental condition for the validity of empirical claims in forecasting research (Roque et al., 2025; Koh et al., 2021).

2.3.3 Validation Protocols and Temporal Consistency

In time series forecasting, the design of validation protocols is a fundamental component of empirical evaluation, as it defines how forecasting performance is estimated and compared. Because time series observations are ordered sequentially by construction, standard validation strategies typically preserve this chronological structure. While cross-validation has been shown to be empirically viable for time series when blocked schemes are used (Bergmeir and Benítez, 2012), and theoretically valid for purely autoregressive models provided the candidate models have uncorrelated errors (Bergmeir et al., 2018), preserving temporal order remains the mandated best practice to prevent future-to-past information leakage in real-world, non-stationary data (Hewamalage et al., 2023).

Standard practice, therefore, involves holdout evaluation or rolling-origin protocols. In a holdout setup, the model is trained on an initial chronological period and assessed on a subsequent, unseen forecast horizon (Cerqueira et al., 2020). Building on this, rolling-origin evaluation (also known as walk-forward validation or time series cross-validation) involves evaluating forecasts from multiple sequential starting points. By re-estimating models across different time frames and shifting the forecast origin forward, this approach enables the aggregation of performance metrics across multiple test horizons, providing a more comprehensive and robust view of model behaviour over time (Hewamalage et al., 2023; Cerqueira et al., 2020).

Taken together, choices regarding windowing strategies, horizon lengths, and update frequencies constitute a critical but often under-reported dimension of experimental configuration. This variation across studies can substantially affect the conclusions drawn about relative model performance (Cerqueira et al., 2020; Hewamalage et al., 2023), further highlighting the need for rigorously standardised benchmarking practices.

2.3.4 Benchmarking Datasets

Benchmark datasets serve as shared reference points for comparing forecasting methods and substantiating methodological claims. The availability of public repositories and standardised benchmarks has substantially improved the reproducibility and transparency of empirical results, enabling large-scale comparative studies across diverse application domains.

The landscape of benchmarking has evolved through a continuous effort to build expansive repositories and organise global forecasting competitions. These initiatives aggregate datasets from various sources, sectors, and temporal scales. Prominent examples include the long-running M-competitions (Makridakis and Hibon, 2000; Makridakis et al., 2020, 2022) and the comprehensive Monash Time Series Forecasting Archive (Godahewa et al., 2021). Alongside these aggregated archives, specific multivariate datasets—such as those encompassing energy consumption, traffic flow, and weather patterns—have become standard benchmarks in the deep learning forecasting literature (Zeng et al., 2023). Collectively, these collections are designed to provide structural diversity, encompassing a wide range of seasonal patterns, temporal frequencies, and dynamic behaviours.

More recently, efforts have shifted towards designing benchmarks explicitly curated for breadth and representativeness, expanding the range of domains, frequencies, and prediction horizons covered by standard evaluation (Aksu et al., 2024; Shchur et al., 2025).

Underlying this design is the assumption that a sufficiently diverse collection of datasets can serve as a proxy for real-world forecasting complexity (Godahewa et al., 2021; Makridakis et al., 2020). When this assumption holds, performance aggregated across a benchmark collection generalises meaningfully to unseen domains. When it does not, because the selected datasets are unrepresentative or systematically skewed, the resulting conclusions may be misleading, regardless of the care taken in model design or metric selection.

In large-scale benchmarking frameworks, performance is assessed by aggregating error metrics across multiple datasets, relying on scale-independent measures such as the Mean Absolute Scaled Error (MASE) or relative errors to ensure comparability across series with differing magnitudes (Hyndman and Koehler, 2006; Makridakis et al., 2020). The validity of these aggregated conclusions, however, depends critically on whether the underlying dataset collection is truly representative of the target domain—a fundamental dependency that is rarely formally audited prior to evaluation.

2.4 Challenges and Mitigation Strategies in Time Series Benchmarking

Benchmarking serves as the primary mechanism for empirical validation in time series forecasting, substantiating scientific claims through shared repositories and standardised evaluation protocols (Makridakis and Hibon, 2000; Godahewa et al., 2021; Qiu et al., 2024; Aksu et al., 2024). However, the reliability of these empirical conclusions often depends on methodological assumptions about the neutrality and composition of the datasets used (Roque et al., 2025).

2.4.1 Current Mitigation Approaches

Several research trajectories have emerged to address dataset selection bias and improve the reliability of the empirical evaluation infrastructure. One prominent approach is expanding data scale through large-scale repositories and competitions. The M-competitions, for instance, have progressively increased evaluation diversity through carefully curated collections (Makridakis et al., 2020, 2022). Concurrently, the introduction of the Monash Forecasting Archive (Godahewa et al., 2021) — inspired by the success of the UCR archive in the adjacent field of time series classification (Dau et al., 2019) — has finally broadened standard access to multi-domain forecasting data. Beyond curated academic environments, forecasting competitions on platforms such as Kaggle provide access to high-dimensional, real-world data. These environments are increasingly recognised as valuable learning opportunities, showcasing the effectiveness of complex machine learning and cross-learning strategies on corporate-style datasets (Bojer and Meldgaard, 2021). However, because their primary focus is on maximising predictive performance rather than enforcing

controlled experimental designs, they represent a different paradigm from formal benchmarking.

Beyond data availability, recent frameworks have focused on reducing experimental variability by standardising preprocessing pipelines, splitting strategies, and reporting conventions (Qiu et al., 2024; Aksu et al., 2024). More sophisticated still, recent benchmarks such as GIFT-Eval (Aksu et al., 2024) and FEV-Bench (Shchur et al., 2025) have been deliberately curated for breadth and representativeness, explicitly recognising that benchmark design is itself a methodological choice with direct consequences for the validity of empirical conclusions. Complementarily, synthetic generation frameworks such as GRATIS have been proposed to explore the instance space by creating series with diverse and controllable statistical properties (Kang et al., 2020).

Despite these significant methodological advancements, a fundamental vulnerability remains. While standardisation frameworks ensure that models are evaluated consistently, the resulting conclusions remain strictly conditional on the benchmark's composition. Providing massive archives, curated benchmarks, or synthetic generators does not prevent researchers from selectively evaluating models on non-representative subsets. Indeed, model rankings remain highly sensitive to the specific datasets included, even when evaluation protocols are rigorously fixed (Roque et al., 2025). Consequently, while current mitigation approaches enhance data availability and protocol standardisation, they do not inherently address the structural vulnerabilities embedded in dataset selection practices.

2.4.2 Methodological Limitations and Bias

A significant gap in current practice remains the absence of a formal framework for subset selection. The extensive scale of modern forecasting archives, such as the Monash Repository, does not guarantee that a selected portion is suitable for a specific experimental objective (Godahewa et al., 2021). In the absence of rigorous selection criteria, convenience sampling remains prevalent.

Recent reviews of empirical evaluation practices vividly illustrate this vulnerability (Roque et al., 2025). Despite the availability of massive archives, state-of-the-art deep learning architectures are frequently evaluated on highly restricted subsets, typically comprising merely three to six datasets. This heterogeneity becomes apparent when examining the benchmark choices adopted by different architectural families. Methods designed for long-horizon forecasting, including Informer (Zhou et al., 2021), NHITS (Challu et al., 2023), and TiDE (Das et al., 2023), are commonly evaluated on datasets such as ETT, Electricity, and Traffic, which contain relatively few time series, each recorded over exceptionally long spans. By contrast, traditional forecasting benchmarks, including the Tourism and M-competition collections (Makridakis and Hibon, 2000; Makridakis et al., 2020), represent the opposite evaluation regime, comprising large numbers of comparatively short series. The transition from N-BEATS to NHITS provides a particularly illustrative example of this divide. Although NHITS is a direct architectural descendant of N-BEATS (Oreshkin et al., 2020), its empirical evaluation shifted almost entirely towards long-horizon benchmarks, whereas N-BEATS itself was originally validated on classical forecasting collections. Consequently, each evaluation setting captures only a limited portion of the diversity encountered in practical forecasting applications.

Within such heterogeneous collections, variations in noise levels, missing observations, and seasonal structures imply that a model's observed performance may reflect the specific composition of the data mix rather than its general algorithmic efficacy (Hyndman and Athanasopoulos, 2021; Godahewa et al., 2021). Empirical evidence confirms that performance rankings derived from these practices are highly unstable (Roque et al., 2025). Sampling different subsets from a single large-scale archive has been shown to alter the perceived ranking of methods, potentially misidentifying certain algorithms as state-of-the-art due to peculiarities of the subset. This sensitivity is often compounded by informal selection practices that overrepresent simpler, well-behaved datasets while neglecting more complex sequences. Additionally, a persistent gap remains between curated historical benchmarks and real-world environments, as the former often mask the structural changes and distribution shifts inherent in live forecasting applications (Koh et al., 2021).

Beyond dataset composition, the interaction between validation protocols and data selection introduces further complexity. While temporally consistent protocols, such as rolling-origin evaluation, provide a structured approach to performance estimation, their outputs remain strictly conditional on the underlying data quality. Empirical evidence suggests that even when validation protocols are rigorously standardised, performance rankings fluctuate significantly depending on the specific subsets and temporal segments selected (Cerqueira et al., 2020; Roque et al., 2025). This dependency indicates that robust validation protocols, though necessary, are insufficient to mitigate the systemic biases introduced by the selection of non-representative datasets. This limitation underscores the critical need for formal, statistically grounded auditing frameworks applied prior to model evaluation.

2.5 Feature-Based Representation of Time Series

The direct comparison of raw time series is unsuitable for dataset-level analysis, as it is highly sensitive to scale, phase, sampling rate, and sequence length. Furthermore, raw representations do not explicitly encode higher-level structural characteristics, such as the strength of seasonality, trend complexity, or the noise regime. A widely adopted alternative is to represent each series as a vector of descriptive features capturing statistical, temporal, spectral, and complexity-related properties. Under this paradigm, a time series dataset is mapped to an empirical distribution in a common feature space, enabling robust comparisons across diverse collections using standard statistical tools.

Feature-based representations differ primarily in their trade-off between dimensionality and interpretability. High-dimensional approaches prioritise domain-independent coverage by extracting a massive range of features across multiple mathematical categories. The *hctsa* framework exemplifies this strategy, cataloguing thousands of candidate metrics for exhaustive time series characterisation (Fulcher and Jones, 2017). Similarly, libraries such as *tsfresh* (Christ et al., 2018) and TSFEL (Barandas et al., 2020) provide large collections of generic features to minimise domain assumptions. While these high-dimensional representations are powerful for exploratory

data analysis, they can complicate dataset auditing. Feature redundancy, multicollinearity, and the mathematical instability of distance metrics in high-dimensional spaces (the curse of dimensionality) often obscure meaningful structural differences between datasets.

Compact alternatives address these challenges by prioritising a smaller, carefully curated feature set. For instance, *catch22* isolates a subset of canonical features specifically selected for their high discriminative power and computational efficiency (Lubba et al., 2019). In the specific context of forecasting, the *tsfeatures* library (Hyndman et al., 2019) isolates aspects directly relevant to predictive modelling, such as trend strength, seasonality, and spectral entropy. The direct link between these temporal features and model performance was empirically validated by the FFORMA framework (Montero-Manso et al., 2020), which demonstrated that forecasting accuracy can be predicted from time series features alone. This pivotal finding establishes that this feature space captures the exact structural properties that drive model selection.

In auditing scenarios, relying on these compact, domain-relevant representations is crucial. They not only stabilise the estimation of distributional distances but also provide clear, diagnostic interpretations of the statistical shifts that occur when a benchmark subset diverges from its reference population.

2.6 Distributional Discrepancy Measures

Assessing whether a benchmark dataset is representative of a target population can be framed as a problem of quantifying distributional discrepancy between two empirical samples. The literature offers several nonparametric approaches for this purpose that operate without strong assumptions about the underlying data-generating processes.

Kernel-based two-sample methods, such as Maximum Mean Discrepancy (MMD), provide a principled framework for assessing whether two empirical samples originate from the same underlying distribution by comparing their mean embeddings in a reproducing kernel Hilbert space (Gretton et al., 2012). Complementarily, Optimal Transport approaches characterise distributional discrepancy through a geometric formulation, measuring the minimal cost required to transform one empirical distribution into another, with Wasserstein distances being among the most widely studied instances (Peyré and Cuturi, 2019; Flamary et al., 2021). These global measures are sensitive to differences in the overall shape and spread of distributions, but do not capture local structural failures such as gaps in coverage or concentrated redundancy within a subset.

To address this, neighbourhood-based approaches analyse local distributional structure through proximity relationships in feature space, enabling the detection of finer-grained representativeness failures — such as a subset that is internally consistent but fails to cover peripheral regions of the reference distribution. Metrics in this family, originally developed in the context of generative model evaluation (Sajjadi et al., 2018), decompose representativeness into complementary dimensions of fidelity and coverage, providing richer diagnostic information than any single scalar distance.

Building on these foundations, the *pyMDMA* framework (Façoco et al., 2025; AICOS, 2024) integrates both global and neighbourhood-based metrics into a unified auditing tool, originally conceived for synthetic data quality assessment. Its multidimensional structure — combining kernel-based distances, optimal transport costs, and local manifold-based metrics — makes it directly applicable to the representativeness assessment of time series benchmark collections, as employed in this dissertation.

2.7 Summary

Forecasting research is shaped not only by model design but also by the statistical properties of the data and the evaluation protocols under which empirical claims are established. This chapter examined these dimensions with particular emphasis on benchmarking practices and dataset selection.

Time series data exhibit structural properties that fundamentally differentiate forecasting from standard predictive modelling (Hyndman and Athanasopoulos, 2021; Box et al., 2015). These properties complicate both model evaluation and cross-dataset comparison (Bergmeir and Benítez, 2012; Bergmeir et al., 2018), motivating careful consideration of metrics (Hyndman and Koehler, 2006), validation protocols (Cerqueira et al., 2020), and benchmark composition. While substantial progress has been made in defining appropriate evaluation methodologies, empirical results remain inherently conditional on the datasets to which they are applied.

Despite the widespread adoption of large-scale repositories and standardised evaluation protocols (Godahewa et al., 2021; Makridakis et al., 2020; Aksu et al., 2024), implicit assumptions regarding dataset representativeness persist. Empirical evidence demonstrates that model rankings and comparative conclusions can be highly sensitive to benchmark composition (Roque et al., 2025), particularly when dataset selection is driven by convenience or historical precedent rather than principled criteria (Koh et al., 2021). This sensitivity undermines the generalisability of reported results and raises concerns regarding the external validity of empirical progress in forecasting research.

Several strategies have been proposed to mitigate these limitations, including expanding benchmark repositories (Makridakis et al., 2022; Godahewa et al., 2021), protocol-level standardisation (Qiu et al., 2024; Aksu et al., 2024; Shchur et al., 2025), and synthetic data generation (Kang et al., 2020). While these approaches improve accessibility, reproducibility, and experimental coverage, they do not provide explicit mechanisms to assess or guarantee the representativeness of a selected benchmark subset with respect to a reference population or a user-defined target domain.

Finally, the chapter reviewed methodological foundations relevant to dataset-level auditing and structuring. Feature-based representations enable heterogeneous time series to be embedded in a common space (Fulcher and Jones, 2017; Lubba et al., 2019; Barandas et al., 2020; Hyndman et al., 2019), where distributional discrepancy can be quantified using mathematically grounded distance measures (Gretton et al., 2012; Peyré and Cuturi, 2019; Façoco et al., 2025). Neighbourhood-based analysis further supports local similarity assessment and controlled subset

construction (Sajjadi et al., 2018). Despite the maturity of these individual components, their integration into a unified *a priori* framework for auditing benchmark composition remains absent in the forecasting literature.

A clear gap exists between current benchmarking practices and the need for principled, dataset-centric methodologies that explicitly quantify representativeness and support the construction of robust evaluation subsets. Addressing this gap constitutes the exact focus of the methodological framework introduced in the following chapter.

Chapter 3

Methodology

3.1 Introduction

This chapter presents the framework, a general methodology for quantifying the representativeness of a benchmark dataset with respect to a reference corpus of time series. The framework operates entirely in feature space: each series is mapped to a vector of interpretable structural characteristics, and representativeness is measured by the similarity between the candidate dataset's feature distributions and those of the reference corpus. This design makes the framework dataset-agnostic — it can be applied to any collection of time series, at any frequency, from any domain. The central premise is that a benchmark poorly covering the structural diversity of the reference corpus will produce model rankings that are unreliable — valid locally but unlikely to generalise beyond the specific characteristics of that dataset.

Figure 3.1 presents a high-level overview of the framework. The methodology proceeds through three conceptual components, each described in the sections that follow: feature extraction and representation (Section 3.2), representativeness auditing (Section 3.3) and rank stability measurement (Section 3.4)

3.2 Feature Extraction and Representation

The first step in the framework transforms each time series into a structured numerical representation. Each series Y_i is mapped to a fixed-length feature vector $\mathbf{z}_i \in \mathbb{R}^F$, where F denotes the number of extracted features. The resulting matrix $\mathbf{Z} \in \mathbb{R}^{N \times F}$ encodes the structural properties of an entire dataset of N series in a form amenable to distributional analysis.

The motivation for this transformation lies in the structural heterogeneity of time series data. Series drawn from different domains tend to occupy distinct regions of the feature space, and a dataset that covers only a narrow portion of that space may yield model rankings that hold locally but fail to generalise. Instability in rankings across datasets, under this view, reflects genuine diversity in the underlying data-generating processes rather than model inconsistency. By embedding

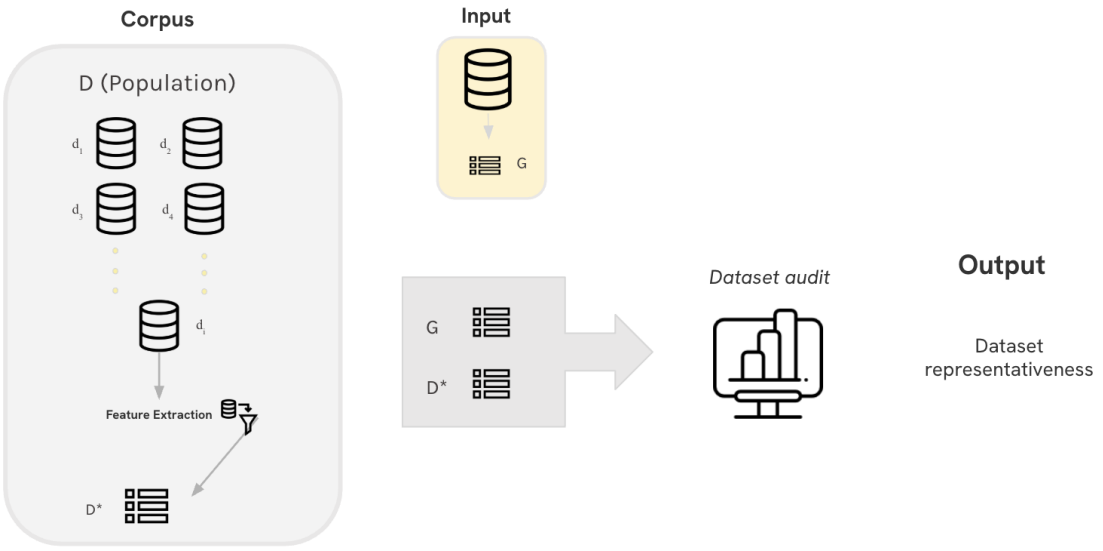


Figure 3.1: High-level overview of the proposed framework. Each dataset $(d_1, d_2, \ldots, d_n)$ in the reference corpus D is mapped via feature extraction into the reference feature set D^*. A candidate dataset G undergoes the same transformation and is compared against D^* in the audit step, which quantifies its representativeness relative to the corpus.

raw series into a common feature space, representativeness becomes an objectively measurable quantity — expressible as the distance and overlap between empirical distributions of $\mathbf{Z}$.

The choice of feature extraction procedure determines which structural properties are made explicit. Among the characteristics commonly employed in the literature, autocorrelation structure, trend and seasonality components, measures of signal complexity, and indicators of stationarity or volatility have proven particularly informative for forecasting model selection (Hyndman and Athanasopoulos, 2021; Montero-Manso et al., 2020).

Normalisation. The representativeness metrics developed in the following section are distance-based. Without normalisation, features with larger numerical ranges would dominate distance calculations irrespective of their discriminative relevance, introducing a measurement artefact rather than a structural signal. To prevent this, $\mathbf{Z}$ has to be normalised through a stage procedure that would map all the features into a common bounded interval. This normalisation step should be applied once to the full reference corpus and then applied without modification to any candidate dataset under evaluation, ensuring that all feature vectors occupy a comparable space.

3.3 Representativeness Auditing

Given a candidate benchmark dataset and a reference corpus, the auditing component quantifies how well the candidate's feature distribution aligns with and covers the reference. The comparison between two datasets is framed as a data quality problem: the candidate dataset is treated as an analogue of synthetic data, and the reference corpus as the real-world distribution it should reflect. This framing allows the direct adoption of metrics developed for synthetic data quality assessment,

which are well-suited to quantifying representativeness as a distributional property. Representativeness is treated as a multidimensional concept, decomposed into three complementary auditing axes.

Fidelity measures whether the candidate dataset's series occupy regions of feature space that the reference also occupies. A high-fidelity dataset provides forecasting models with structural conditions that the reference distribution considers typical. It answers the question: *does the candidate stay within the bounds of the reference?*

Diversity measures whether the candidate dataset spans the full variability of the reference corpus. A dataset with high coverage exercises models across the complete range of structural conditions present in the reference. It answers the complementary question: *does the candidate reach all regions of the reference?*

Distributional distance (Fidelity and Diversity) provides global measures of divergence between the two feature distributions, independent of local neighbourhood structure. These metrics capture the overall structural gap between distributions rather than local alignment or span.

Fidelity and Diversity are conceptually independent: a dataset can have high fidelity — its series are structurally plausible given the reference — while having low Diversity, sampling only a narrow subset of the reference space. Both dimensions are necessary for a complete representativeness profile.

The three auditing axes are operationalised through six metrics applied to the normalised feature matrices described in Section 3.2.

Fidelity: Improved Precision. Improved Precision measures the fraction of series in the candidate dataset whose feature vectors fall within the support of the reference distribution. Support is estimated locally via k-nearest neighbour manifolds: for each reference series x_i, a neighbourhood of radius $r_i = \|x_i - x_i^{(k)}\|$ is defined, where $x_i^{(k)}$ is its k-th nearest neighbour in the reference. A candidate series I_j is considered to fall within the reference support if it lies within the neighbourhood of at least one reference series:

$$\text{Precision} = \frac{1}{|I|} \sum_{I_j \in I} \mathbb{I}(\exists x_i \in X \text{ such that } \|I_j - x_i\| \leq r_i)$$

Figure 3.2 illustrates this computation geometrically.

Fidelity: Density. Density is a density-weighted generalisation of precision. Rather than a binary check, it counts how many reference neighbourhoods each candidate series falls within, measuring whether the candidate not only lies within the reference support but matches its local concentration:

$$\text{Density} = \frac{1}{|I|} \sum_{I_j \in I} \left(\frac{1}{k} \sum_{x_i \in X} \mathbb{I}\big(\|I_j - x_i\| \leq r_i\big) \right)$$

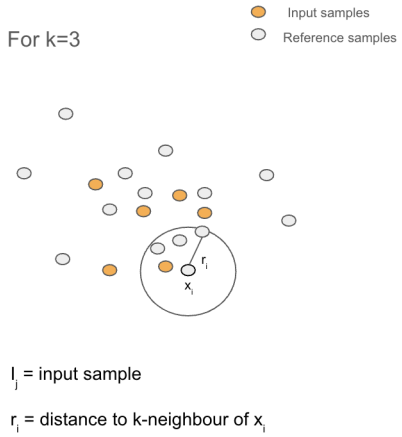


Figure 3.2: Illustration of the Improved Precision metric for $k = 3$. A candidate point (cross) is counted as precise if it falls within the neighbourhood of at least one reference point (circle).

A candidate dataset with density greater than 1 concentrates more tightly in feature space than the reference itself — a signal of a homogeneous, potentially unrepresentative dataset. Figure 3.3 illustrates this metric.

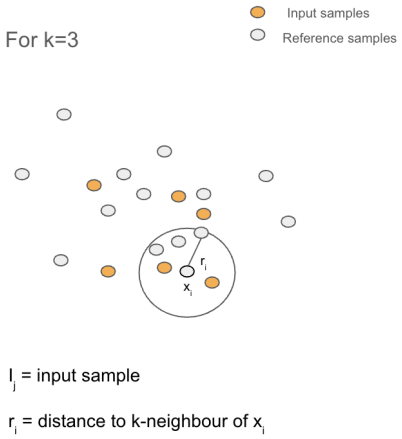


Figure 3.3: Illustration of the Density metric for $k = 3$.

Diversity: Improved Recall. Improved Recall reverses the direction of the precision check: it measures the fraction of the reference series that the candidate's manifold covers. For each reference series x_i, recall checks whether it falls within the neighbourhood of at least one candidate series, where neighbourhoods are now defined around candidate points:

$$\text{Recall} = \frac{1}{|X|} \sum_{x_i \in X} \mathbb{I}\left(\exists I_j \in I \text{ such that } \|x_i - I_j\| \leq r_j\right)$$

Low recall indicates that large regions of the reference feature space are not represented in the candidate dataset. Figure 3.4 illustrates this metric.

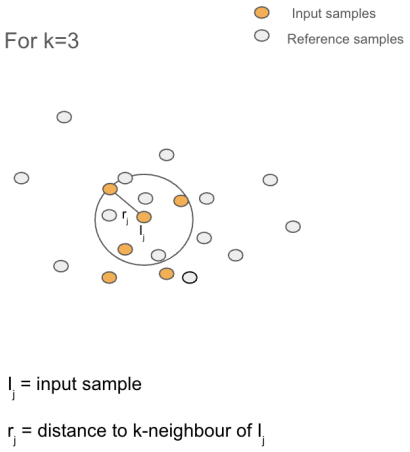


Figure 3.4: Illustration of the Improved Recall metric for $k = 3$. A reference point (circle) is counted as covered if it falls within the neighbourhood of at least one candidate point (cross).

Diversity: Coverage. The Coverage metric is a threshold-based variant of recall that uses reference neighbourhoods rather than candidate neighbourhoods. For each reference series x_i, it checks whether the nearest candidate series falls within x_i's own neighbourhood radius:

$$\text{Coverage} = \frac{1}{|X|} \sum_{x_i \in X} \mathbb{I}\left(\min_{I_j \in I} \|x_i - I_j\| \leq r_i \right)$$

This provides a stricter coverage criterion: a reference region is considered covered only if a candidate series is sufficiently close to the reference's local scale. Figure 3.5 illustrates this metric.

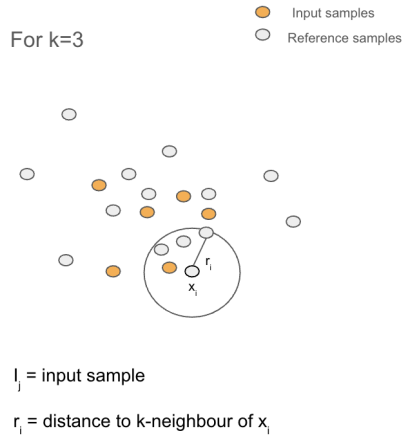


Figure 3.5: Illustration of the Coverage metric for $k = 3$.

Distributional distance: Maximum Mean Discrepancy. Maximum Mean Discrepancy (MMD) is a kernel-based measure of the distance between two distributions (Gretton et al., 2012). It compares the mean embeddings of the two distributions in a reproducing kernel Hilbert space: if the two distributions are identical, their mean embeddings coincide and MMD is zero; the larger the

distributional gap, the larger the MMD value. Unlike the neighbourhood-based metrics above, MMD captures global differences in distributional shape without requiring explicit density estimation or a choice of neighbourhood radius.

Distributional distance: Wasserstein Distance. The Wasserstein distance is the optimal transport cost between two empirical distributions (Peyré and Cuturi, 2019; Flamary et al., 2021). It measures the minimum total effort required to transform one distribution into the other by moving probability mass through the feature space. This makes it sensitive to the geometry of the feature space and to differences in distributional shape that MMD may not fully capture. Together, MMD and Wasserstein provide complementary perspectives on global distributional divergence.

Together, these six metrics constitute a representativeness profile for any candidate dataset. However, whether this profile is meaningful depends on whether it predicts something of practical consequence — namely, whether a less representative benchmark yields less stable model rankings. Establishing this link requires an external criterion, independent of the auditing metrics themselves, against which representativeness can be validated.

3.4 Rank Stability Metrics

Rank stability is not a functional component of the proposed framework, but it plays a central role in its empirical evaluation. The core hypothesis of this dissertation is that benchmark representativeness should predict evaluation reliability — specifically, that datasets with higher representativeness scores yield model rankings that are more consistent with those produced by the broader corpus. To test this, an external criterion independent of the auditing metrics themselves is needed. Rank stability serves this purpose, quantifying the degree to which a candidate dataset's model rankings agree with a consensus derived from the reference corpus. Concretely, it is measured by comparing two rank vectors: the **local ranking** $\mathbf{r} \in \mathbb{R}^M$, produced by evaluating models solely on the candidate dataset, and the **consensus ranking** $\mathbf{c} \in \mathbb{R}^M$, derived from all remaining datasets in the reference corpus, where M denotes the number of models under evaluation. These metrics are used in Chapter 4 to empirically validate the auditing framework.

Five complementary metrics to the framework are used to measure rank stability, each capturing a distinct aspect of ranking agreement. No single metric is sufficient: correlation-based measures capture ordinal agreement but are insensitive to the magnitude of rank differences, while distance-based measures capture magnitude but not direction. Used together, they provide a complete picture of how a dataset's ranking deviates from the consensus.

Spearman's rank correlation (ρ) (Zar, 2005) measures the ordinal agreement between two rank vectors — whether the relative ordering of models is the same across both rankings. It is robust to non-linear relationships and ranges from -1 (perfect inversion) to 1 (perfect agreement).

Kendall's τ (Kendall, 1938) provides an alternative measure of ordinal concordance, based on the number of concordant and discordant pairs of models. It is more conservative than Spearman's ρ and penalises inversions differently, making it a complementary rather than redundant measure.

Weighted rank correlation (Pinto da Costa and Soares, 2005) is a variant that assigns greater importance to agreement at the top of the ranking. This reflects the practical concern of benchmark evaluation: correctly identifying the best-performing model matters more than resolving the ordering among poor performers.

Euclidean distance between rank vectors measures the geometric separation between two rankings, penalising large absolute positional differences more heavily than small ones. Unlike correlation-based metrics, it is sensitive to the magnitude of rank shifts rather than their direction:

$$d_E(\mathbf{r},\mathbf{c}) = \sqrt{\sum_{i=1}^{M}(r_i - c_i)^2} \tag{3.1}$$

Mean rank difference quantifies the average absolute positional displacement across all models, offering an interpretable summary of how much rankings shift on average:

$$d_{\text{mean}}(\mathbf{r},\mathbf{c}) = \frac{1}{M}\sum_{i=1}^{M}|r_i - c_i| \tag{3.2}$$

Maximum rank difference captures the worst-case positional displacement of any individual model. A dataset may yield a ranking that correlates well overall yet still displaces one model dramatically — a scenario that matters in practice when the goal is to identify a specific top performer:

$$d_{\max}(\mathbf{r},\mathbf{c}) = \max_{i\in\{1,\ldots,M\}}|r_i - c_i| \tag{3.3}$$

Chapter 4

Experimental Setup and Results

4.1 Introduction

This chapter presents the empirical instantiation and validation of the Framework described in Chapter 3. The chapter is organised as follows.

Section 4.2 describes the experimental setup in full: the reference corpus constructed for this study, the forecasting models evaluated, the evaluation protocol, and the audit procedure used to validate whether representativeness predicts rank stability.

Sections 4.3–4.5 present the empirical results across the three stages of the analysis: the representativeness audit, the rank stability analysis, and the correlation between the two.

Section 4.6 describes a subsequent experiment designed to probe the limits of the framework. To further investigate the behaviour of the framework under controlled conditions, assessing the extent to which the feature-based representation captures meaningful structural differences that lead to cherry-picking, this experiment replaces the feature-based representation using tsfeatures with a feature-based representation using the ranks feature space, using the per-series model rankings as a direct geometric proxy for cherry-picking susceptibility. This allows the framework's auditing capacity to be tested under conditions where the feature space is guaranteed to carry the relevant signal relative to the cherry-picking behaviour.

Section 4.7 discusses the findings across both experiments in relation to the central hypothesis of the dissertation.

4.2 Experimental Setup

4.2.1 Reference Corpus

The reference corpus is assembled from the Chronos benchmark collection (Ansari et al., 2024), which aggregates time series from multiple domains and sampling frequencies. The collection draws from well-established benchmarks in the forecasting literature: the M-competition series (Makridakis and Hibon, 2000; Makridakis et al., 2020), the Tourism dataset (Athanasopoulos et al.,

2011), the CIF 2016 competition datasets (Štěpnička and Burda, 2017), the FRED-MD macroeconomic database (McCracken and Ng, 2016), and several domain-specific collections distributed through the Monash Forecasting Archive (Godahewa et al., 2021), including the Hospital, NN5, Electricity, Traffic, and Weather datasets. To increase frequency diversity, a subset of datasets is resampled to coarser frequencies via temporal aggregation: flow-type variables are aggregated by summation; state-type variables by mean.

A maximum cap of 5,000 time series per dataset is enforced. This constraint serves a dual purpose: it prevents massive collections from disproportionately dominating the reference distribution, and it ensures that the framework remains computationally tractable on standard hardware, preserving its practical accessibility. For datasets exceeding this limit, a random subsample of 5,000 series is selected. The resulting filtered corpus comprises 38,900 series across 26 datasets and six frequency groups, summarised in Table 4.1. The full unfiltered corpus is provided in Table A.1.

Each series in the reference corpus and in any candidate dataset under evaluation is mapped to a feature vector using the *tsfeatures* R package (Hyndman et al., 2019), following the standardised feature set reviewed by Hyndman and Athanasopoulos (2021) and Montero-Manso et al. (2020). This yields a vector of $F = 42$ features per series, spanning autocorrelation structure, trend and seasonality components, spectral entropy, stationarity tests, and conditional heteroscedasticity statistics. Seasonal periods are passed dynamically based on each dataset's temporal frequency.

Missing value handling. Certain features are undefined for specific series types. For non-seasonal series, all seasonal components — including seasonal strength, peak and trough magnitudes, and seasonal autocorrelation coefficients — are set to zero. For very short series, autocorrelation coefficients at higher lags are likewise set to zero, as their estimates are statistically unreliable. The Hurst exponent defaults to 0.5, corresponding to a random walk baseline (Box et al., 2015), and all conditional heteroscedasticity statistics default to zero in the absence of detectable volatility clustering effects.

Normalisation. The feature matrix is normalised through a two-stage procedure: `RobustScaler` — based on the median and interquartile range — is applied first to reduce the influence of outliers, followed by `MinMaxScaler` that maps all features into $[0, 1]$. Both scalers are fitted once on the full reference corpus and applied without modification to any candidate dataset under evaluation, ensuring all feature vectors occupy a common, comparable space.

4.2.2 Forecasting Models and Evaluation Protocol

A fixed set of six deep learning models is evaluated across all datasets: NHITS (Challu et al., 2023), TiDE (Das et al., 2023), NBEATS (Oreshkin et al., 2020), NLinear (Zeng et al., 2023), MLP, and DLinear (Zeng et al., 2023), all implemented via the NeuralForecast library (Olivares et al., 2022). Deep learning models are selected because such architectures are particularly sensitive to dataset composition and evaluation conditions (Roque et al., 2025), making them well-suited for studying the relationship between benchmark representativeness and rank stability.

Table 4.1: Datasets used in the forecasting and representativeness experiments (filtered corpus).

	Dataset	# Time Series	# Observations	H	Seasonal Period
Yearly	M1	181	4,515	2	1
	M3	645	18,319	3	1
	M4	5,000	185,393	3	1
	Tourism	503	12,592	2	1
Quarterly	M1	203	9,944	2	4
	M3	756	37,004	4	4
	M4	5,000	498,561	4	4
	Tourism	427	42,544	4	4
	CIF 2016	55	2,095	4	4
	FRED-MD	107	26,001	4	4
Monthly	M1	617	55,998	6	12
	M3	1,428	167,562	12	12
	M4	5,000	1,164,797	12	12
	Tourism	366	109,280	12	12
	Hospital	767	64,428	12	12
	CIF 2016	69	7,022	12	12
	FRED-MD	107	77,896	12	12
	Weather	2,989	1,413,594	12	12
Weekly	M4	359	371,579	13	52
	NN5	111	87,801	8	52
	Electricity	321	50,076	8	52
	Traffic	862	90,510	8	52
	Weather	3,010	6,149,904	8	52
Daily	M4	4,227	10,023,836	14	365
	Electricity	321	352,137	30	365
	Traffic	862	630,122	30	365
	Weather	3,010	43,032,000	30	365
Hourly	M4	414	373,372	48	24
	Electricity	321	8,443,584	48	24
	Traffic	862	15,122,928	48	24
Total		**38,900**	**88,625,394**	–	–

Rather than strictly targeting the absolute SOTA, these models were chosen because they serve as well-established, computationally efficient baselines in the modern forecasting literature.

Each model undergoes automatic hyperparameter optimisation via 2 trials of random search – a computational constraint acknowledged as a limitation of this study. Models are evaluated using a temporal hold-out strategy, where the final h observations of each series are strictly reserved for testing, and all preceding data is used for training. Performance is measured using the Mean Absolute Scaled Error (MASE):

$$\text{MASE} = \frac{\text{MAE}}{\frac{1}{T-m}\sum_{t=m+1}^{T}|y_t - y_{t-m}|}$$

where m is the seasonal period and T is the length of the training series. MASE is selected for its scale-independence, making it suitable for aggregation across datasets with different magnitudes and units (Hyndman and Koehler, 2006).

4.2.3 Audit and Validation Protocol

As introduced earlier, the framework is validated by testing whether the representativeness metrics defined in Chapter 3 predict rank stability. The central hypothesis is that datasets with lower representativeness produce less stable model rankings — and that a practitioner auditing a benchmark can therefore use the framework's output to anticipate whether that benchmark is likely to yield rankings that generalise.

Representativeness auditing is performed using the *pyMDMA* library (Façoco et al., 2025), which implements the six metrics described in Section 3.3. The library was originally developed for synthetic data quality assessment and is applied here without modification, but with a different interpretation to compare empirical feature distributions.

To operationalise this, the framework is applied using a leave-one-out strategy: when auditing dataset d, it is excluded from the reference corpus and its feature distribution is compared against the remaining datasets. Two audit variants are computed in parallel:

- **Global audit:** the reference is the full corpus excluding d, measuring representativeness with respect to the overall corpus diversity regardless of frequency.
- **Same-frequency audit:** the reference consists only of datasets sharing the same temporal frequency as d, isolating structural differences from frequency-driven distributional shifts.

The global audit aligns with the framework's intended use: it measures how representative a benchmark is of the overall corpus, treating temporal frequency as a legitimate dimension of that representativeness. The same-frequency audit is a control rather than an alternative recommendation — it holds frequency constant to test whether the representativeness–stability relationship survives once this dominant axis of variation is removed.

Rank stability for each dataset is measured by comparing its model ranking — derived from per-series MASE scores — with the consensus ranking across all other datasets under the same

leave-one-out protocol. To account for negligible performance differences, two models are considered tied on a given series if their absolute MASE difference is less than 0.05, and they share the average rank. Dataset-level ranks are obtained by averaging per-series ranks across all series within the dataset.

The relationship between representativeness and rank stability is then assessed by computing Spearman correlations between each of the six representativeness metrics and each of the six rank stability metrics, across all 26 datasets, under both audit variants.

4.3 Overview of Representativeness Metrics

Figure 4.1 presents the distribution of the six representativeness metrics across all 26 datasets under both audit variants. Precision clusters at high values with constrained dispersion under both audit variants, suggesting that most datasets occupy regions locally consistent with the reference space. Density exhibits considerably wider dispersion, particularly under the global audit, reflecting greater heterogeneity in how densely datasets populate the reference space. Recall and Coverage show the lowest medians under the global audit, indicating that individual datasets cover only a fraction of the reference distribution. Both improve under the same-frequency audit, though the gain is substantially more pronounced for recall than for coverage. Among the distance metrics, MMD distance narrows considerably under the same-frequency audit, while Wasserstein distance remains more resistant to this correction.

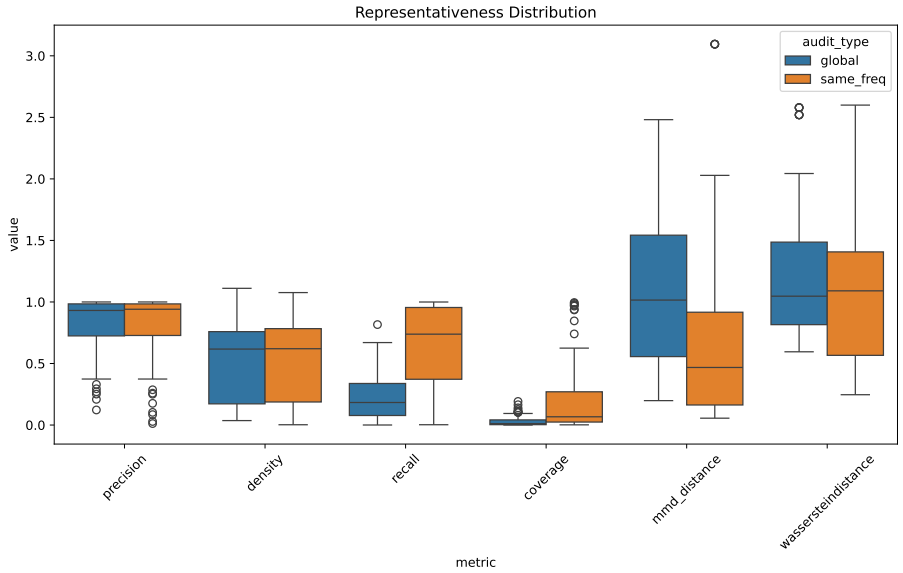


Figure 4.1: Distribution of the six representativeness metrics across all 26 datasets, for the global and same-frequency audit variants. Each box shows the interquartile range across datasets; circles denote outliers. Higher precision, density, recall, and coverage indicate greater representativeness; lower MMD and Wasserstein distances indicate greater similarity to the reference distribution.

The per-metric distributions broken down by dataset and frequency group are provided in Appendix B.2. Taken together, the results presented in this section establish that representativeness

varies substantially across datasets and is not reducible to a single dimension. Some datasets score high across all six metrics, suggesting broad structural alignment with the reference corpus, while others exhibit low coverage or high distributional distance regardless of the audit variant, indicating structural isolation. This variability is a necessary condition for the analysis that follows: without meaningful dispersion in representativeness scores across datasets, it would not be possible to assess whether more representative benchmarks yield more stable model rankings.

4.4 Rank Stability Results

4.4.1 Forecasting Results and Cherry-Picking Validation

In an ideal setting, evaluating which forecasting model performs best would be straightforward: given unlimited computational resources, data access, and implementation effort, one could simply run every model on the entire population of possible time series and unambiguously identify the best performer. In practice, however, these constraints make exhaustive evaluation infeasible. Practitioners instead rely on a sample of benchmark datasets — a necessarily finite collection that stands in for the broader population of series a model may encounter in deployment.

This substitution raises a representativeness concern. If the benchmark collection is not a faithful sample of the population — whether through deliberate curation, domain bias, or inadvertent cherry-picking — the model rankings it produces may not generalise beyond it. A model that ranks first on a narrow or atypical benchmark may perform very differently when deployed on series with different structural characteristics.

Rank stability across datasets is the empirical lens through which this concern is examined. If rankings are consistent across a diverse set of benchmarks, this suggests that the underlying performance ordering is robust and not an artefact of any particular dataset's composition. The empirical results confirm substantial variance in rankings across the corpus, establishing that the collection is sufficiently diverse for stability differences to be meaningful — and for their relationship with representativeness to be informative.

Figure 4.2 illustrates the average rank of each model per dataset, computed from per-series MASE scores. The results reveal significant heterogeneity in model performance: no single architecture establishes absolute dominance or consistent inferiority across the entire corpus. While TiDE frequently ranks lowest, demonstrating the weakest relative performance, its placement remains dataset-dependent. Conversely, the top-ranking positions alternate predominantly between NHITS and N-BEATS, whereas the MLP baseline shows high variance, fluctuating between mean ranks of 2.7 and 5.1 across datasets.

The methodological implications of this instability are highlighted in Figure 4.3, which identifies the top-performing model for each dataset. The distribution of the best model within the group is highly bifurcated, with NHITS ranking first across 13 datasets and N-BEATS across 12. This empirical parity clearly illustrates the cherry-picking vulnerability documented in the literature (Roque et al., 2025). By selectively sampling from this corpus, an evaluator could construct

disjoint benchmark subsets that systematically favour either NHITS or N-BEATS, or even other models, in specific datasets, yielding completely contradictory scientific conclusions about general algorithmic superiority. This baseline instability confirms that the constructed corpus successfully replicates the volatility observed in real-world benchmarks, providing a robust foundation for the subsequent representativeness-stability analysis.

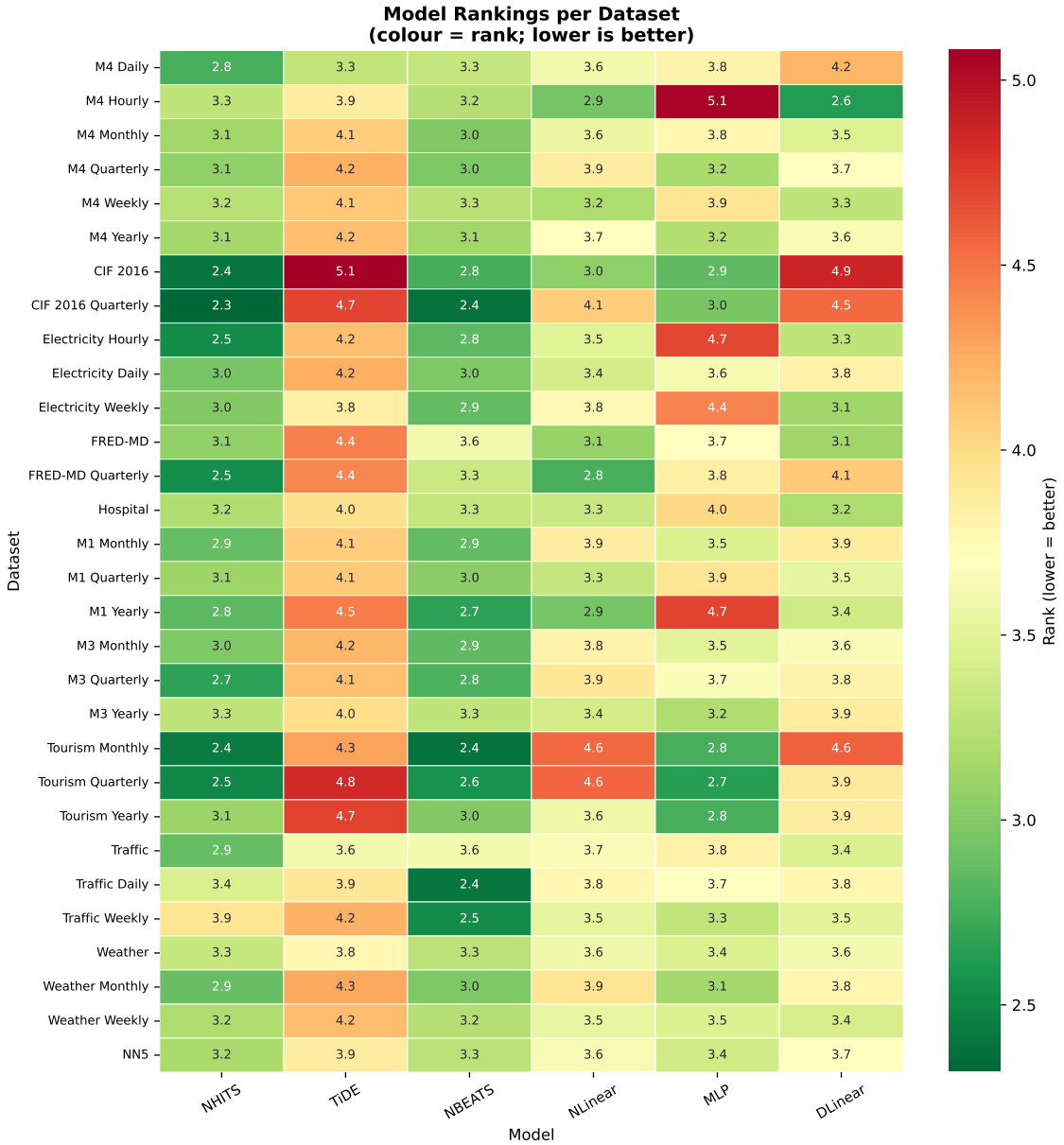


Figure 4.2: Average rank of each model per dataset, derived from per-series MASE scores. Lower rank (greener) indicates better performance. The colour variation across rows confirms that model rankings are not consistent across datasets.

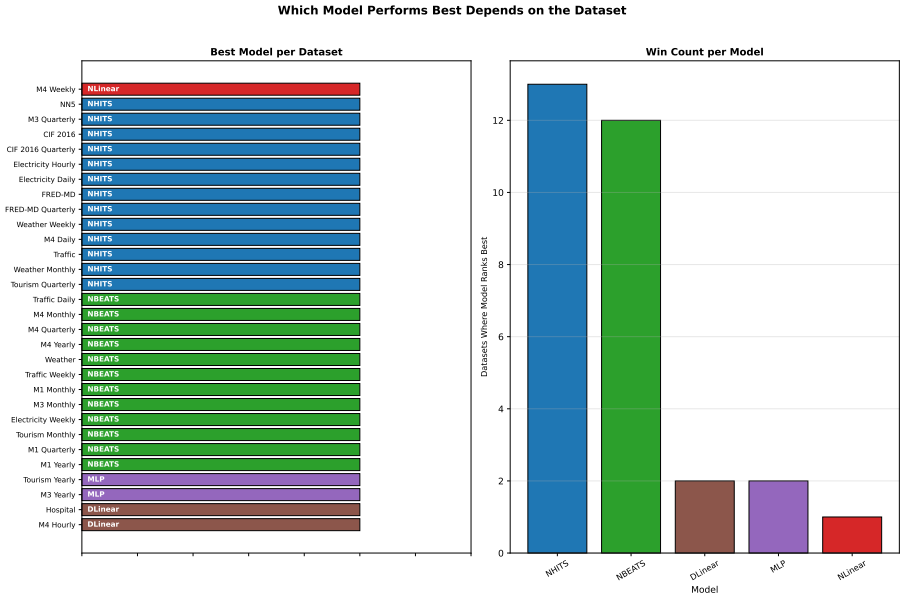


Figure 4.3: Best-performing model per dataset (left) and number of datasets where each model ranks first (right). NHITS and NBEATS together dominate, yet disagree on which datasets each performs best — evidence of dataset-dependent ranking behaviour.

4.4.2 Rank Stability Across Datasets

Figure 4.4 presents the distribution of all six stability metrics across the 26 datasets, evaluated under both the global and same-frequency consensus definitions. The correlation-based measures exhibit strong overall agreement and remain relatively robust across both audit types. Spearman correlation is the highest (median $\sim$0.80), whereas Kendall's τ and weighted rank correlation present more conservative estimates (medians $\sim$0.60). However, the distance and difference-based metrics reveal a distinct structural shift between the two variants. When evaluating stability against a strictly same-frequency consensus, the medians and upper quartiles for Euclidean distance, mean rank difference, and maximum rank difference notably increase. This indicates that while the broad ordinal agreement (correlation) remains intact, model rankings exhibit greater positional volatility and larger displacement magnitudes when compared strictly against their temporal peers rather than the global baseline.

Importantly, the wide interquartile ranges and the presence of long tails in the distance metrics — as well as distinct lower-bound outliers in the correlation metrics — confirm that ranking instability is not a systemic baseline behaviour, but rather a pathology concentrated within a specific subset of datasets. This stark contrast between highly stable benchmarks and severe outliers prompts the central question of this study: whether this empirical instability can be systematically explained by the structural representativeness of the underlying datasets.

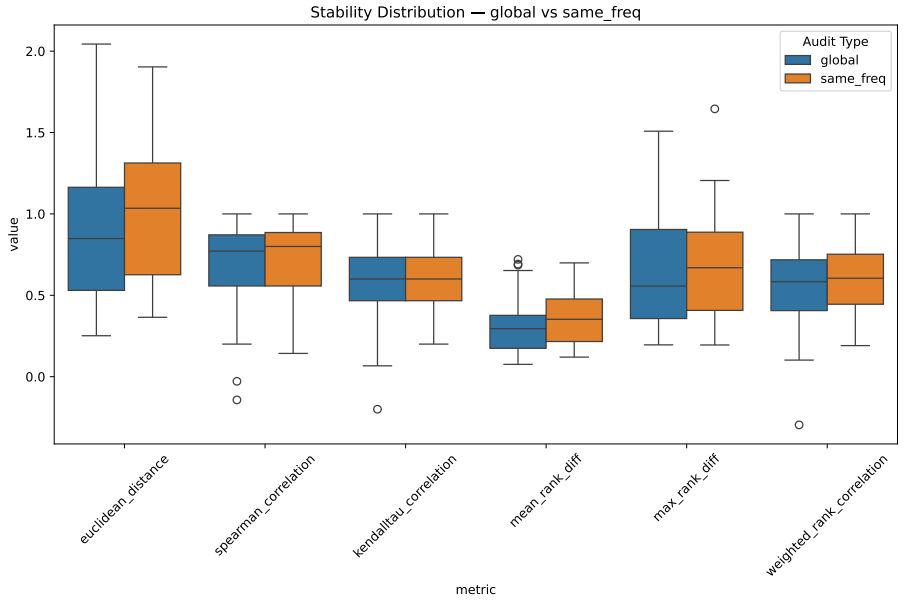


Figure 4.4: Distribution of the six rank stability metrics across all 26 datasets, for the global and same-frequency audit variants. For correlation-based metrics, higher values indicate greater stability; for distance-based metrics, lower values indicate greater stability.

4.5 Representativeness–Stability Correlation

4.5.1 Global Audit

Figure 4.5 presents the Spearman correlation between each representativeness metric and each rank stability metric at $k = 10$ under the global audit. The overall picture is one of weak to moderate relationships — no metric pair exceeds $|\rho| = 0.35$ — and while the strongest signals are consistent with theoretical expectations, some of the weaker relationships exhibit noisy directionality.

Coverage emerges as the most predictive representativeness metric under the global audit, particularly concerning positional volatility. Datasets with higher coverage scores tend to produce more stable rankings by minimising severe displacement: coverage correlates negatively with the mean rank difference ($\rho = -0.35$), the Euclidean distance ($\rho = -0.33$), and the maximum rank difference ($\rho = -0.30$). While its relationship with pure ordinal correlation is positive but weak (e.g., Spearman $\rho = 0.11$), the strong negative correlation with the distance metrics is highly interpretable. A dataset that covers a larger fraction of the reference feature space is more likely to present the models with a representative distribution of forecasting conditions, thereby preventing severe, erratic deviations from the global consensus.

Wasserstein distance shows the next most consistent pattern, exhibiting moderate positive correlations with Euclidean distance ($\rho = 0.31$), mean rank difference ($\rho = 0.29$), and maximum rank difference ($\rho = 0.26$). This confirms that datasets that are more distributionally distant from the reference tend to produce rankings with greater displacement. MMD follows a similar, but slightly weaker, pattern ($\rho = 0.21$–0.28 across the distance-based stability metrics).

Precision, density, and recall show consistently weak correlations across all stability metrics (mostly below $|\rho| = 0.25$) and occasional counter-intuitive directionality. This suggests that pure fidelity to the reference distribution — whether a dataset's series strictly falls within the reference support — is less predictive of rank stability than the coverage and distributional distance metrics. A dataset can have high precision while still occupying a narrow, unrepresentative region of the feature space; it is precisely this structural isolation, captured by distance and coverage metrics, that appears most detrimental to ranking behaviour.

Finally, part of the variance in these global comparisons reflects temporal frequency, which is itself a legitimate component of representativeness but also a dimension one may wish to hold constant when isolating purely structural effects. This motivates the same-frequency control presented next.

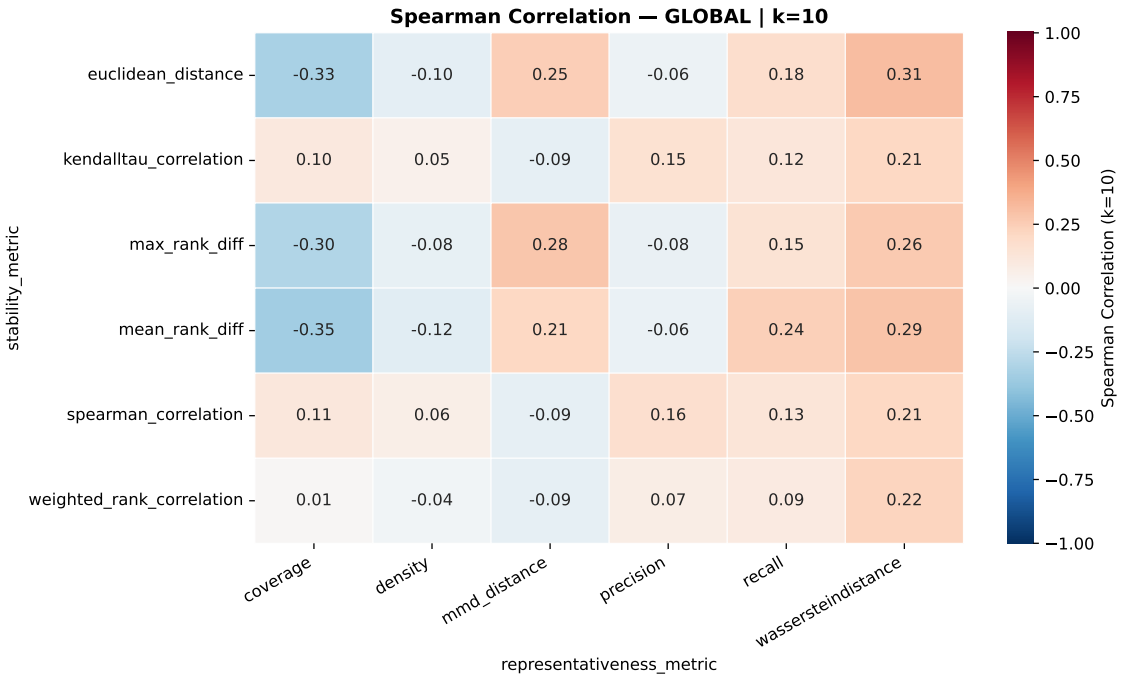


Figure 4.5: Spearman correlation between representativeness metrics and rank stability metrics at $k = 10$, global audit. Positive values indicate that higher representativeness is associated with greater stability; negative values indicate the opposite. For distance-based representativeness metrics (MMD, Wasserstein), higher values indicate lower representativeness, so positive correlations with distance-based stability metrics are consistent with the hypothesis.

4.5.2 Same-Frequency Audit

Figure 4.6 presents the same analysis under the same-frequency audit. Removing the frequency confound alters the pattern of correlations in two highly interpretable ways.

First, the distributional distance metrics (MMD and Wasserstein) now exhibit theoretically consistent relationships with the ordinal correlation stability metrics. Datasets with higher same-frequency MMD tend to produce less stable rankings relative to their frequency peers, as evidenced

by negative correlations with Kendall's τ ($\rho = -0.29$), weighted rank correlation ($\rho = -0.25$), and Spearman correlation ($\rho = -0.21$). Crucially, Wasserstein distance undergoes a sign reversal compared to the global audit (where it previously showed anomalous positive correlations), now correctly aligning with the expected negative direction across all ordinal correlation metrics (e.g., $\rho = -0.19$ with Kendall's τ). Furthermore, Wasserstein maintains positive correlations with the displacement metrics (e.g., $\rho = 0.26$ with Euclidean distance and $\rho = 0.25$ with max rank difference). This indicates that within a clean, frequency-isolated context, structural divergence is a consistent driver of rank instability.

Second, coverage retains strong predictive value for ordinal agreement ($\rho = 0.29$ with Kendall's τ, $\rho = 0.28$ with weighted rank correlation), and precision gains newfound relevance ($\rho = 0.26$ with weighted rank correlation, up from 0.07 globally). This suggests that within a specific frequency group, the pure fidelity of a dataset's distribution to its peers (precision) carries more information about ranking behaviour than it does when assessed against the highly heterogeneous global corpus.

Density and recall remain largely uninformative under both variants, with most correlations hovering around or below $|\rho| = 0.15$.

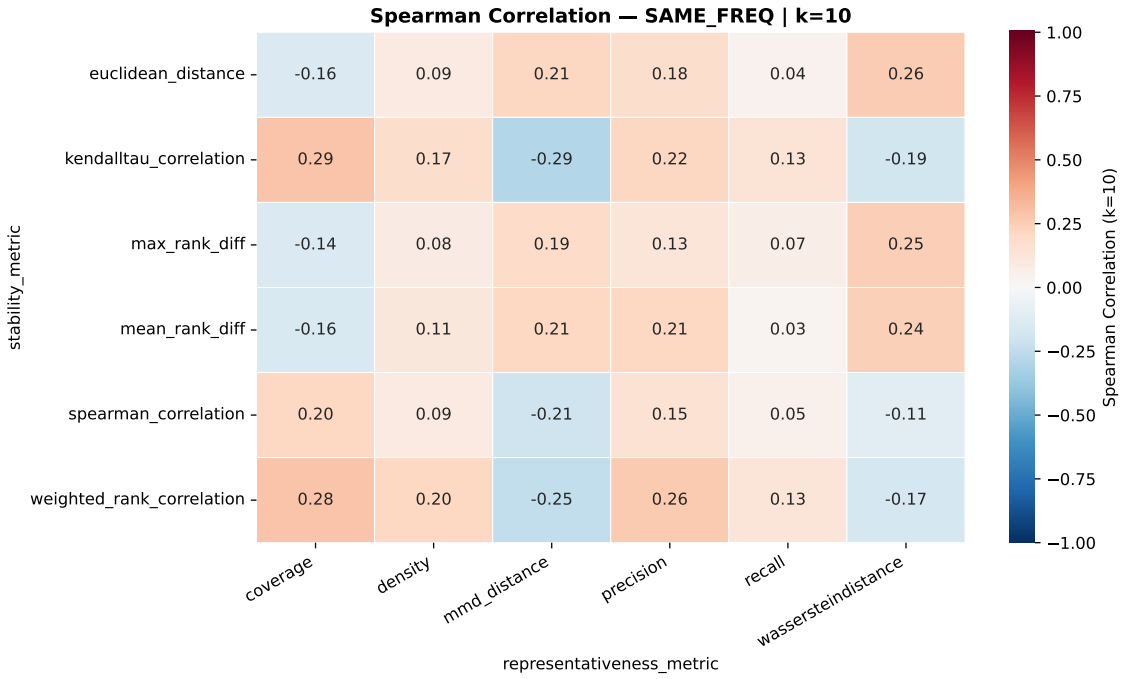


Figure 4.6: Spearman correlation between representativeness metrics and rank stability metrics at $k = 10$, same-frequency audit.

4.6 Rank-Based Feature Analysis

The correlation results reported in Section 4.5 provide only weak to moderate support for the central hypothesis. While the direction of the observed relationships is consistent with theoretical

expectations, the modest magnitudes raise a natural question: do the results reflect a genuine limitation of the auditing framework, or are the tsfeatures-based representations simply insufficiently informative about the factors that drive rank instability?

To disentangle these two explanations, a subsequent experiment replaces the tsfeatures representation with a rank-based feature space. Instead of characterising each series by its structural properties, each series is represented by the vector of per-model MASE ranks it induces. This representation encodes, for each series, which model performs best — the information most directly relevant to cherry-picking susceptibility.

This substitution serves a specific diagnostic purpose: it tests whether the auditing framework can detect cherry-picking when the feature space is guaranteed to contain the relevant signal. If the framework produces strong representativeness–stability correlations in this rank-based space, it provides evidence that the methodology is sound and that the weak results in the tsfeatures experiment are attributable to the choice of feature representation rather than to a fundamental limitation of the auditing approach. Conversely, if the correlations remain weak even with this more informative representation, it would suggest a deeper limitation in the framework's ability to connect distributional representativeness to ranking reliability at this corpus scale.

The rank-based features are constructed as follows: for each series, the six neural forecasting models are evaluated, and their MASE scores are converted into discrete ranks. Each series is thus embedded as a six-dimensional vector of model ranks, serving as a direct geometric proxy for the dataset's internal consensus. The auditing framework is then applied to this new representation using the same leave-one-out protocol and correlation analysis described in Section 4.2. Notably, unlike the `tsfeatures` space — where temporal frequency acts as a severe structural confounder — the rank-based projection maps all series into a uniform, frequency-agnostic combinatorial space. Because the explicit goal of this experiment is purely diagnostic, the audit is performed exclusively at the global level, providing a sufficient, unfiltered test of the framework's ability to detect ranking bias.

4.6.1 Results of the Rank-Based Audit

The results of the rank-based feature experiment provide overwhelming support for the methodological soundness of the auditing framework. Figure 4.7 presents the Spearman correlations between the representativeness and stability metrics under this new representation. Compared to the `tsfeatures` experiment, the correlations exhibit an increase in magnitude, with the strongest relationships shifting from moderate ($|\rho| \approx 0.35$) to strong ($|\rho| > 0.80$).

This substancial increase confirms the diagnostic hypothesis: the framework's geometric engine accurately captures the conditions that lead to ranking instability. The weaker signals observed in Section 4.5 were not a failure of the methodology, but rather a limitation of the features to fully encode algorithmic preference. When the feature space explicitly contains the relevant combinatorial signal, the representativeness audit becomes highly predictive of empirical stability.

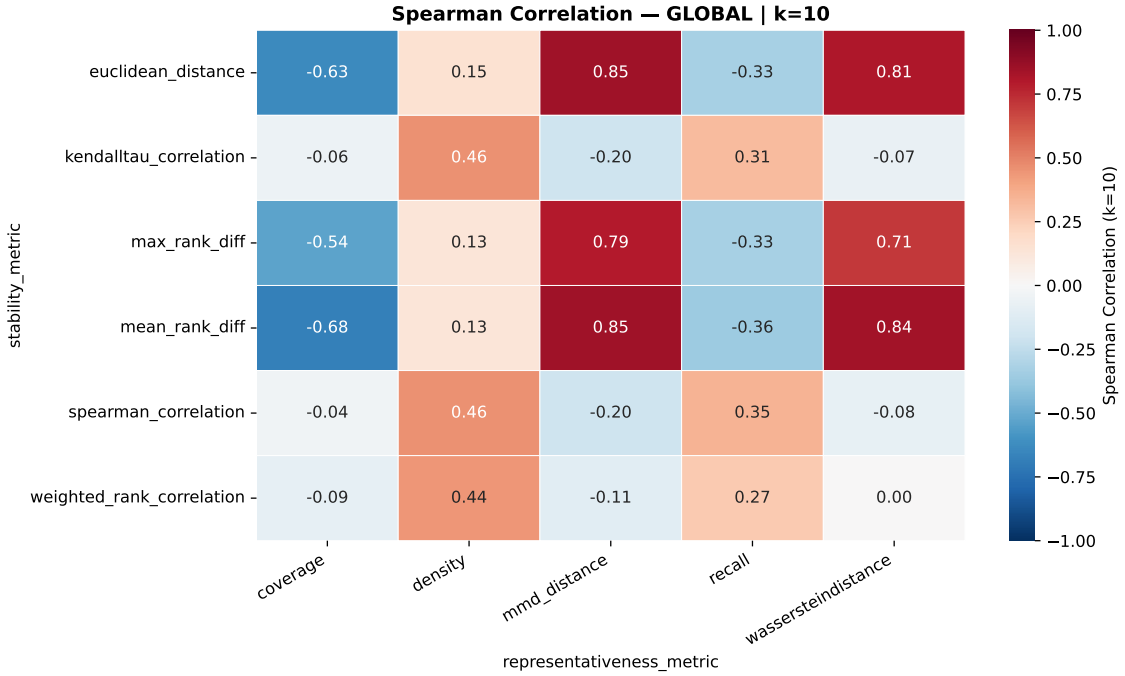


Figure 4.7: Spearman correlation between representativeness and stability metrics using rank-based features (Global Audit, $k = 10$).

The distributional distance metrics emerge as the most powerful predictors of positional volatility. Both MMD and Wasserstein distance show strong positive correlations with Euclidean distance ($\rho = 0.85$ and $\rho = 0.81$, respectively), mean rank difference ($\rho = 0.85$ and $\rho = 0.84$), and max rank difference ($\rho = 0.79$ and $\rho = 0.71$). The scatter plots for MMD and Wasserstein (Figures 4.8 and 4.9) reveal highly monotonic relationships: as a candidate dataset's distribution of model ranks diverges from the global consensus distribution, the magnitude of the resulting ranking displacement increases predictably and severely.

Coverage also demonstrates strong predictive power. It correlates strongly and negatively with mean rank difference ($\rho = -0.68$) and Euclidean distance ($\rho = -0.63$). As illustrated in Figure 4.10, datasets that cover a broader combinatorial space of the global rank permutations are significantly less prone to wild positional swings.

In contrast, boundary-based metrics such as Precision and Recall carry almost no predictive signal in this specific experiment. As detailed in Appendix B.3, these metrics yield flat distributions, with nearly all datasets scoring close to 1.0. This is an expected artefact of the underlying geometry: the rank-based feature space is a bounded, low-dimensional combinatorial space. Given the massive scale of the global reference corpus, virtually all possible rank permutations are present somewhere within the reference support, naturally driving precision and recall toward perfection. In such dense combinatorial spaces, evaluating simple overlap (Precision/Recall) is insufficient; analysing the full distributional mass and its shifts (MMD, Wasserstein, Coverage) becomes the most informative approach.

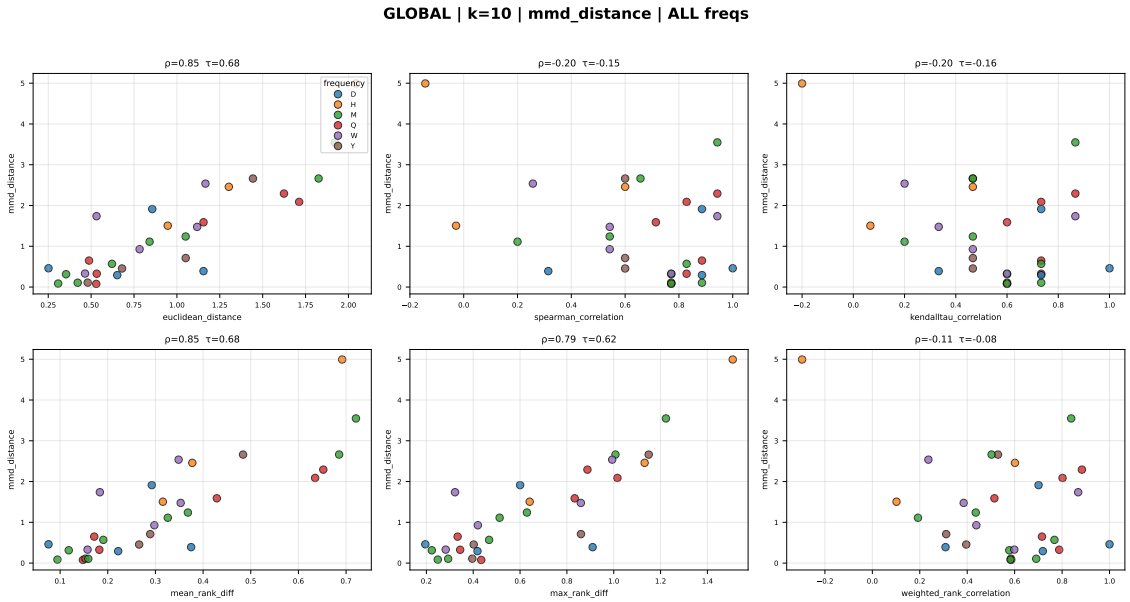


Figure 4.8: Scatter plots detailing the strong monotonic relationship between MMD distance and rank stability metrics in the rank-based feature space.

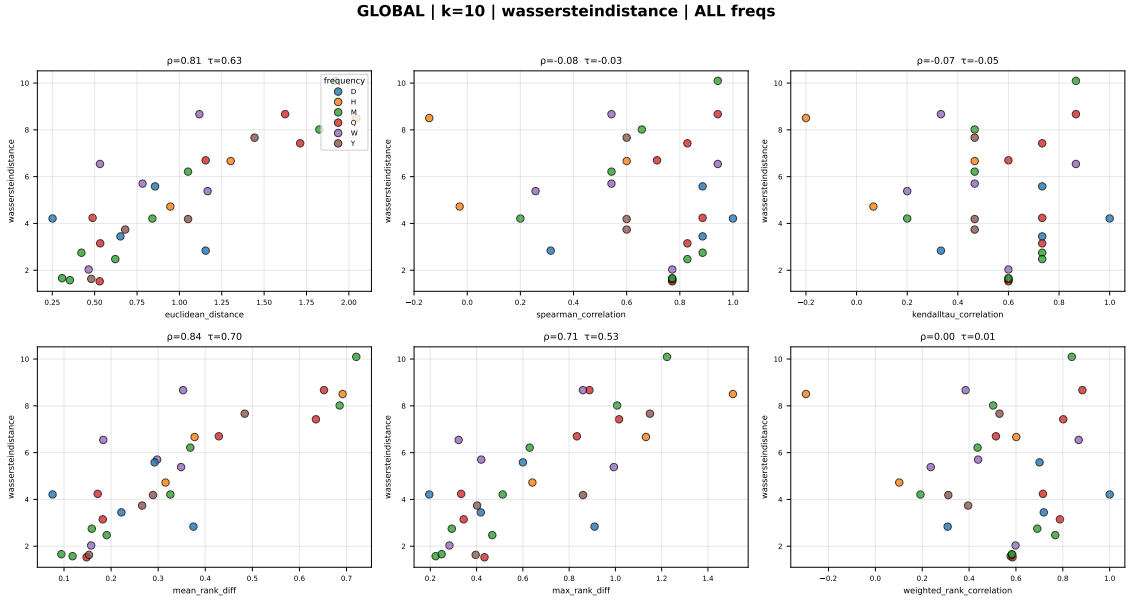


Figure 4.9: Scatter plots demonstrating the relationship between Wasserstein distance and rank stability metrics.

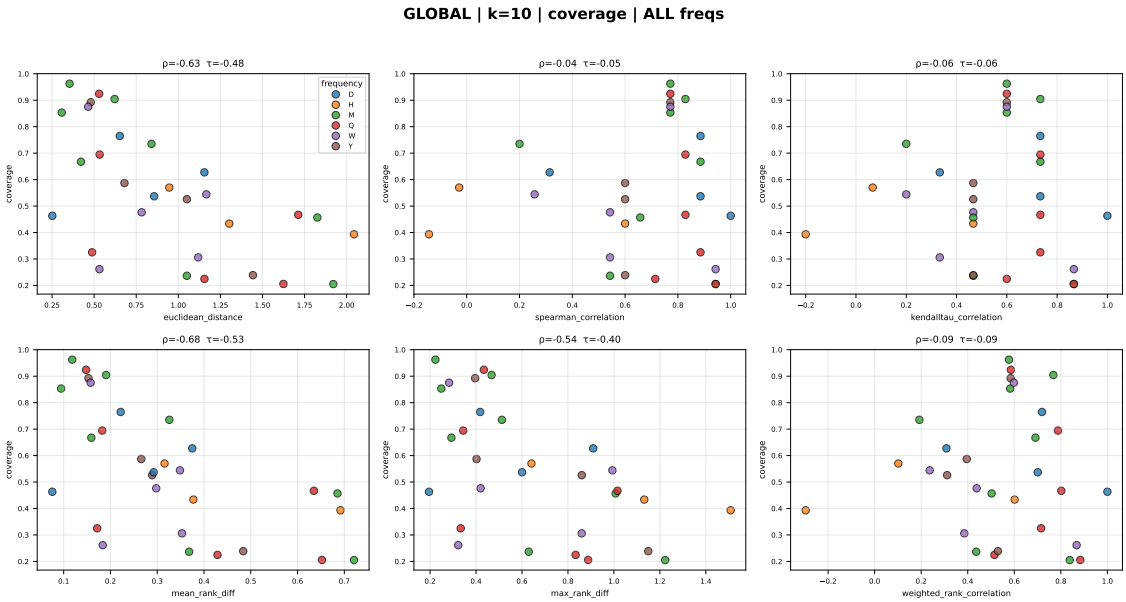


Figure 4.10: Scatter plots demonstrating the negative correlation between Coverage and ranking displacement metrics.

4.7 Discussion of Findings

The empirical results presented in this chapter provide clear validation for the central hypothesis of this thesis: structural representativeness is intrinsically linked to benchmarking stability. By mapping the landscape of the global forecasting corpus, the pyMDMA framework characterised the structural position of each dataset, distinguishing those occupying narrow or atypical regions of the reference space. Across the corpus, datasets with lower coverage and greater distributional distance tended to exhibit higher ranking volatility, consistent with the correlations reported above.

This section revisits the research questions posed in Chapter 1 and examines them against the evidence gathered from the experiments.

The first research question (**RQ1**) addressed the extent to which structural time series features can characterise the composition of a benchmark dataset relative to a broader reference population. The representativeness audit produced discriminative profiles across all 26 datasets, with substantial variation in every metric. Precision, density, recall, coverage, and the two distributional distance metrics each captured distinct and complementary dimensions of structural composition, none of which was reducible to the others. This dispersion confirms that the `tsfeatures` representation carries sufficient structural signal to differentiate benchmarks at varying distances from the reference corpus, though its capacity to capture the inductive biases of modern neural architectures remains bounded.

Complementing this, the second research question (**RQ2**) investigated how effectively feature-based distributional metrics can quantify and detect lack of representativeness within benchmark datasets. Coverage and distributional distance metrics (MMD and Wasserstein) proved the most informative, consistently identifying datasets occupying narrow or atypical regions of the reference

space, while precision, density, and recall provided complementary but less discriminative signals. The same-frequency audit further confirmed that the detected isolation is meaningfully distinct from mere frequency mismatch: structural divergence persisted for certain datasets even when the reference was restricted to same-frequency peers, indicating that the metrics capture genuine compositional differences rather than artefacts of temporal frequency.

With respect to the third research question (**RQ3**), the evidence from Figures 4.5 and 4.6 shows that datasets characterised by low coverage and high distributional distance from the corpus consensus consistently induced higher ranking volatility. Under the global audit, coverage showed the strongest predictive signal for rank displacement; under the same-frequency audit, distributional distance metrics aligned more consistently with ordinal stability measures. That said, the overall magnitude of the predictive association was weak to moderate when standard `tsfeatures` were used, which initially raised questions about the framework's predictive ceiling. The rank-based diagnostic experiment resolved this ambiguity. As shown in Figure 4.7, when the feature space was replaced with a direct combinatorial proxy of algorithmic preference, representativeness–stability correlations amplified to substantially stronger relationships ($|\rho| > 0.80$). This confirms that the auditing framework is mathematically sound and highly effective at detecting susceptibility to dataset cherry-picking. Ultimately, the bottleneck in anticipating model performance does not lie in the methodology of measuring overlap and distance, but rather in the capacity of generic statistical features to fully capture the complex, non-linear patterns exploited by modern neural networks.

Crucially, the analysis also addresses the role of temporal frequency, one of the most salient axes of variation in the feature space. The same-frequency audit was included as a control rather than as a recommended alternative to the global audit: by holding frequency constant, it tests whether the link between structural divergence and rank instability persists once this dominant dimension is removed. It does — confirming that the relationship is structural in nature and not merely an artefact of frequency differences, while the global audit remains the one that reflects representativeness as a practitioner would actually assess it.

Together, these findings validate the necessity and utility of systematic dataset auditing prior to model evaluation. They demonstrate that cherry-picking and benchmarking bias can be geometrically anticipated, provided the right feature space and temporal controls are applied. The broader implications of these results for the machine learning community and the design of future forecasting benchmarks are discussed in the final chapter.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

This dissertation addressed a fundamental vulnerability in the empirical evaluation of machine learning models for time series forecasting: the implicit assumption that benchmark datasets are neutral, universally representative testing grounds. In reality, the composition of a dataset inherently favours certain inductive biases over others. Without a principled way to measure this structural bias, the research community remains susceptible to dataset cherry-picking, where models appear superior simply because they were evaluated on datasets that narrowly align with their architectural strengths.

To address this, the work proposed and validated a comprehensive auditing framework that quantifies the structural representativeness of forecasting datasets relative to a large reference corpus. By projecting time series into a measurable feature space, the framework successfully transitioned dataset assessment from a subjective, ad hoc practice to a formal methodology.

The empirical application of this framework provided evidence supporting the central hypothesis: structural representativeness is related to benchmarking reliability. The answers to the three research sub-questions consolidate this conclusion.

RQ1 asked whether structural time series features can characterise a benchmark's composition relative to a broader reference population. The framework produced discriminative representativeness profiles across all 26 datasets, with substantial variation confirming that they can — and revealing that datasets routinely treated as interchangeable members of a single corpus in fact occupy distinct structural regions, a distinction that is invisible to performance-based evaluation alone. **RQ2** concerned whether distributional metrics can quantify and detect a lack of representativeness; the metrics consistently flagged structurally isolated datasets, and the same-frequency audit confirmed that this isolation is meaningfully distinct from mere frequency mismatch, establishing that the diagnostic responds to genuine structural divergence rather than re-encoding sampling frequency as a proxy. **RQ3** examined whether representativeness carries predictive information about ranking stability; datasets with low coverage and high distributional distance from the corpus consensus consistently induced greater ranking volatility, establishing that it does, though

with weak to moderate effect sizes when standard structural features are used. The rank-based diagnostic experiment traced this ceiling to the expressive limits of classical time series features relative to the inductive biases of modern neural networks, rather than to any limitation of the auditing methodology itself — a distinction that matters because it locates the path to a stronger diagnostic in richer or learned feature representations, not in the auditing approach, which the rank-based results show to be sound.

Taken together, this dissertation demonstrates that the methodology for dataset auditing is fully realised and highly effective. The proposed framework provides a robust, operational tool for detecting benchmarking bias, provided it is supplied with the appropriate structural signal. While standard statistical features currently limit its predictive ceiling, the framework's core engine is fundamentally sound. Moving forward, the primary challenge is no longer how to audit a benchmark, but rather what representations best capture the inductive biases of modern neural networks. By decoupling the auditing methodology from the feature space, this work establishes a foundation for more transparent and trustworthy forecasting research.

5.2 Limitations

Several limitations constrain the conclusions that can be drawn from this study.

The corpus comprises 26 datasets and 38,900 series, which is sufficient to demonstrate the framework's operation but too small to achieve consistent statistical significance in the correlation analysis. With $n = 26$ data points, effect sizes below $|\rho| = 0.40$ are unlikely to be reliably detected, and the frequency confound further reduces the effective sample size within each group.

The forecasting experiment used 2 trials of random configuration search per model during hyperparameter optimisation — a computational constraint that limits the quality of model fits and may introduce noise into the rank estimates. A more thorough hyperparameter search could produce cleaner rankings and stronger stability signals.

Only six deep learning models were evaluated. The selection covers a range of architectures — MLP-based, linear, and basis-expansion — but excludes statistical models, hybrid approaches, and foundation models. The relationship between representativeness and rank stability may differ for other model families, particularly those that rely on different inductive biases.

5.3 Further Work

The framework established in this dissertation provides a robust methodological foundation for dataset auditing and opens several promising avenues for future research, particularly regarding the evolution of feature representations and the expansion of the scope of benchmarking.

The most immediate priority is to explore richer feature spaces. Since the framework's geometric engine has been proven sound, its predictive ceiling is now bounded only by the expressiveness of the input representations.

A crucial direction is the deliberate discovery and design of *meta-features* explicitly engineered to capture a dataset's algorithmic susceptibility. Since standard statistical properties often fail to capture the nuances that drive model rankings, future research should focus on extracting meta-features that serve as direct proxies for forecasting difficulty and predictability. As one concrete possibility, the community could adapt *landmarking* techniques from the broader meta-learning literature to time series forecasting (Pfahringer et al., 2000). This would provide a highly efficient meta-feature vector that explicitly captures the ranking signal, mimicking the success of the rank-based diagnostic experiment without the prohibitive computational cost of training deep neural networks.

5.4 Impact on Sustainable Development Goals

The work presented in this dissertation primarily advances rigorous, reproducible scientific methodology in machine learning and forecasting research. While not directly targeting social or environmental outcomes, the framework developed here has meaningful indirect implications for several United Nations Sustainable Development Goals (SDGs), through the domains in which time series forecasting is applied and through the broader principles of open and trustworthy science.

SDG 3 — Good Health and Well-Being. Healthcare forecasting — including epidemic modelling, patient demand prediction, and resource allocation — relies increasingly on machine learning methods (Chimmula and Zhang, 2020; Kaushik et al., 2020). The validity of these models in clinical or public health contexts is critically sensitive to the data used to evaluate them. A methodology that detects unrepresentative benchmarks reduces the risk of deploying models whose reported performance does not generalise to real-world patient populations, contributing to more reliable health data science.

SDG 7 — Affordable and Clean Energy. Energy consumption forecasting is one of the most active application domains of time series forecasting, supporting the planning of electricity grids, the integration of renewable sources, and the reduction of energy waste (Deb et al., 2017; Martínez-Álvarez et al., 2011). The reliability of models deployed in these contexts depends directly on the quality of the benchmarks used to validate them. By providing tools to audit whether a benchmark dataset is representative of a target energy domain, the proposed framework contributes to more trustworthy model selection in energy forecasting applications, and by extension to the informed deployment of forecasting in sustainable energy systems.

SDG 9 — Industry, Innovation and Infrastructure. The framework advances the state of empirical evaluation infrastructure in forecasting research, directly supporting Goal 9's emphasis on fostering innovation and building reliable scientific infrastructure. Reproducible and unbiased benchmarking is a prerequisite for meaningful scientific progress: without it, reported advances may reflect dataset artefacts rather than genuine algorithmic improvements. By formalising the

assessment of benchmark representativeness, this work strengthens the foundations on which the broader research community builds.

It is important to note that the contributions of this dissertation are methodological rather than applied. The direct beneficiaries are the forecasting research community and the practitioners who rely on benchmarking results to select models for deployment. The downstream impact on the SDGs is therefore indirect, mediated by the framework's improved reliability of empirical results.

References

Ratnadip Adhikari and Ramesh K Agrawal. An introductory study on time series modeling and forecasting. *arXiv preprint arXiv:1302.6613*, 2013.

Fraunhofer AICOS. pyMDMA: Multimodal data metrics for auditing real and synthetic datasets, 2024. URL https://github.com/fraunhoferportugal/pymdma.

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. GIFT-Eval: A benchmark for general time series forecasting model evaluation. In *NeurIPS 2024 Workshop on Time Series in the Age of Large Models (TSALM)*, 2024. URL https://arxiv.org/abs/2410.10393.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://arxiv.org/pdf/2403.07815https://openreview.net/forum?id=gerNCVqqtR.

George Athanasopoulos, Rob J Hyndman, Haiyan Song, and Doris C Wu. The tourism forecasting competition. *International Journal of Forecasting*, 27(3):822–844, 2011.

Marília Barandas, Duarte Folgado, Letícia Fernandes, Sara Santos, Mariana Abreu, Patrícia Bota, Hui Liu, Tanja Schultz, and Hugo Gamboa. Tsfel: Time series feature extraction library. *SoftwareX*, 11:100456, 1 2020. ISSN 2352-7110. doi: 10.1016/J.SOFTX.2020.100456.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

Christoph Bergmeir and José M. Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 5 2012. doi: 10.1016/j.ins.2011.12.028.

Christoph Bergmeir, Rob J. Hyndman, and Bonsoo Koo. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83, 2018. doi: 10.1016/j.csda.2017.11.003.

Casper Solheim Bojer and Jens Peder Meldgaard. Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*, 37(2):587–603, 4 2021. doi: 10.1016/j.ijforecast.2020.07.007.

Gianluca Bontempi, Souhaib Ben Taieb, and Yann-Aël Le Borgne. Machine learning strategies for time series forecasting. In *Business Intelligence*, volume 138 of *Lecture Notes in Business Information Processing*, pages 62–77. Springer, 2013. doi: 10.1007/978-3-642-36318-4_3.

G. E.P. Box and David A. Pierce. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, 65:1509–1526, 1970. ISSN 1537274X. doi: 10.1080/01621459.1970.10481180;PAGE: STRING:ARTICLE/CHAPTER.

George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 5 edition, 2015. ISBN 978-1-118-67502-1.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Vitor Cerqueira, Luis Torgo, and Igor Mozetič. Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, 109(11):1997–2028, 2020. doi: 10.1007/s10994-020-05910-7.

Cristian Challu, Kin G Olivares, Boris N Oreshkin, Federico Garza Ramirez, Max Mergenthaler Canseco, and Artur Dubrawski. Nhits: Neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 6989–6997, 2023.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. Association for Computing Machinery, 2016. doi: 10.1145/2939672.2939785.

Vinay Kumar Reddy Chimmula and Lei Zhang. Time series forecasting of covid-19 transmission in canada using lstm networks. *Chaos, Solitons & Fractals*, 135:109864, 6 2020. ISSN 0960-0779. doi: 10.1016/J.CHAOS.2020.109864.

Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W. Kempa-Liehr. Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package). *Neurocomputing*, 307:72–77, 9 2018. ISSN 0925-2312. doi: 10.1016/J.NEUCOM.2018.03.067.

Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with tide: Time-series dense encoder. *arXiv preprint arXiv:2304.08424*, 2023.

Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Annh Ratanamahatana, and Eamonn Keogh. The ucr time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6:1293–1305, 11 2019. ISSN 23299274. doi: 10.1109/JAS.2019.1911747.

Chirag Deb, Fan Zhang, Junjing Yang, Siew Eang Lee, and Kwok Wei Shah. A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews*, 74:902–924, 7 2017. ISSN 1364-0321. doi: 10.1016/J.RSER.2017.02.085.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Ivo S. Façoco, Joana Rebelo, Pedro Matias, Nuno Bento, Ana C. Morgado, Ana Sampaio, Luís Rosado, and Marília Barandas. pyMDMA: Multimodal data metrics for auditing real and synthetic datasets. *SoftwareX*, 31:102256, 9 2025. doi: 10.1016/j.softx.2025.102256.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T. H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *Journal of Machine Learning Research*, 22 (78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Ben D. Fulcher and Nick S. Jones. hctsa: A computational framework for automated time-series phenotyping using massive feature extraction. *Cell Systems*, 5:531.e3, 11 2017. ISSN 24054720. doi: 10.1016/j.cels.2017.10.001.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*. Neural information processing systems foundation, 5 2021. URL https://arxiv.org/pdf/2105.06643.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 3 2012. doi: 10.5555/2188385.2188410.

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

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9: 1735–1780, 11 1997. ISSN 08997667. doi: 10.1162/NECO.1997.9.8.1735.

Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia, 3 edition, 2021. URL https://otexts.com/fpp3/.

Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006. doi: 10.1016/j.ijforecast.2006.03.001.

Rob J. Hyndman, Anne B. Koehler, J. Keith Ord, and Ralph D. Snyder. *Forecasting with Exponential Smoothing: The State Space Approach*. Springer, 2008. doi: 10.1007/978-3-540-71918-2.

Rob J Hyndman, Earo Wang, and Nikolay Laptev. tsfeatures: Time series feature extraction. *R package version 1.1*, 2019. URL https://CRAN.R-project.org/package=tsfeatures.

Tim Januschowski, Jan Gasthaus, Yuyang Wang, David Salinas, Valentin Flunkert, Michael Bohlke-Schneider, and Laurent Callot. Criteria for classifying forecasting methods. *international Journal of forecasting*, 36(1):167–177, 2020.

Tim Januschowski, Yuyang Wang, Kari Torkkola, Timo Erkkilä, Hilaf Hasson, and Jan Gasthaus. Forecasting with trees. *International Journal of Forecasting*, 38(4):1473–1481, 2022.

Yanfei Kang, Rob J. Hyndman, and Feng Li. Gratis: Generating time series with diverse and controllable characteristics. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 13(4):354–376, 8 2020. doi: 10.1002/sam.11461.

Nataliia Kashpruk, Cezary Piskor-Ignatowicz, and Jerzy Baranowski. Time series prediction in industry 4.0: A comprehensive review and prospects for future advancements. *Applied Sciences 2023, Vol. 13, Page 12374*, 13:12374, 11 2023. ISSN 2076-3417. doi: 10.3390/APP132212374.

Shruti Kaushik, Abhinav Choudhury, Pankaj Kumar Sheron, Nataraj Dasgupta, Sayee Natarajan, Larry A. Pickett, and Varun Dutt. Ai in healthcare: Time-series forecasting using statistical, neural, and ensemble architectures. *Frontiers in Big Data*, 3:475663, 3 2020. ISSN 2624909X. doi: 10.3389/FDATA.2020.00004/BIBTEX.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

Kyoung Jae Kim. Financial time series forecasting using support vector machines. *Neurocomputing*, 55:307–319, 9 2003. ISSN 0925-2312. doi: 10.1016/S0925-2312(03)00372-2.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379, 4 2021. ISSN 1364503X. doi: 10.1098/RSTA.2020.0209/41189.

Carl H. Lubba, Sarab S. Sethi, Philip Knaute, Simon R. Schultz, Ben D. Fulcher, and Nick S. Jones. catch22: Canonical time-series characteristics. *Data Mining and Knowledge Discovery*, 33:1821–1852, 8 2019. ISSN 1573-756X. doi: 10.1007/S10618-019-00647-X.

Spyros Makridakis and Michèle Hibon. The m3-competition: results, conclusions and implications. *International Journal of Forecasting*, 16:451–476, 10 2000. ISSN 0169-2070. doi: 10.1016/S0169-2070(00)00057-1.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36:54–74, 1 2020. ISSN 0169-2070. doi: 10.1016/J.IJFORECAST.2019.04.014.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38:1346–1364, 10 2022. ISSN 0169-2070. doi: 10.1016/J.IJFORECAST.2021.11.013.

Francisco Martínez-Álvarez, Alicia Troncoso, José C. Riquelme, and Jesús S. Aguilar Ruiz. Energy time series forecasting based on pattern sequence similarity. *IEEE Transactions on Knowledge and Data Engineering*, 23:1230–1243, 2011. ISSN 10414347. doi: 10.1109/TKDE.2010.227.

Michael W McCracken and Serena Ng. Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589, 2016.

Pablo Montero-Manso and Rob J. Hyndman. Principles and algorithms for forecasting groups of time series: Locality and globality. *International Journal of Forecasting*, 37:1632–1653, 10 2021. ISSN 0169-2070. doi: 10.1016/J.IJFORECAST.2021.03.004.

Pablo Montero-Manso, George Athanasopoulos, Rob J Hyndman, and Thiyanga S Talagala. Fforma: Feature-based forecast model averaging. *International Journal of Forecasting*, 36(1): 86–92, 2020.

Kin G. Olivares, Cristian Challú, Azul Garza, Max Mergenthaler Canseco, and Artur Dubrawski. NeuralForecast: User friendly state-of-the-art neural forecasting models. PyCon Salt Lake City, Utah, US 2022, 2022. URL https://github.com/Nixtla/neuralforecast.

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

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data sciences. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/2200000073.

Bernhard Pfahringer, Hilan Bensusan, and Christophe G Giraud-Carrier. Meta-learning by landmarking various learning algorithms. In *ICML*, pages 743–750, 2000.

Joaquim Pinto da Costa and Carlos Soares. A weighted rank measure of correlation. *Australian & New Zealand Journal of Statistics*, 47(4):515–529, 2005.

Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, Zhenli Sheng, and Bin Yang. TFB: Towards comprehensive and fair benchmarking of time series forecasting methods. *Proceedings of the VLDB Endowment*, 17(9):2363–2377, 2024. doi: 10.14778/3665844.3665863.

Luis Roque, Vitor Cerqueira, Carlos Soares, and Luis Torgo. Cherry-picking in time series forecasting: How to select datasets to make your model shine. In Luís Roque, Vitor Cerqueira, Carlos Soares, and Luís Torgo, editors, *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20192–20199. Association for the Advancement of Artificial Intelligence, 4 2025. ISBN 157735897X. doi: 10.1609/AAAI.V39I19.34224.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36:1181–1191, 7 2020. ISSN 0169-2070. doi: 10.1016/J.IJFORECAST.2019.07.001.

Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. Financial time series forecasting with deep learning : A systematic literature review: 2005–2019. *Applied Soft Computing*, 90:106181, 5 2020. ISSN 1568-4946. doi: 10.1016/J.ASOC.2020.106181.

Oleksandr Shchur, Abdul Fatir Ansari, Caner Turkmen, Lorenzo Stella, Nick Erickson, Pablo Guerron, Michael Bohlke-Schneider, and Yuyang Wang. fev-bench: A realistic benchmark for time series forecasting. *arXiv preprint arXiv:2509.26468*, 2025.

Martin Štěpnička and Michal Burda. On the results and observations of the time series forecasting competition cif 2016. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6. IEEE, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 6778–6786. International Joint Conferences on Artificial Intelligence Organization, 2023. doi: 10.24963/ijcai.2023/759.

Jerrold H Zar. Spearman rank correlation. *Encyclopedia of biostatistics*, 7, 2005.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2023.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

Appendix A

Reference Corpus and Experimental Setup

A.1 Full Reference Corpus

Table A.1: Full reference corpus used for representativeness auditing.

	Dataset	# Time Series	# Observations	H	Seasonal Period
Yearly	M1	181	4,515	2	1
	M3	645	18,319	3	1
	M4	23,000	854,279	3	1
	Tourism	518	12,757	2	1
Quarterly	M1	203	9,944	2	4
	M3	756	37,004	4	4
	M4	24,000	2,406,108	4	4
	Tourism	427	42,544	4	4
	CIF 2016	55	2,095	4	4
	FRED-MD	107	26,001	4	4
Monthly	M1	617	55,998	6	12
	M3	1,428	167,562	12	12
	M4	48,000	11,246,411	12	12
	Tourism	366	109,280	12	12
	Hospital	767	64,428	12	12
	CIF 2016	72	7,108	12	12
	FRED-MD	107	77,896	12	12
	Weather	2,989	1,413,594	12	12
Weekly	M4	359	371,579	13	52
	NN5	111	87,801	8	52
	Electricity	321	50,076	8	52
	Traffic	862	90,510	8	52
	Weather	3,010	6,149,904	8	52
Daily	M4	4,227	10,023,836	14	365
	Electricity	321	352,137	30	365
	Traffic	862	630,122	30	365
	Weather	3,010	43,032,000	30	365
Hourly	M4	414	373,372	48	24
	Electricity	321	8,443,584	48	24
	Traffic	862	15,122,928	48	24
Total		**118,918**	**101,283,692**	–	–

Appendix B

Supplementary Results

B.1 Feature Distributions per Dataset

This appendix presents the scaled feature distributions for each dataset in the filtered corpus. Features are normalised to $[0, 1]$ using the two-stage scaling pipeline described in Section 3.2. Each plot shows the distribution of all 42 features across the dataset's series, enabling visual inspection of the structural properties captured in the feature space.

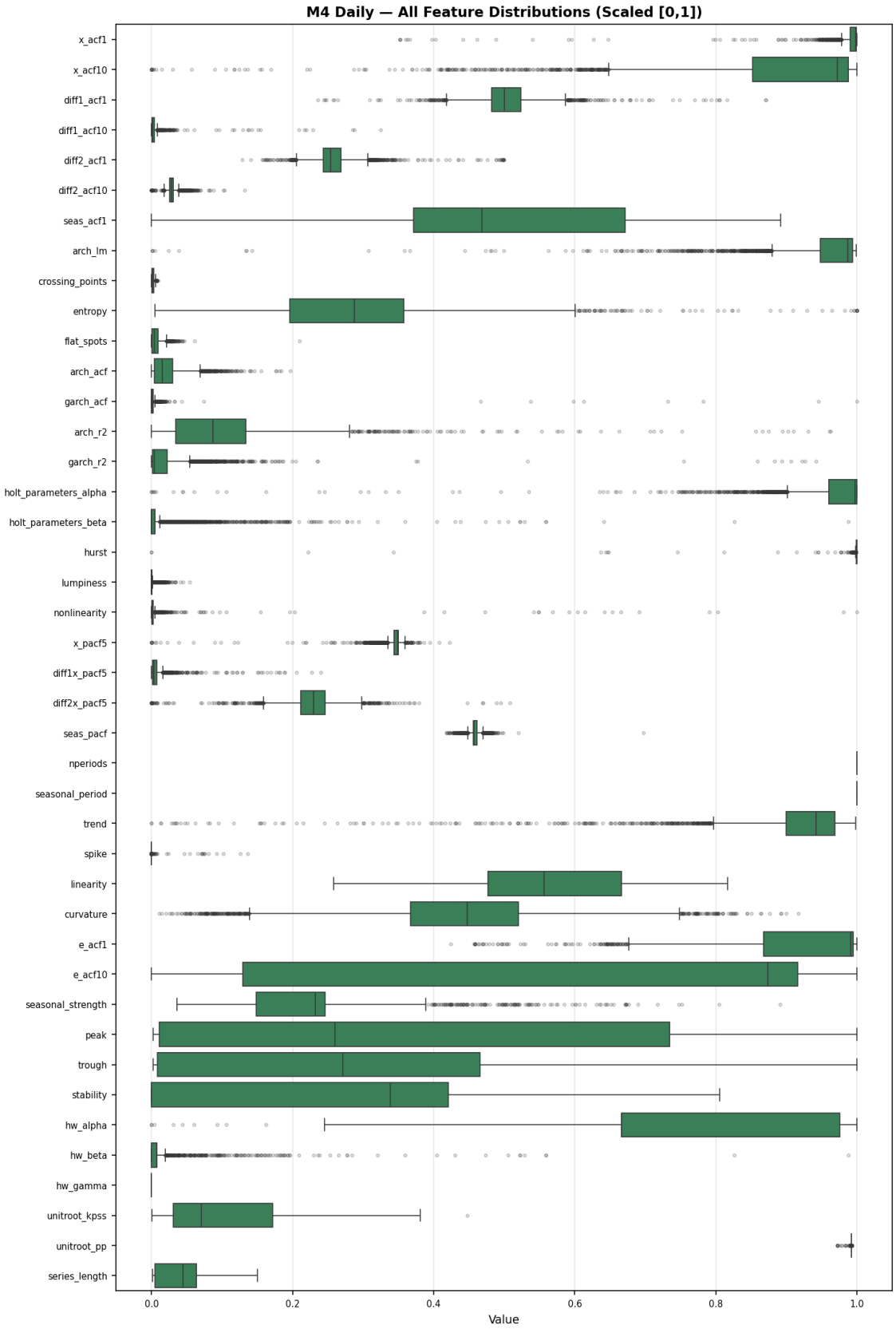


Figure B.1: Feature distributions for M4 Daily.

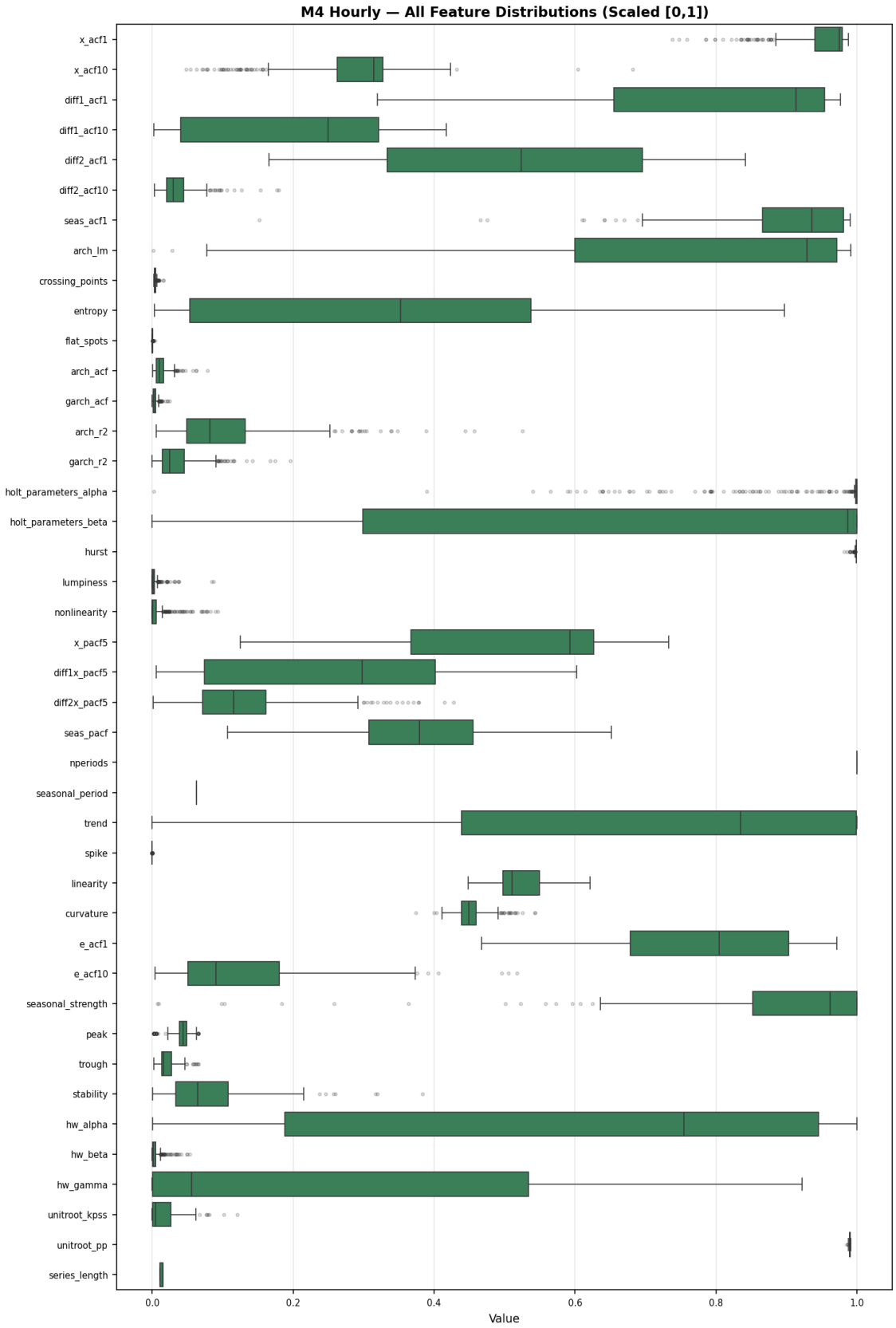


Figure B.2: Feature distributions for M4 Hourly.

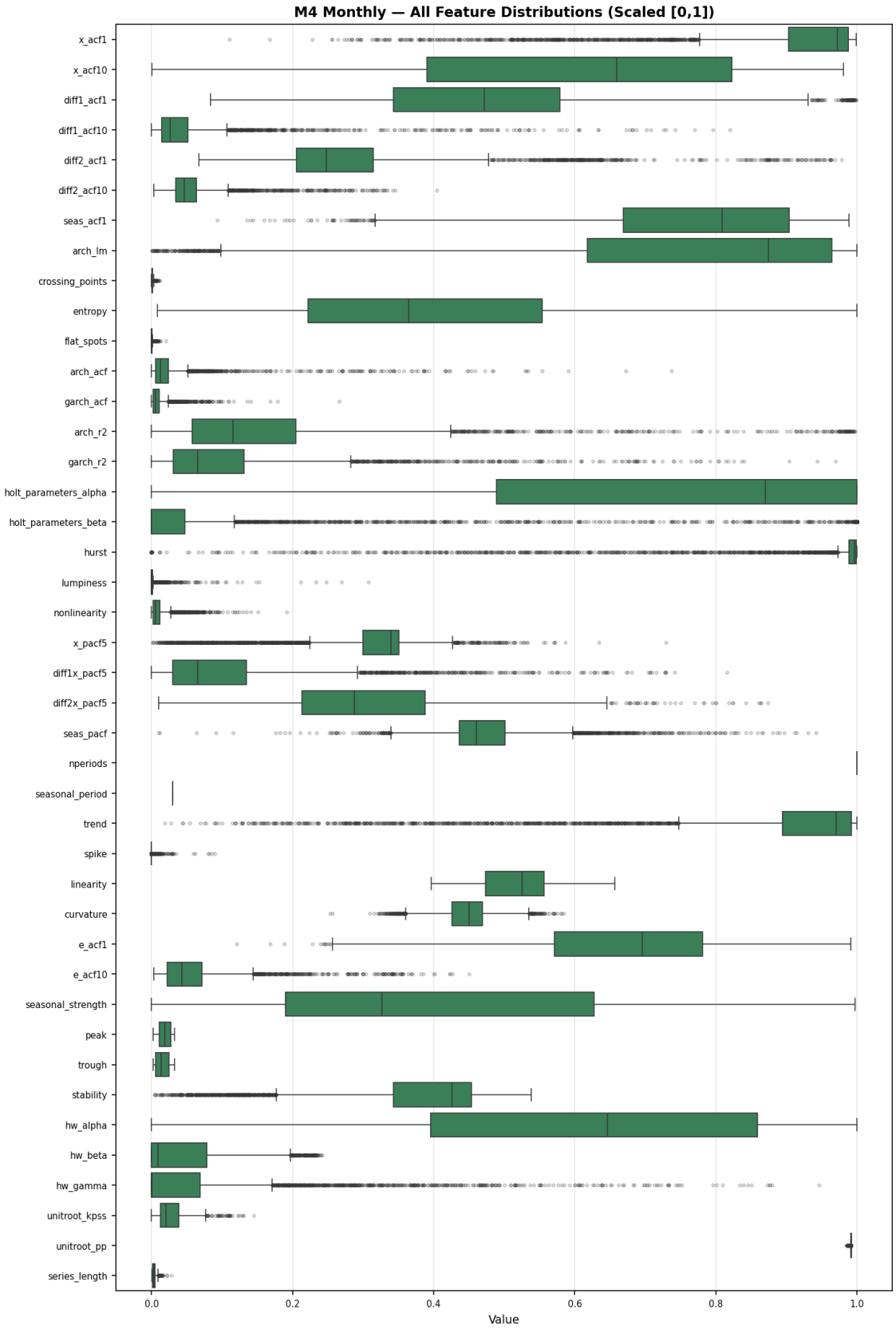


Figure B.3: Feature distributions for M4 Monthly.

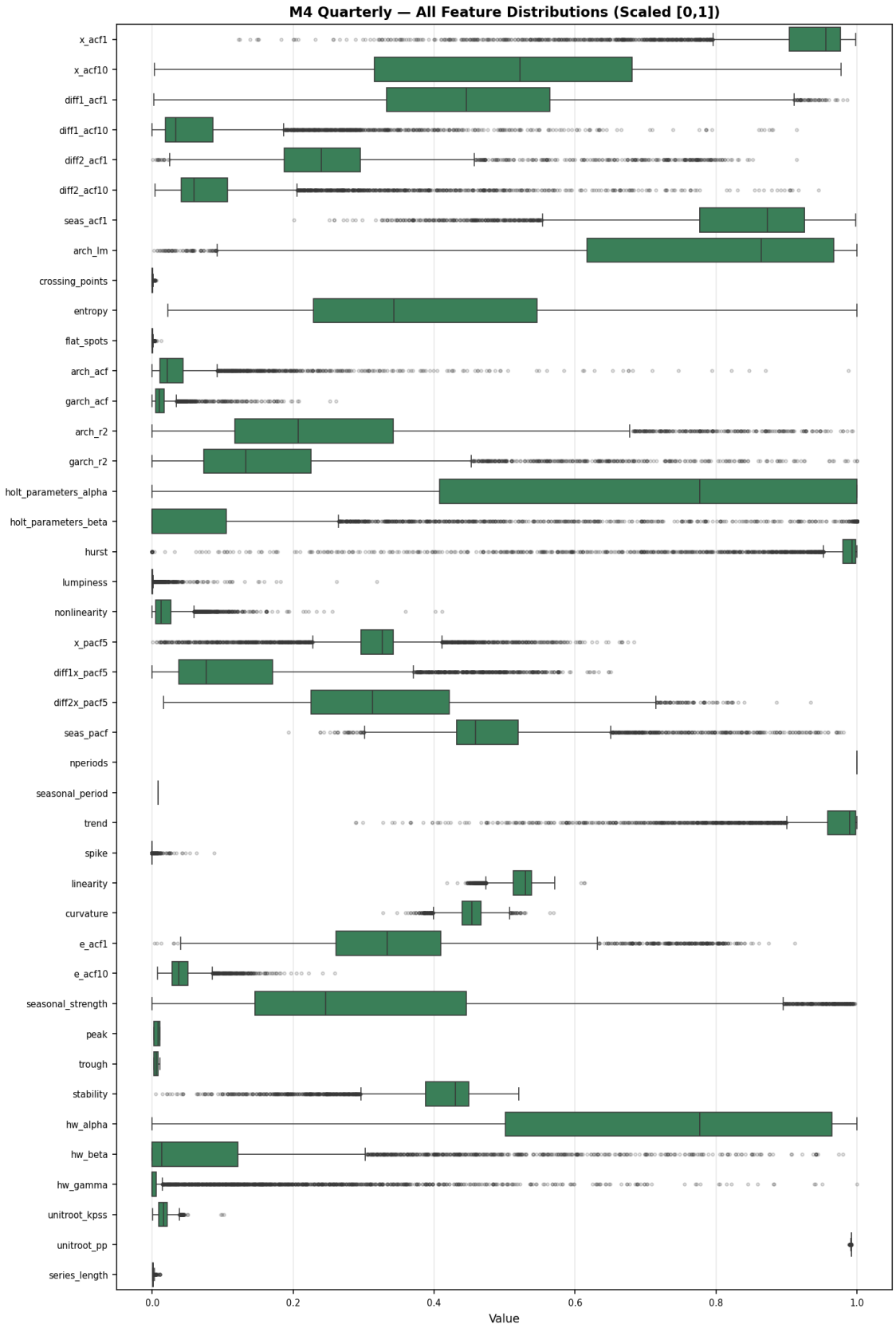


Figure B.4: Feature distributions for M4 Quarterly.

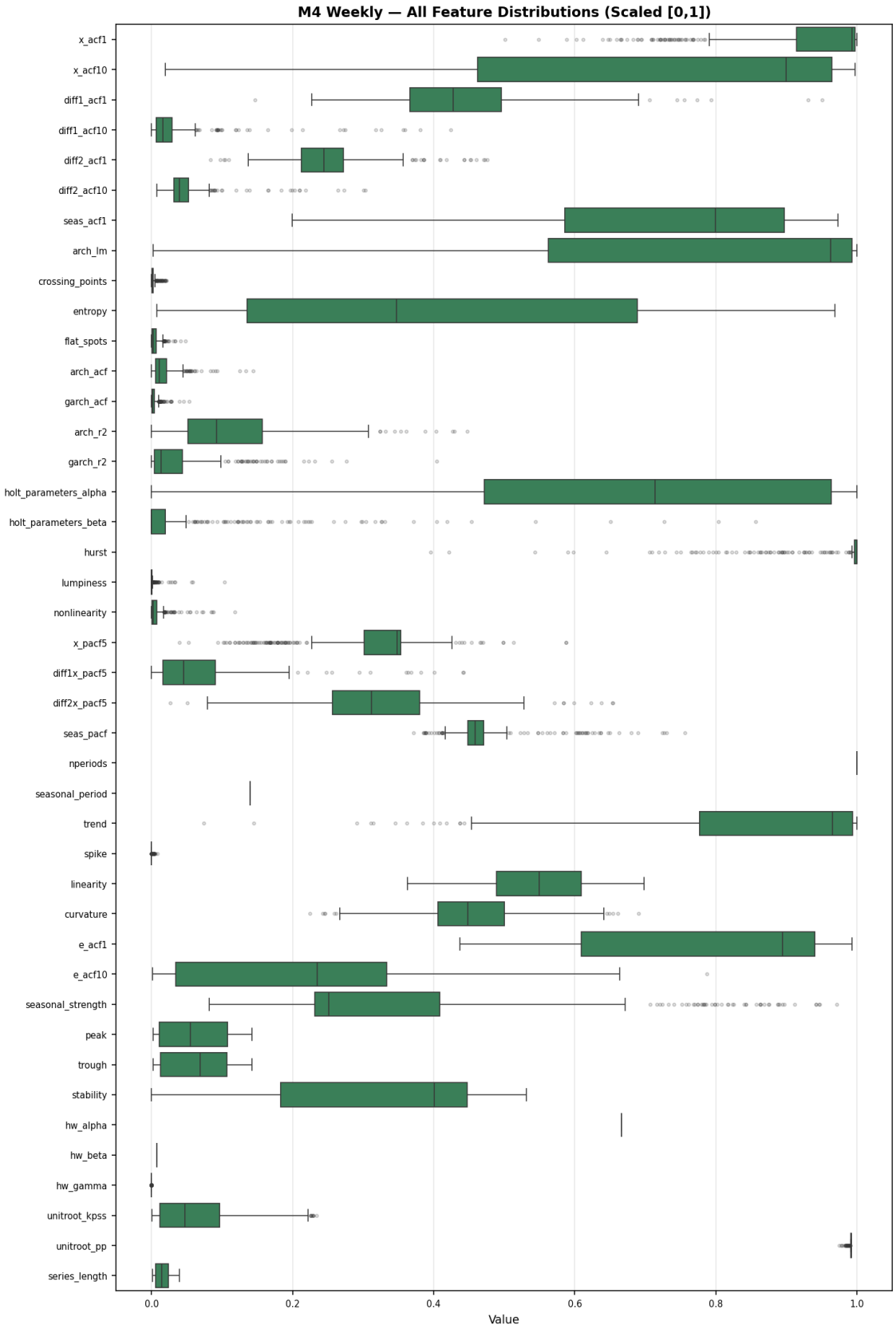


Figure B.5: Feature distributions for M4 Weekly.

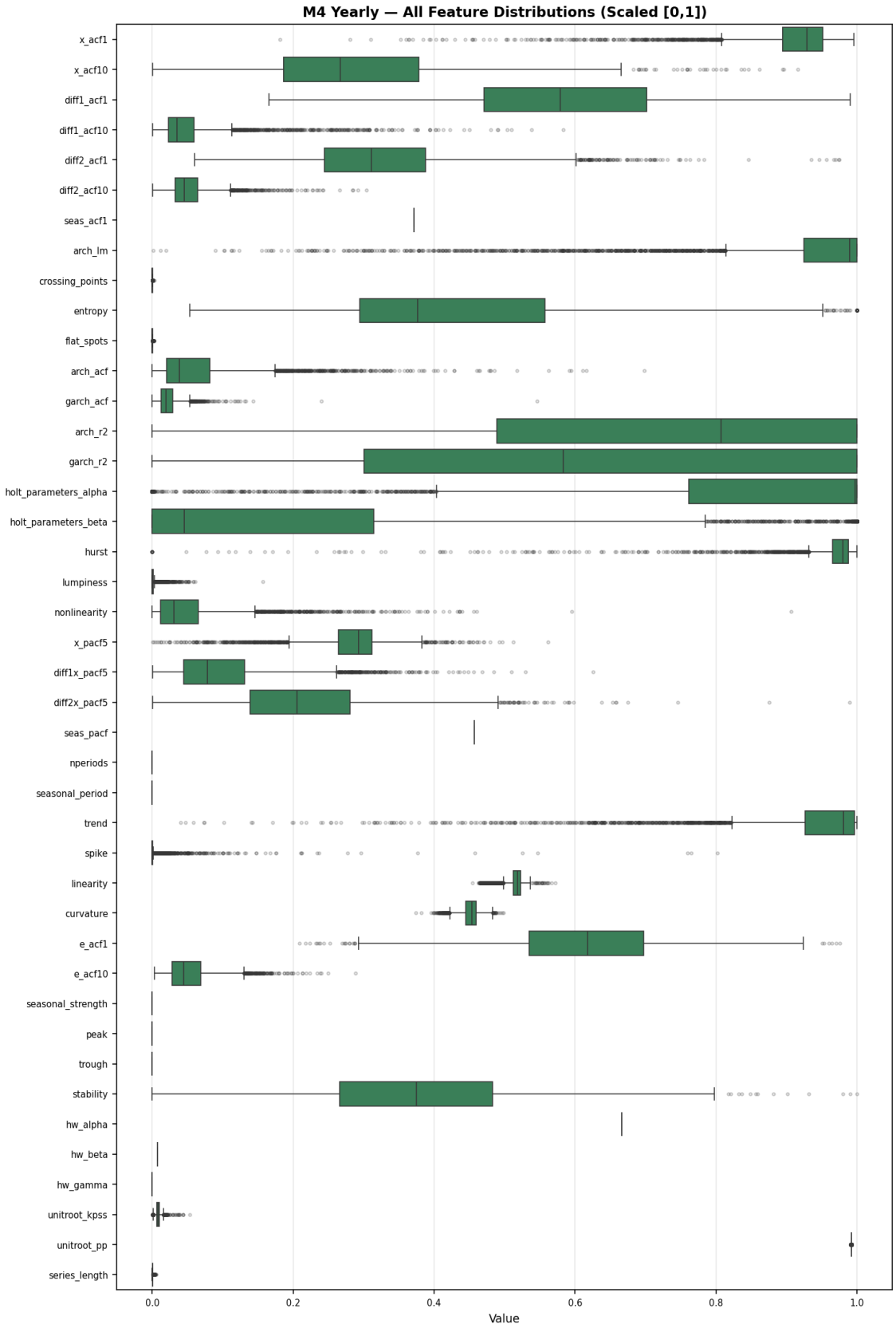


Figure B.6: Feature distributions for M4 Yearly.

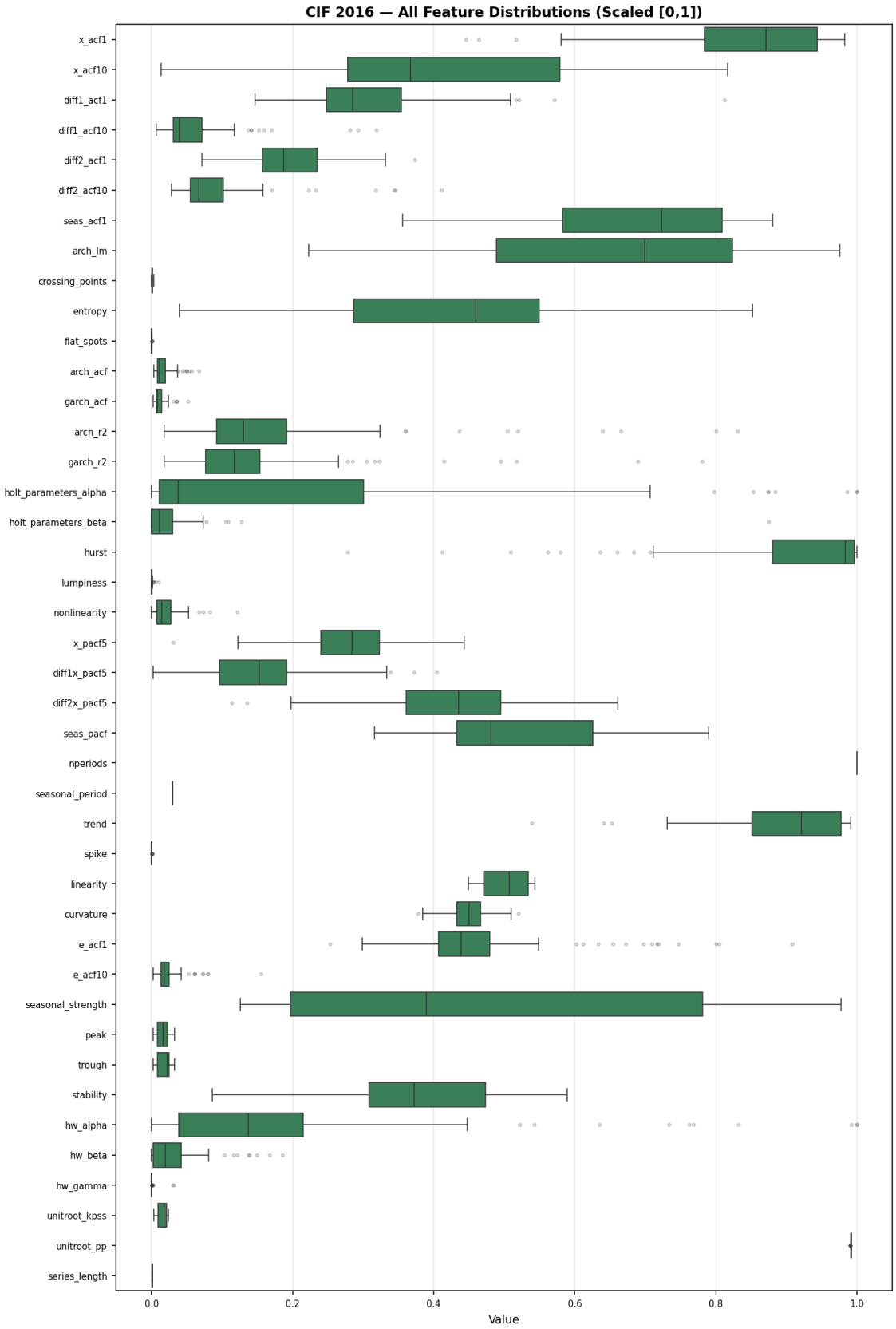


Figure B.7: Feature distributions for CIF 2016 (Monthly).

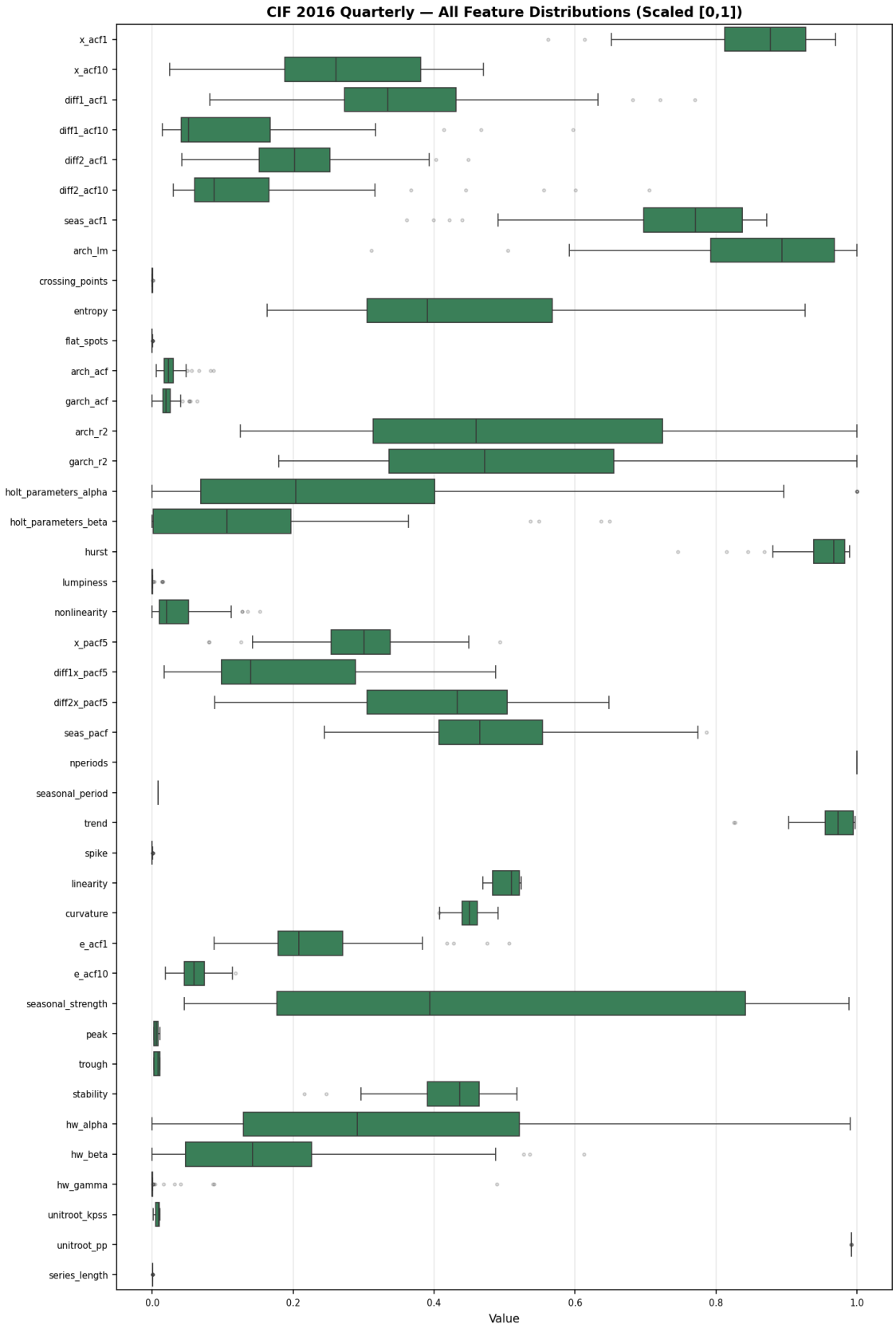


Figure B.8: Feature distributions for CIF 2016 (Quarterly).

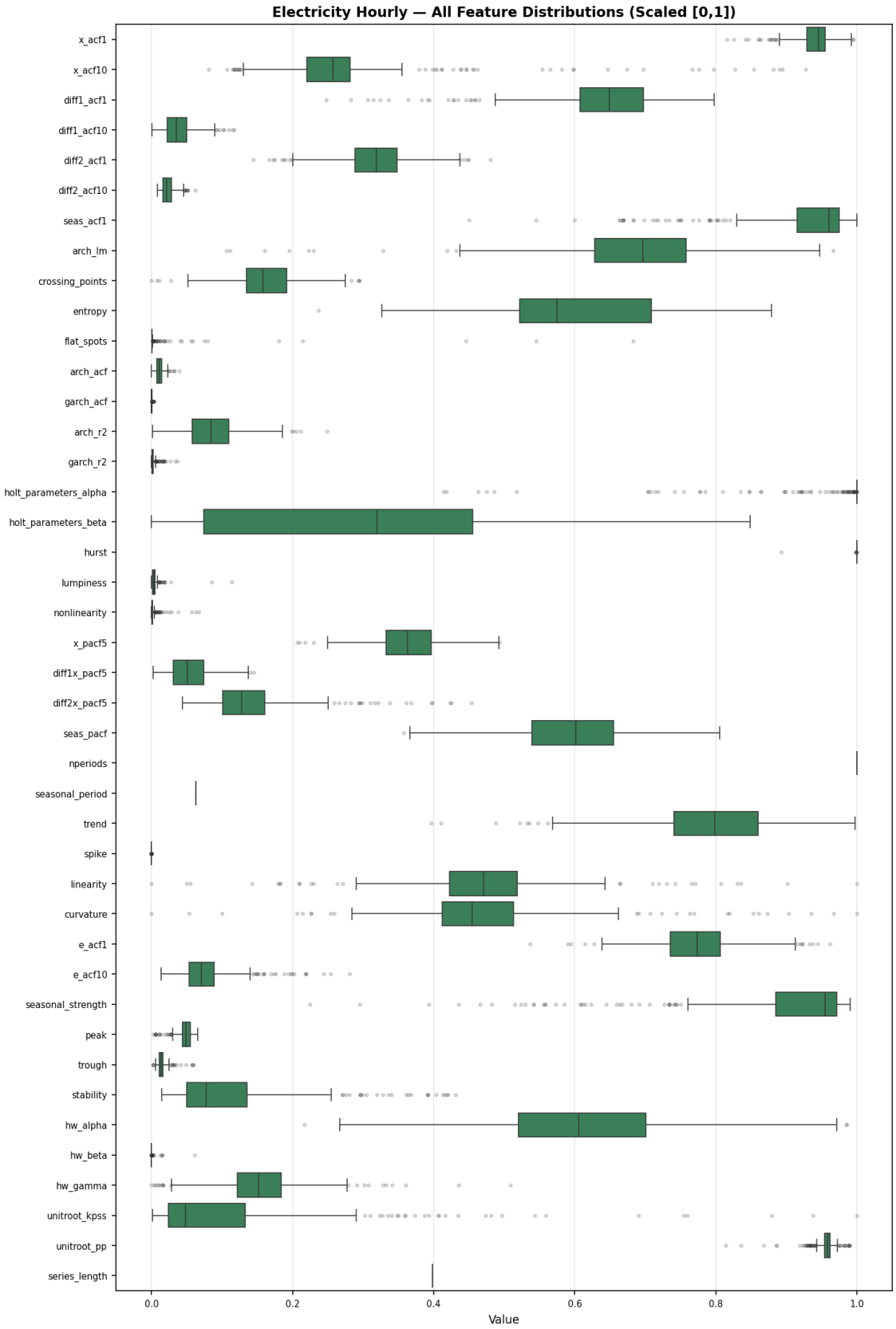


Figure B.9: Feature distributions for Electricity (Hourly).

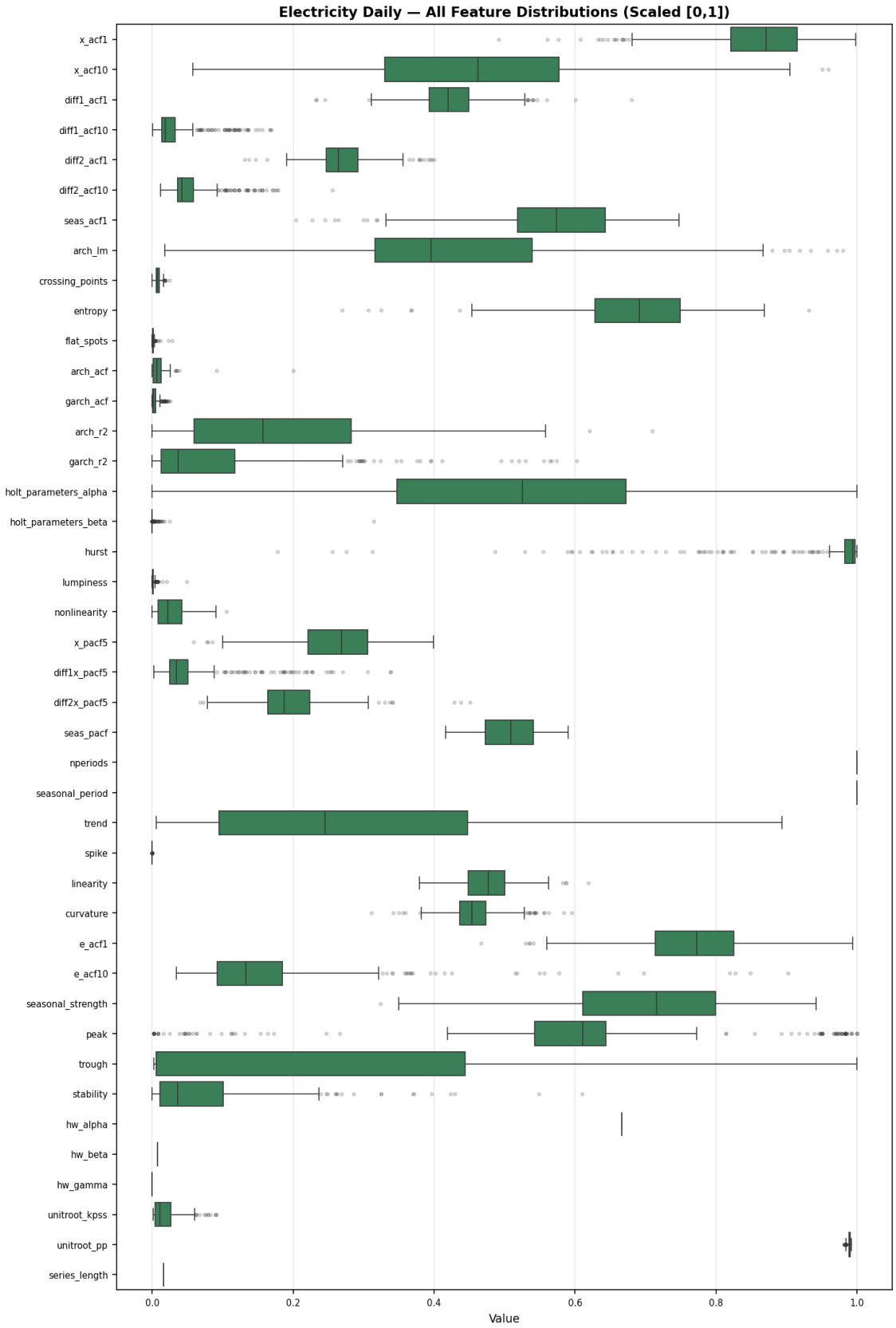


Figure B.10: Feature distributions for Electricity (Daily resampled).

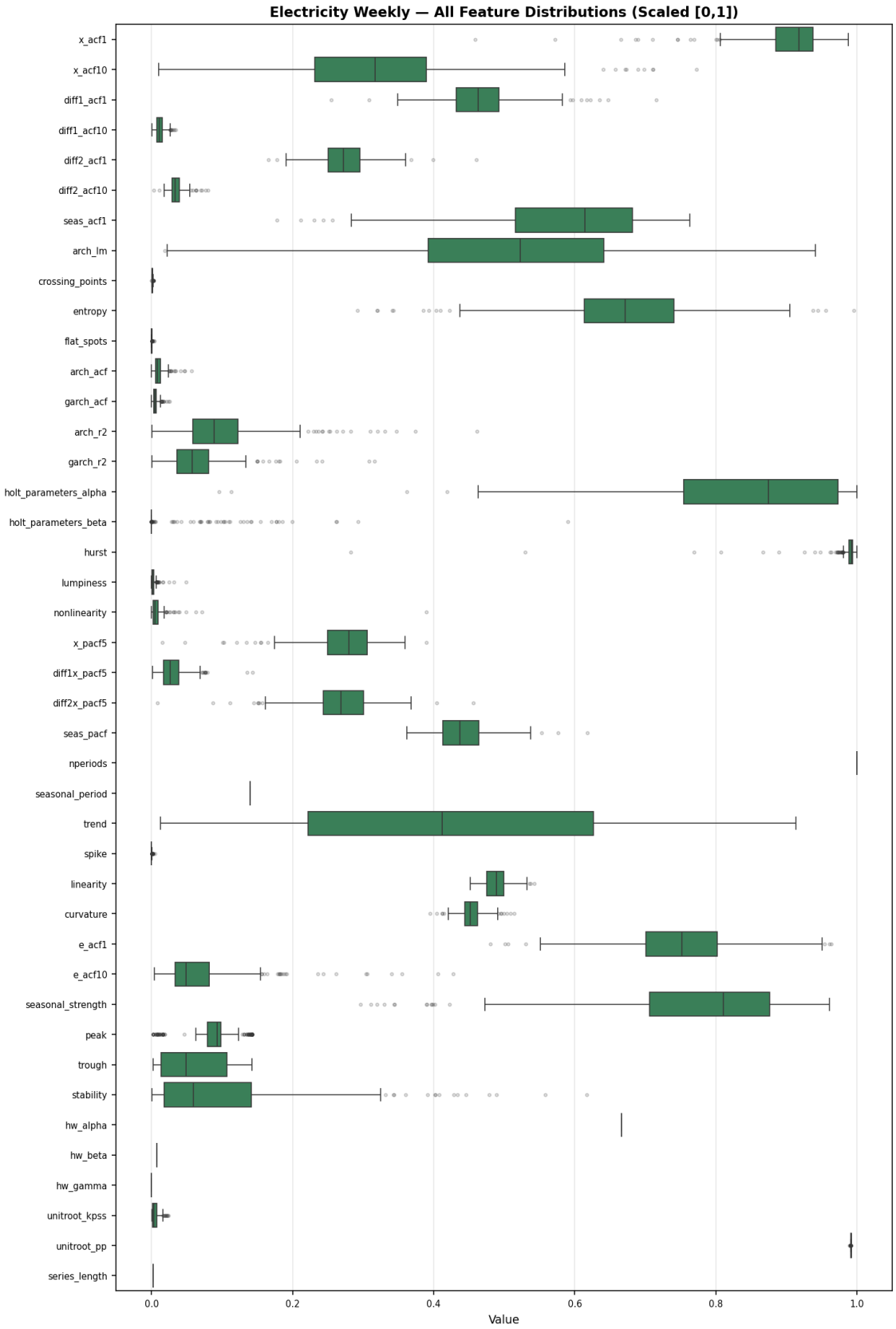


Figure B.11: Feature distributions for Electricity (Weekly).

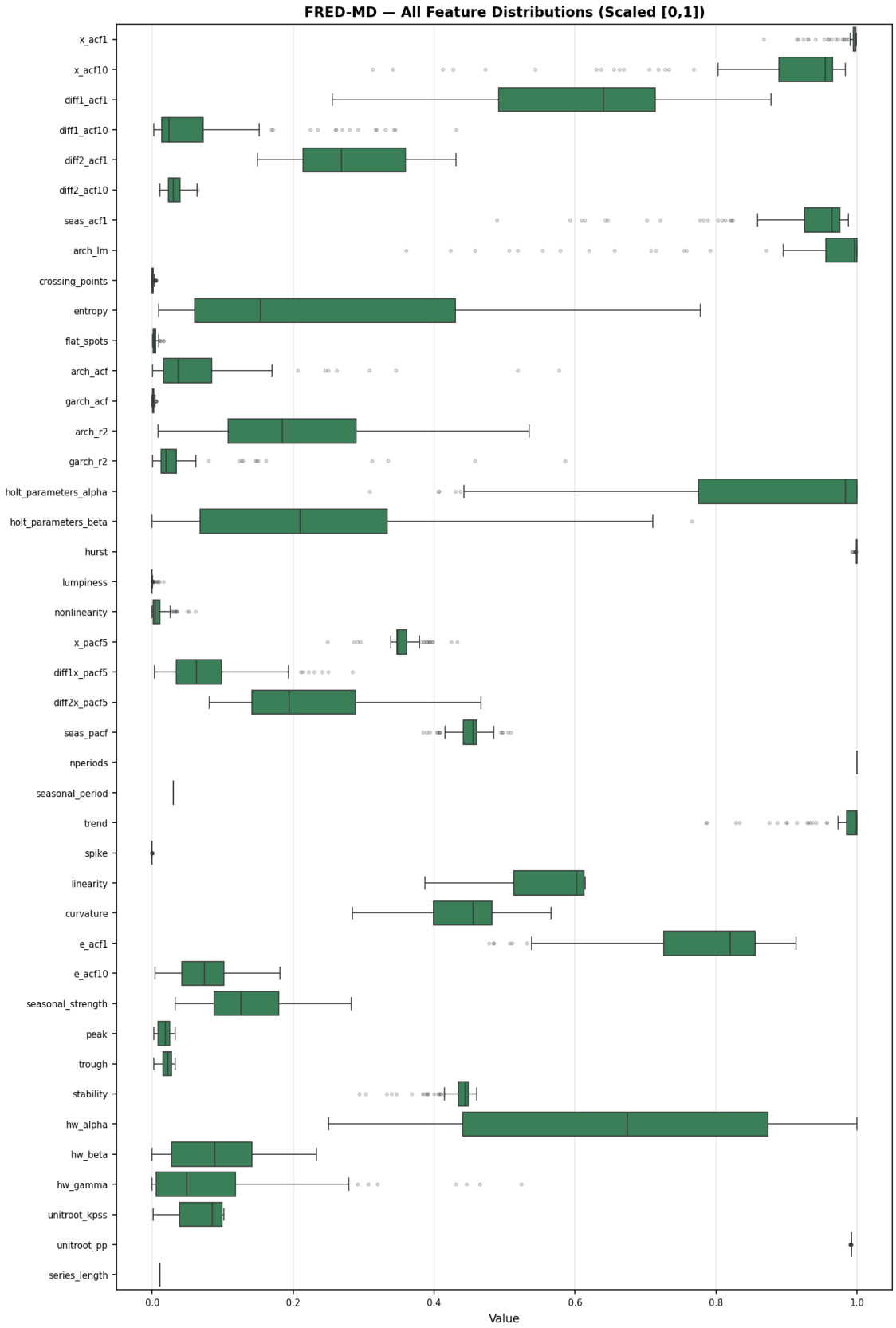


Figure B.12: Feature distributions for FRED-MD (Monthly).

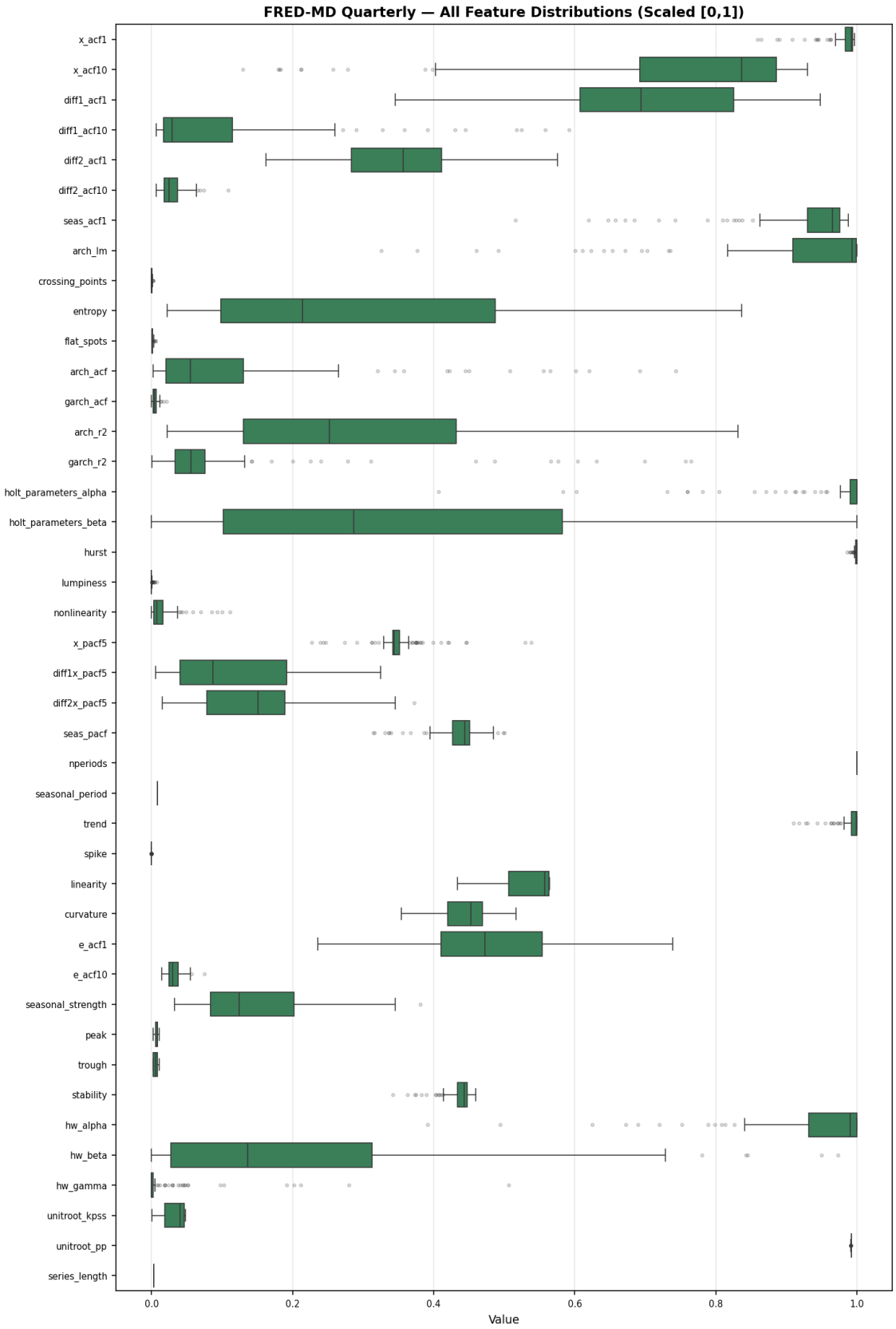


Figure B.13: Feature distributions for FRED-MD (Quarterly).

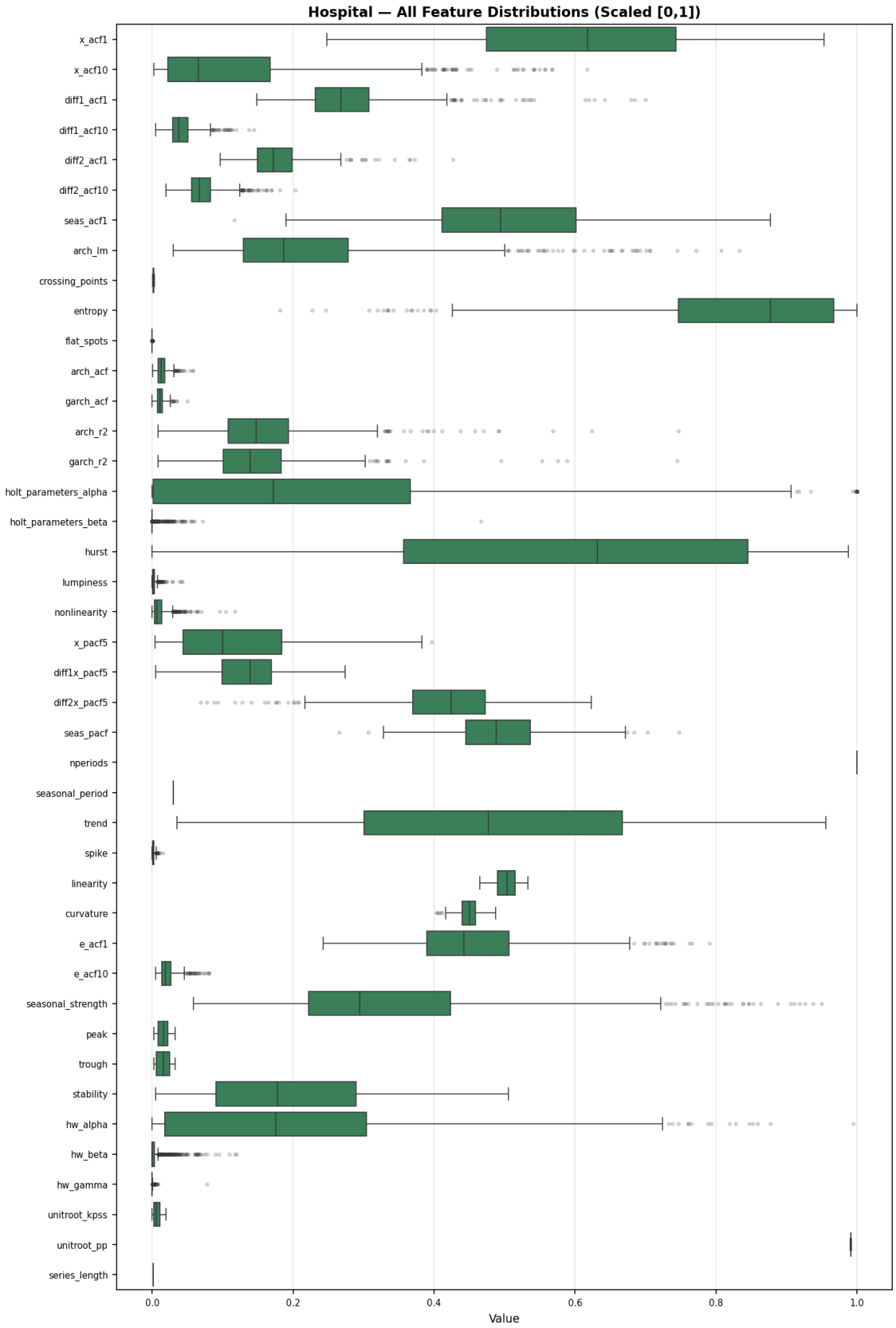


Figure B.14: Feature distributions for Hospital.

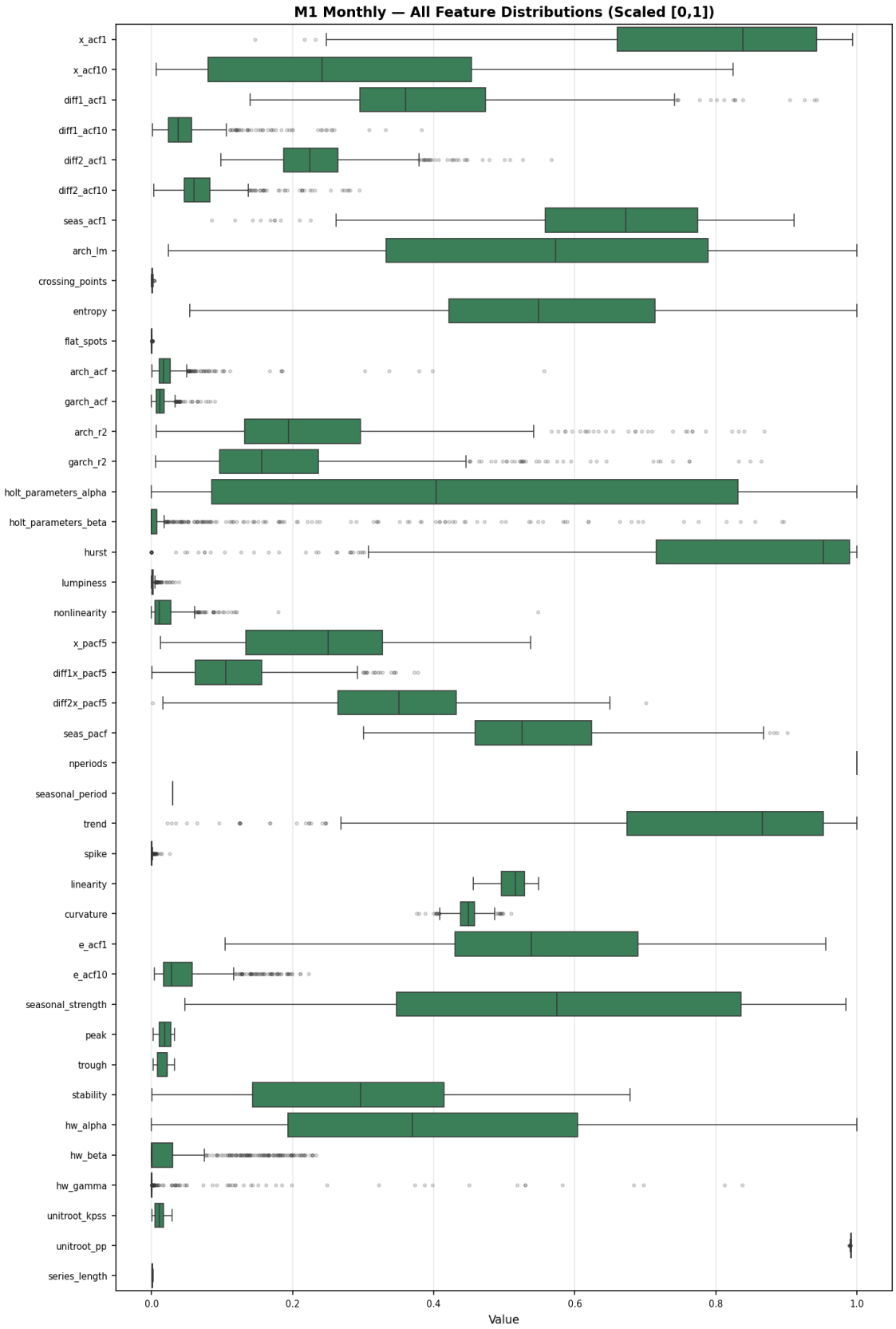


Figure B.15: Feature distributions for M1 (Monthly).

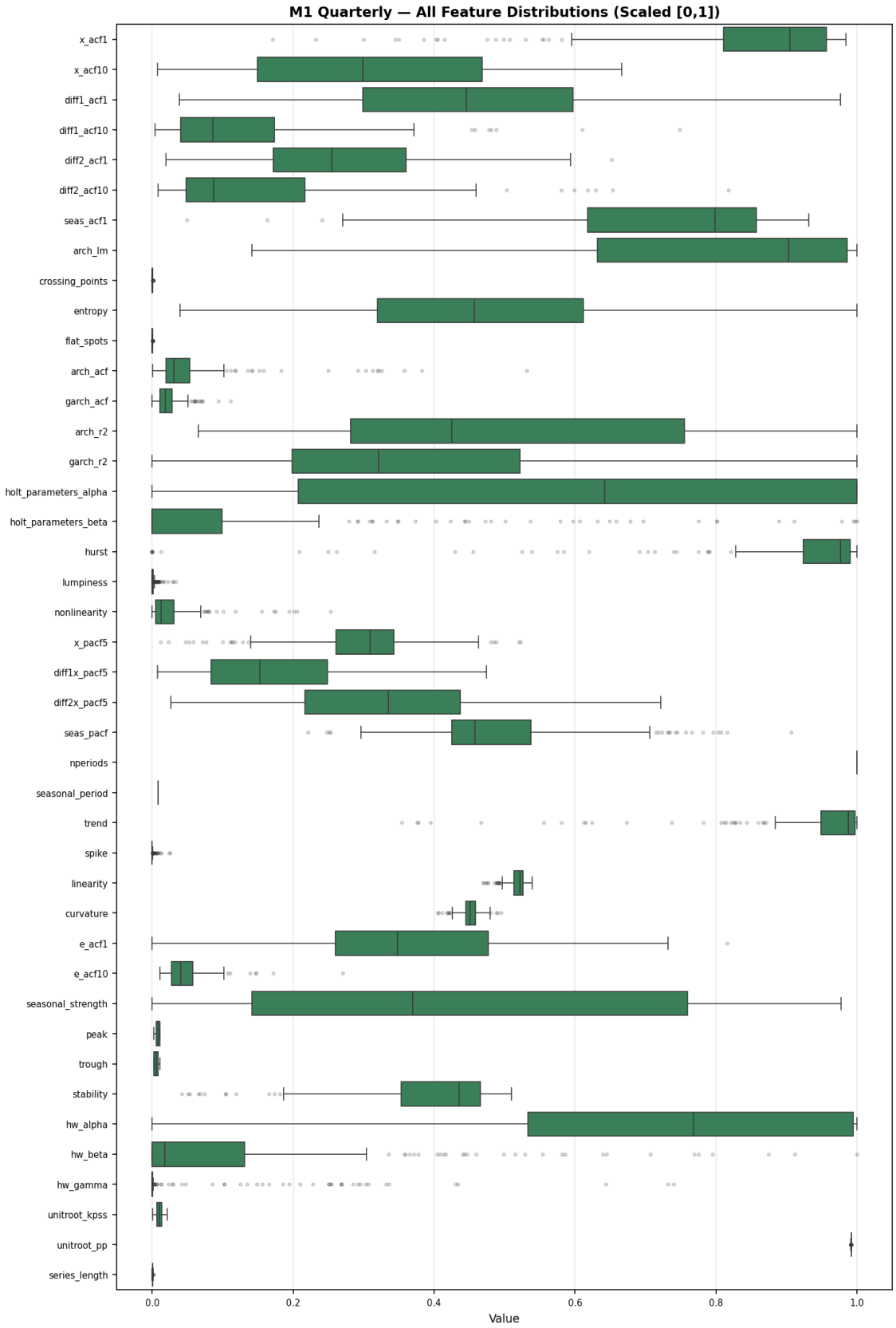


Figure B.16: Feature distributions for M1 (Quarterly).

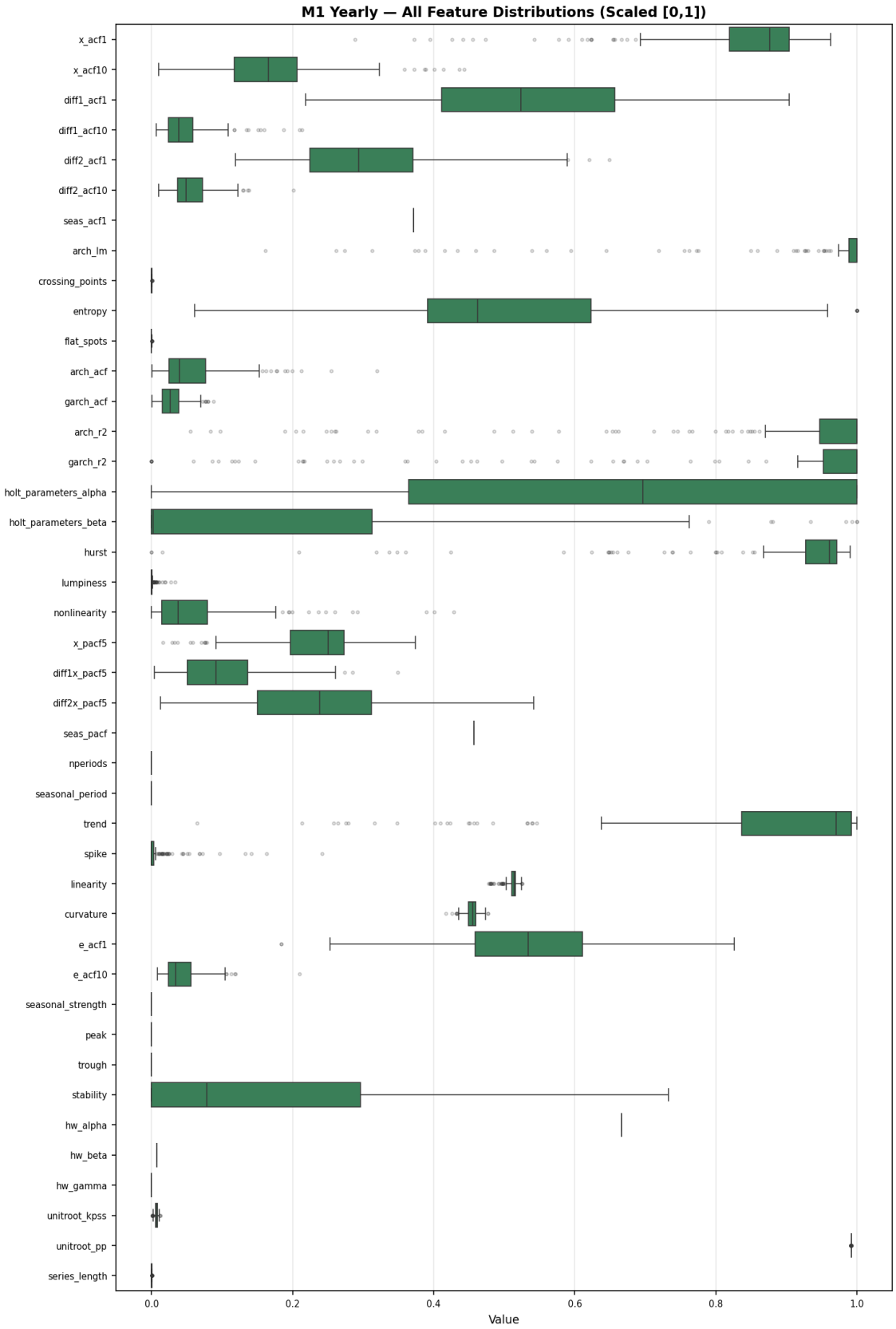


Figure B.17: Feature distributions for M1 (Yearly).

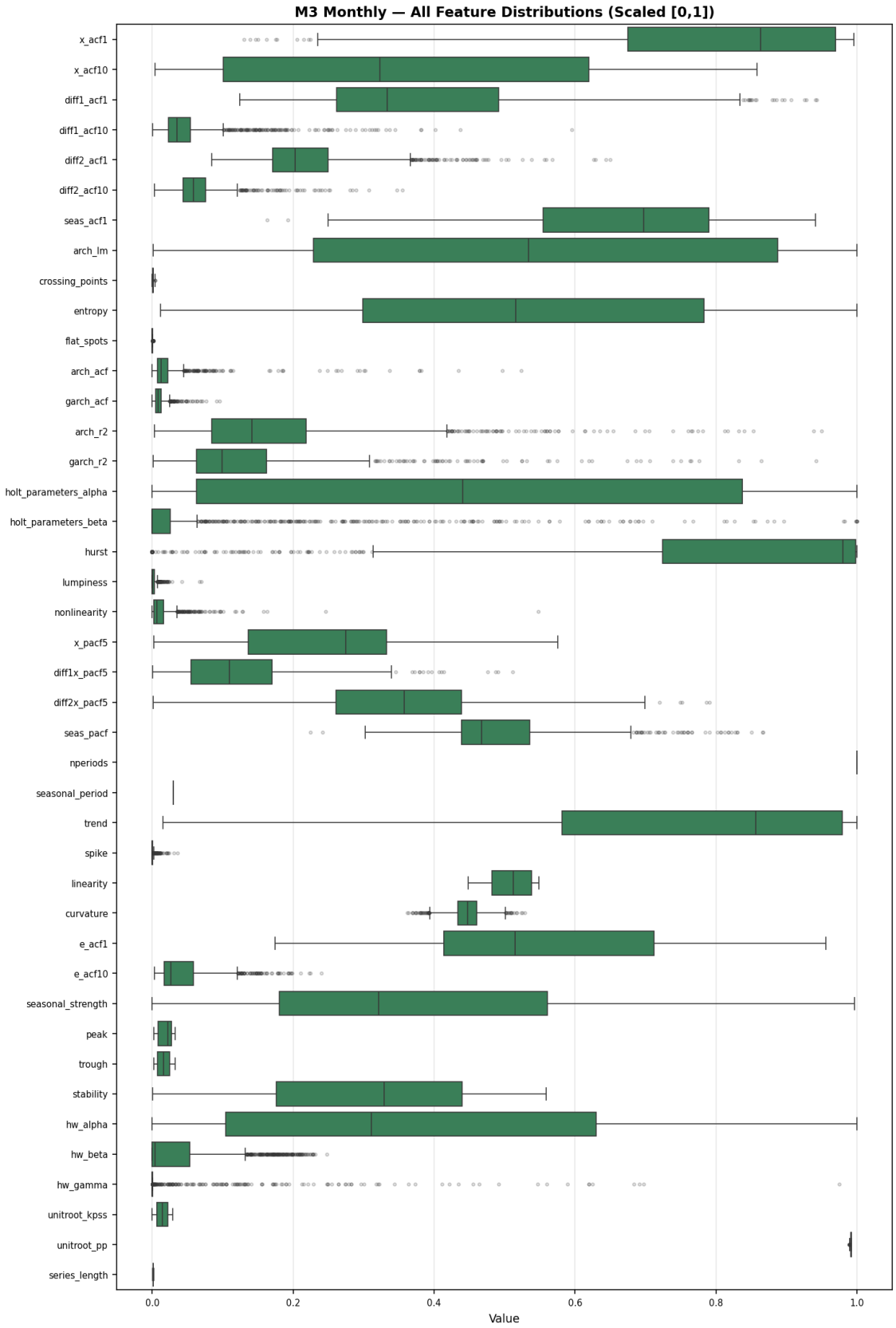


Figure B.18: Feature distributions for M3 (Monthly).

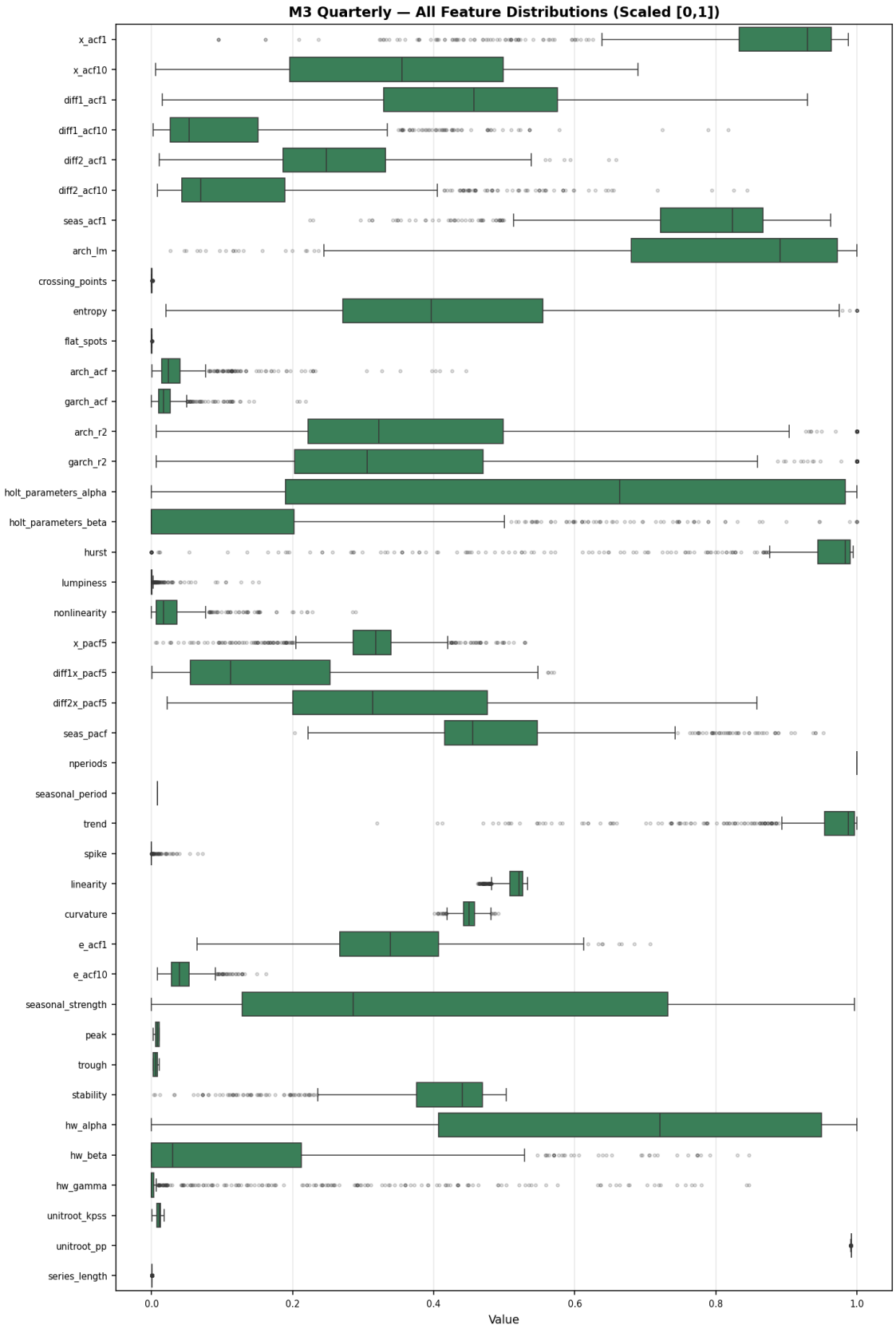


Figure B.19: Feature distributions for M3 (Quarterly).

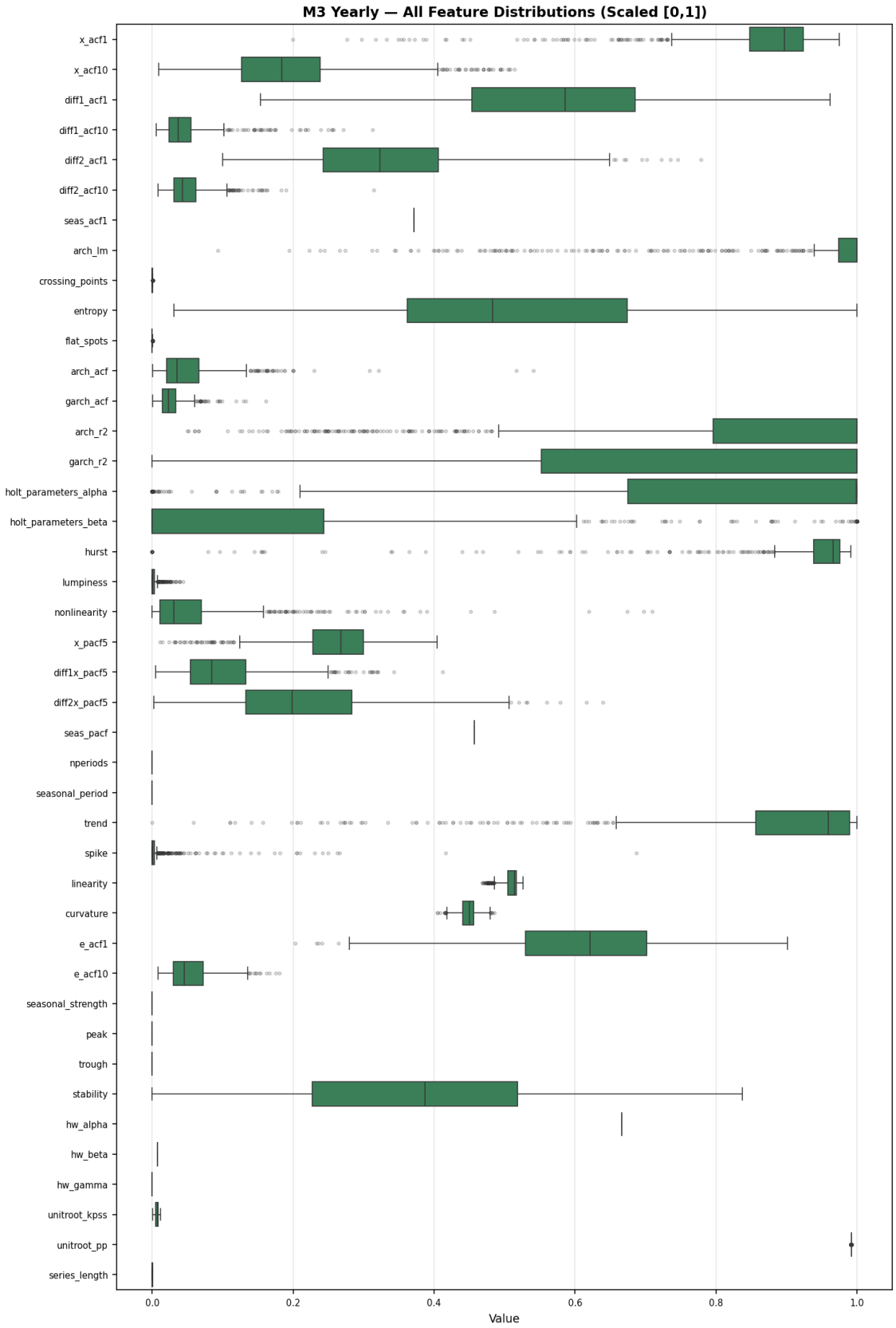


Figure B.20: Feature distributions for M3 (Yearly).

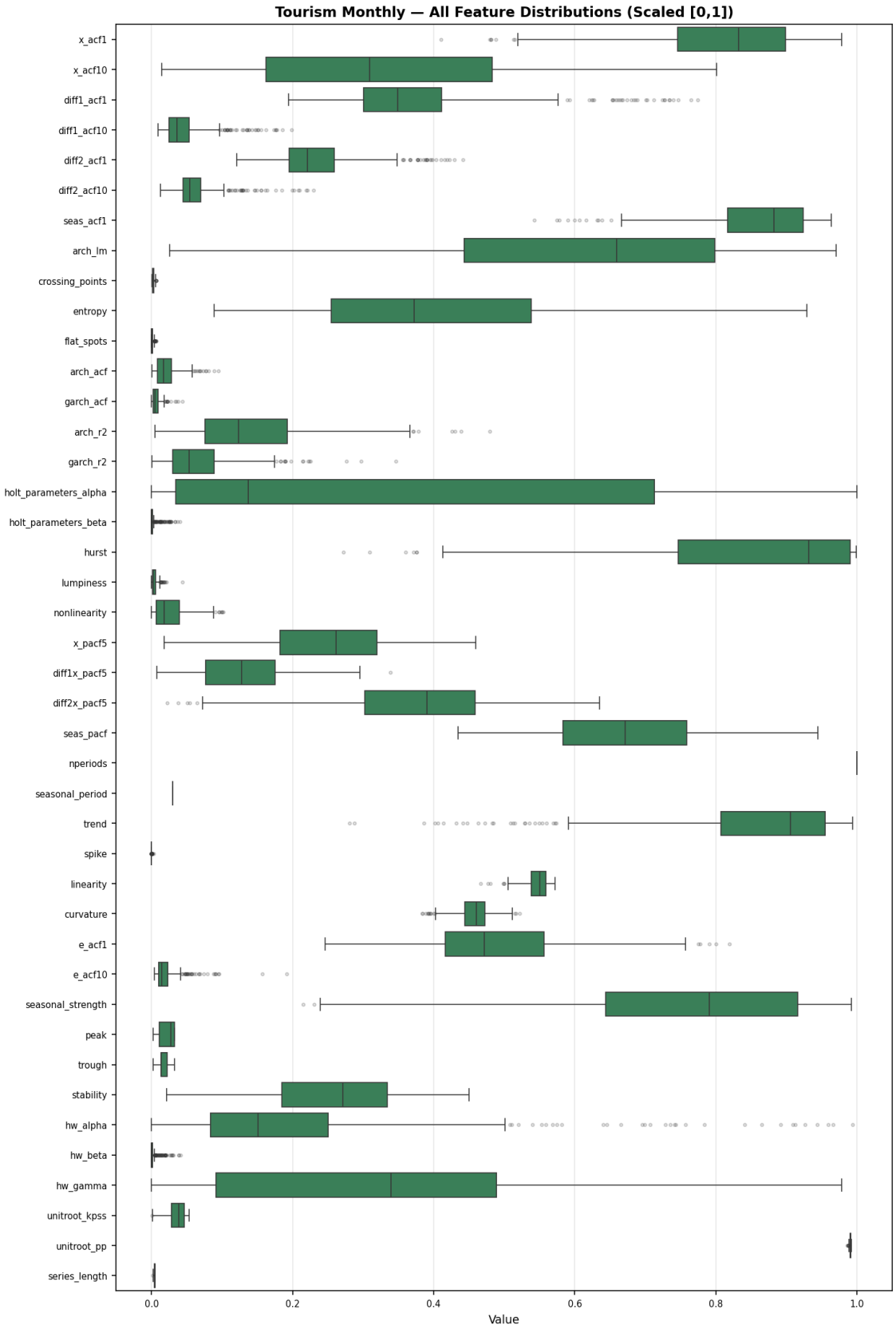


Figure B.21: Feature distributions for Tourism (Monthly).

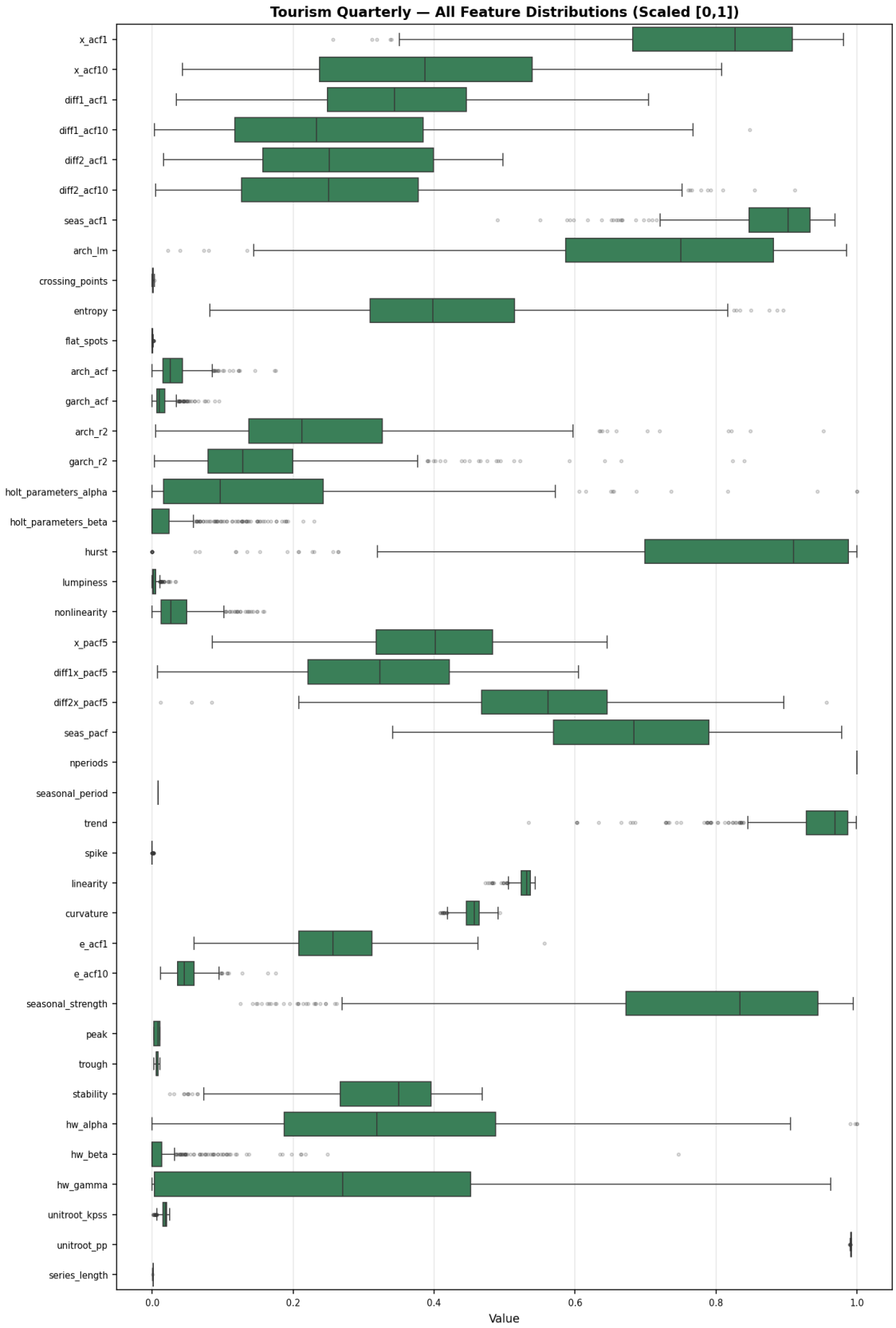


Figure B.22: Feature distributions for Tourism (Quarterly).

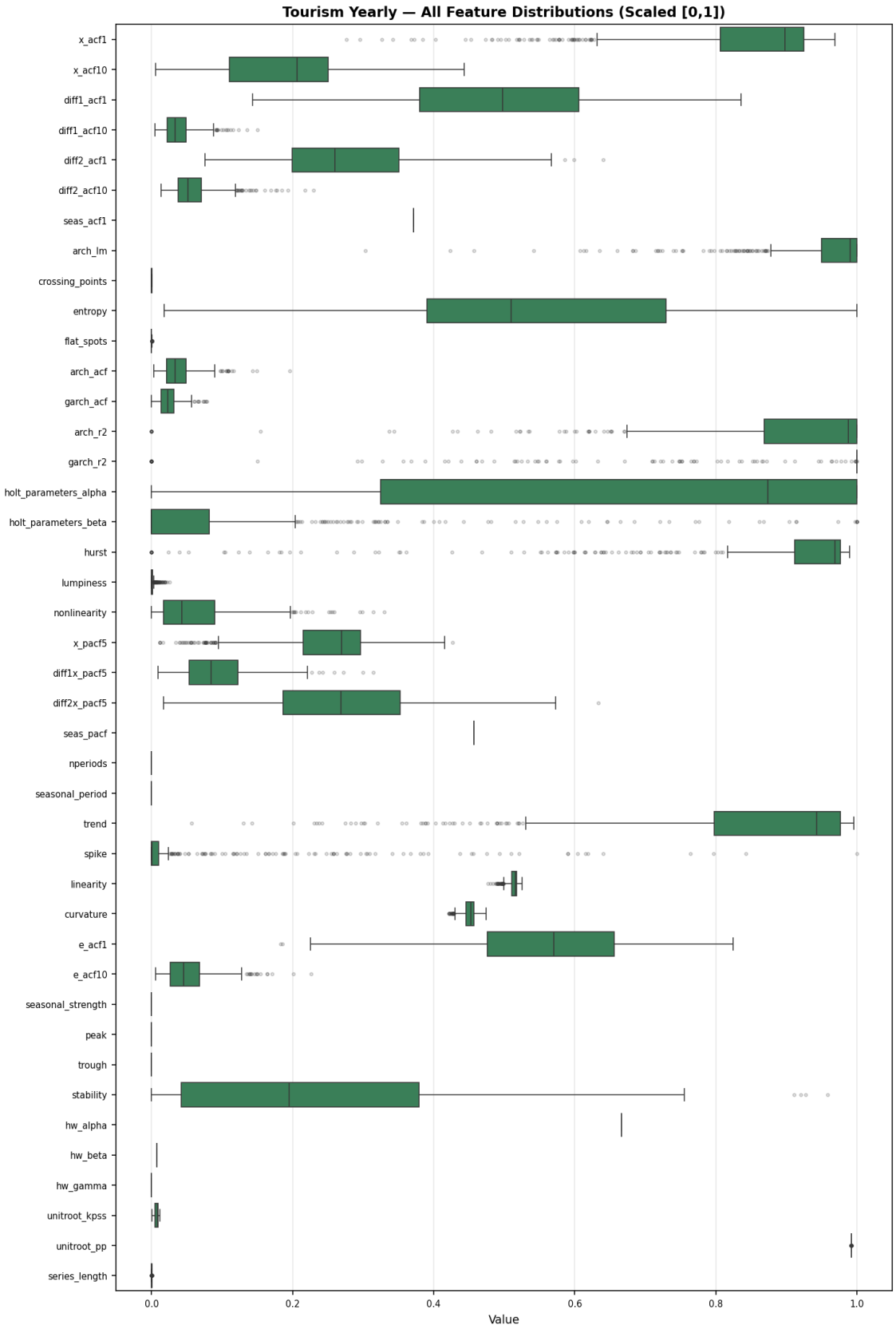


Figure B.23: Feature distributions for Tourism (Yearly).

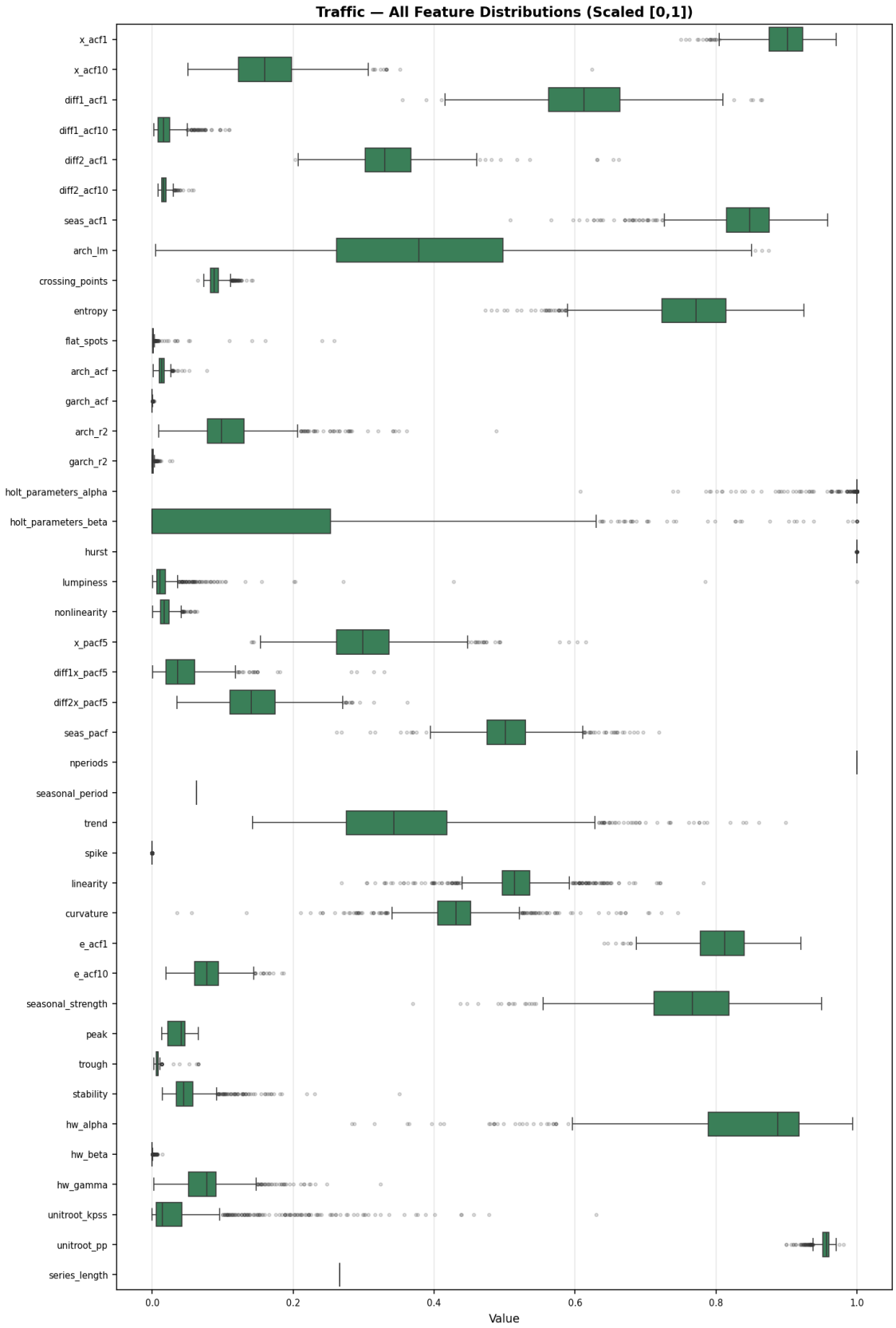


Figure B.24: Feature distributions for Traffic (Hourly).

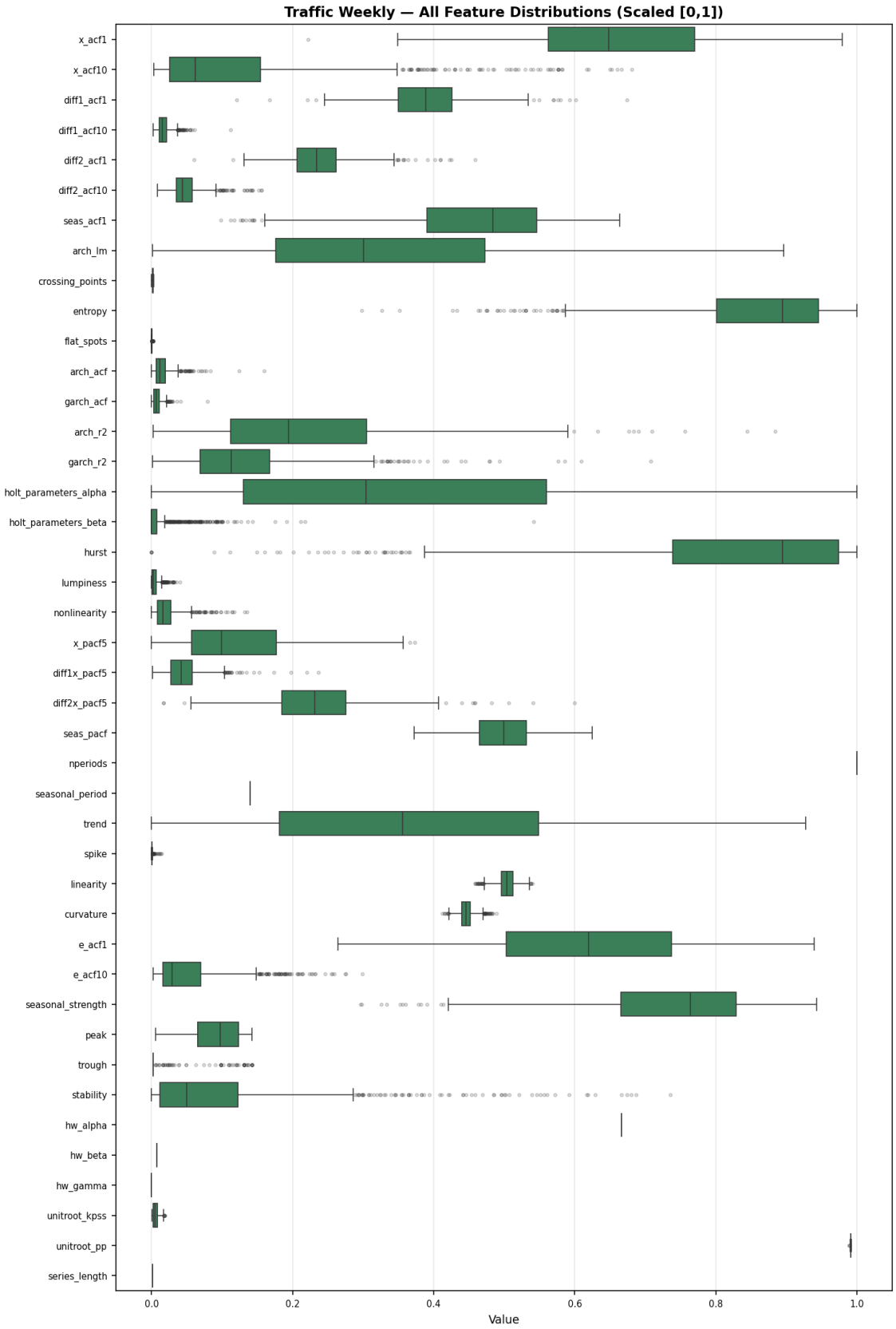


Figure B.26: Feature distributions for Traffic (Weekly resampled).

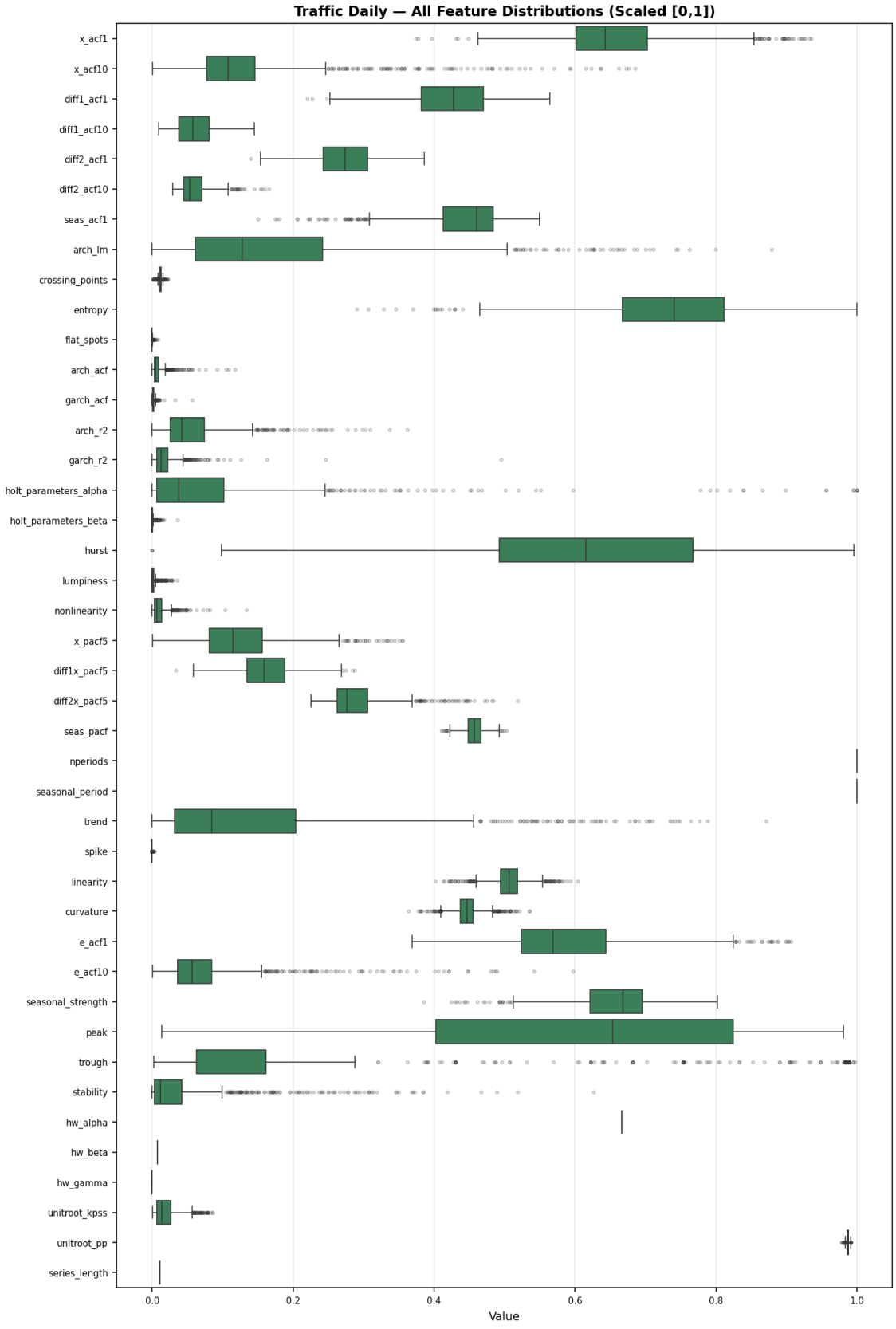


Figure B.25: Feature distributions for Traffic (Daily resampled).

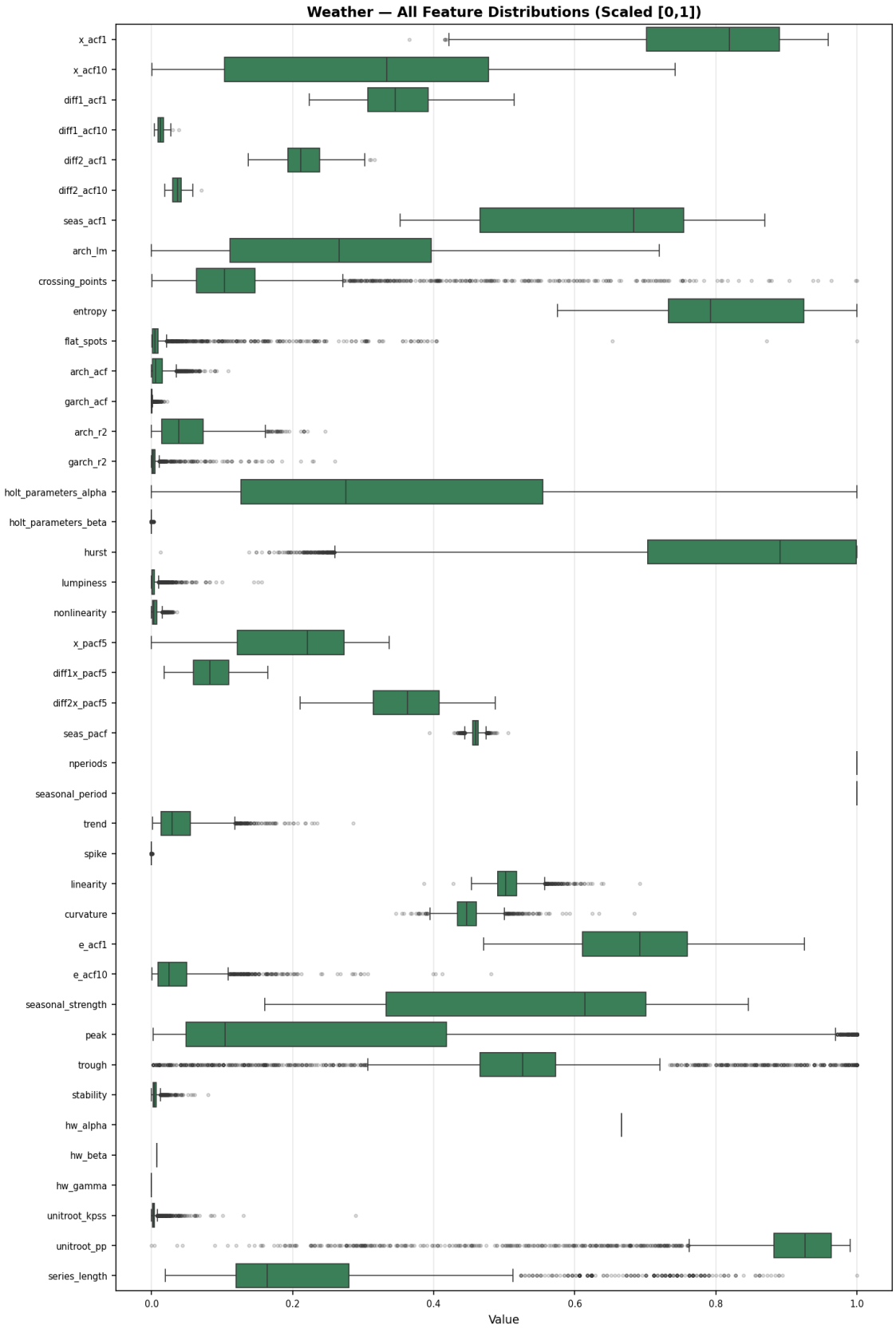


Figure B.27: Feature distributions for Weather (Daily).

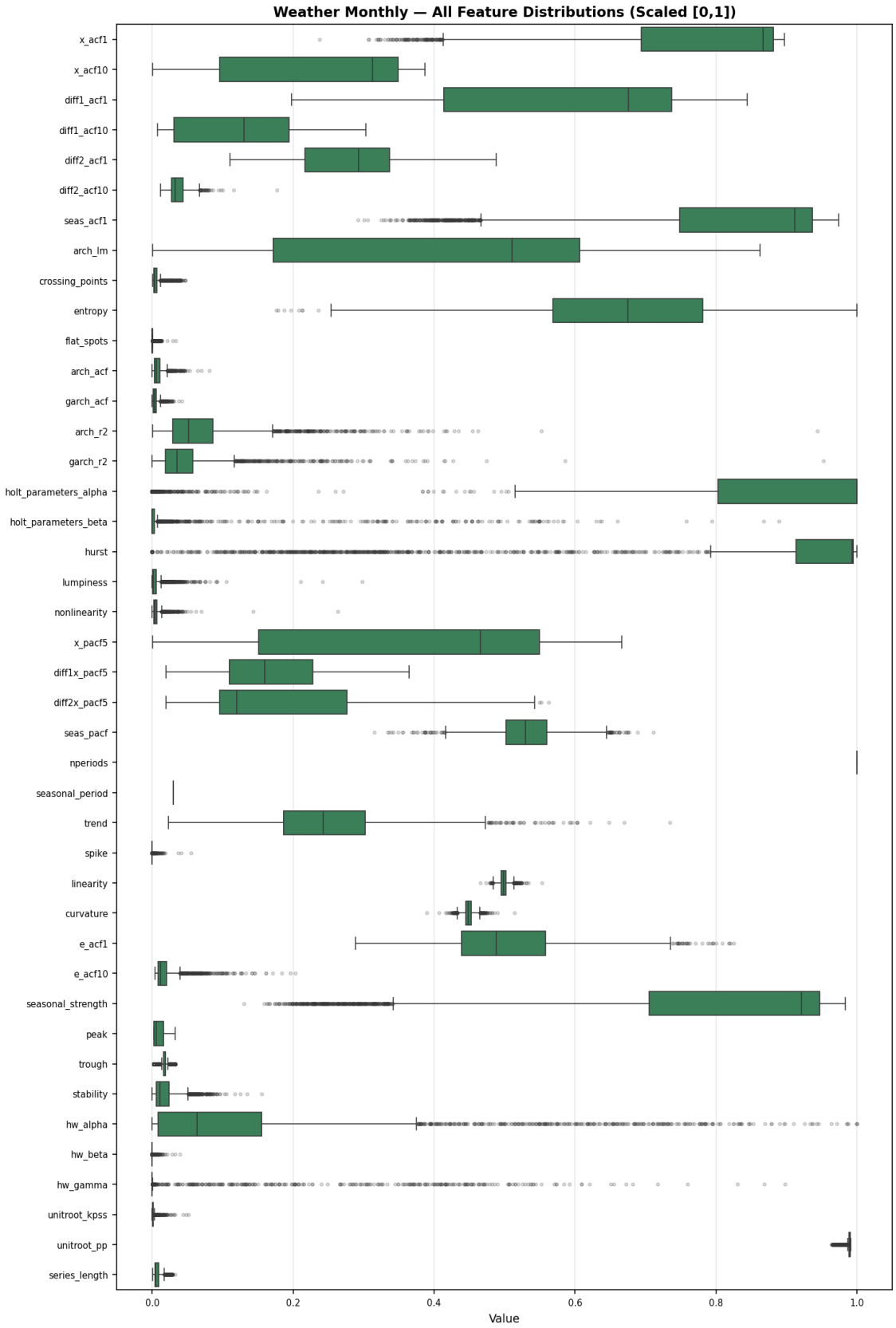


Figure B.28: Feature distributions for Weather (Monthly resampled).

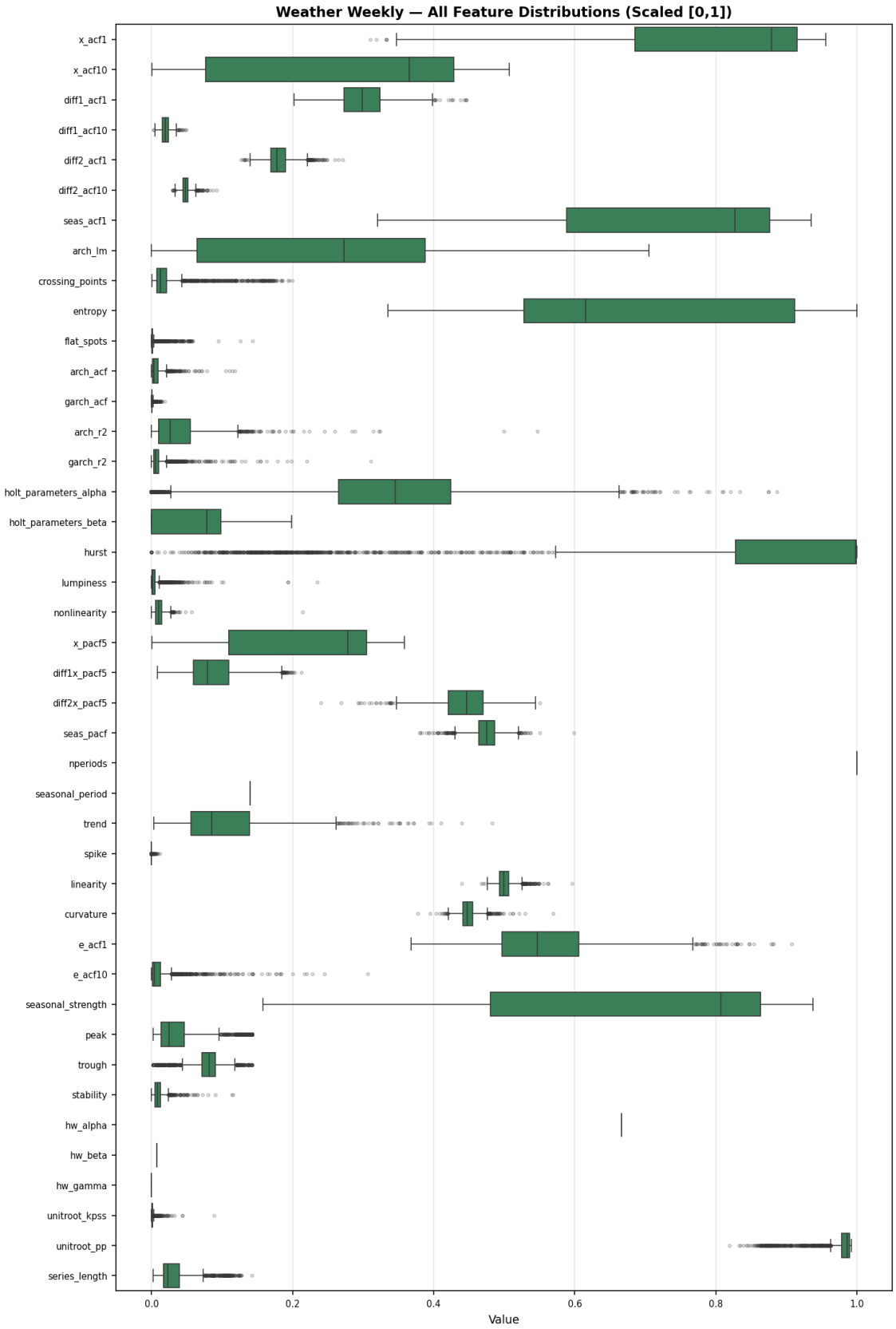


Figure B.29: Feature distributions for Weather (Weekly resampled).

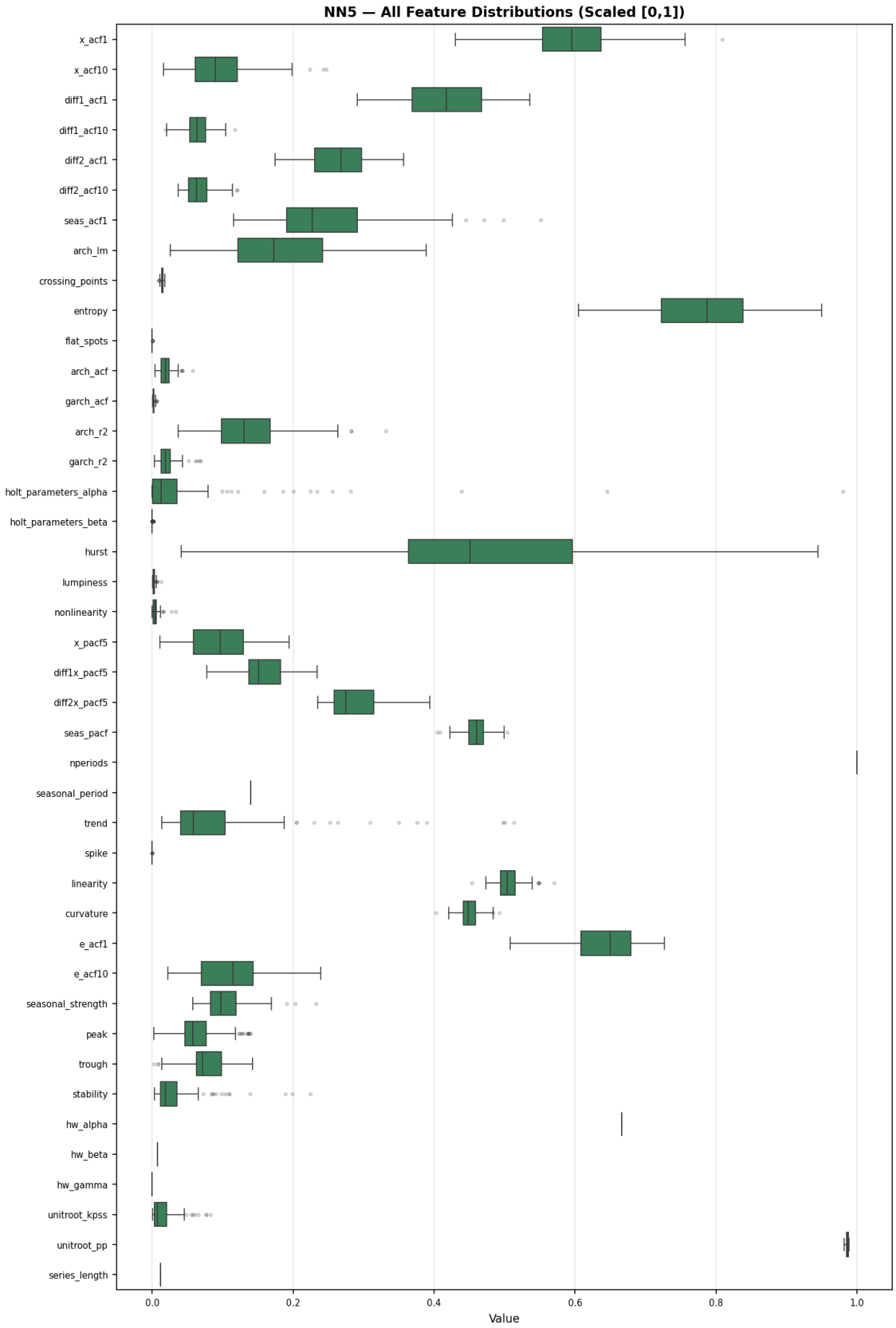


Figure B.30: Feature distributions for NN5.

B.2 Per-Metric Dataset Representativeness Distributions

This section presents the per-metric representativeness distributions broken down by dataset and temporal frequency group, for both the global and same-frequency audit variants. Each figure corresponds to one of the six auditing metrics and displays the scores of all 26 datasets, ordered by frequency group. These plots complement the aggregate distributions presented in Figure 4.1 and support the dataset-specific observations discussed in Section 4.3.

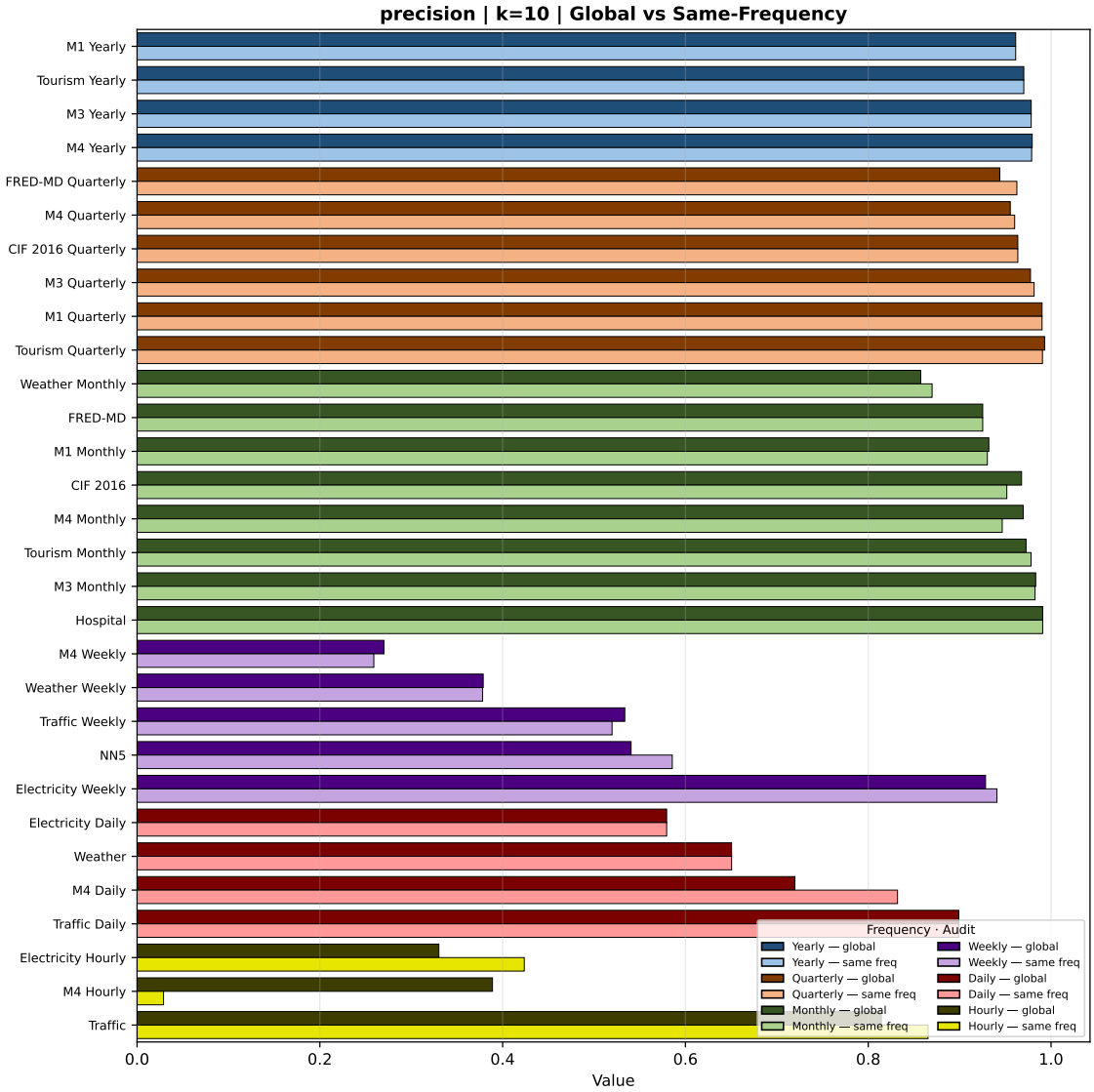


Figure B.31: Precision scores per dataset at $k = 10$, ordered by frequency group. Dark bars = global audit; light bars = same-frequency audit.

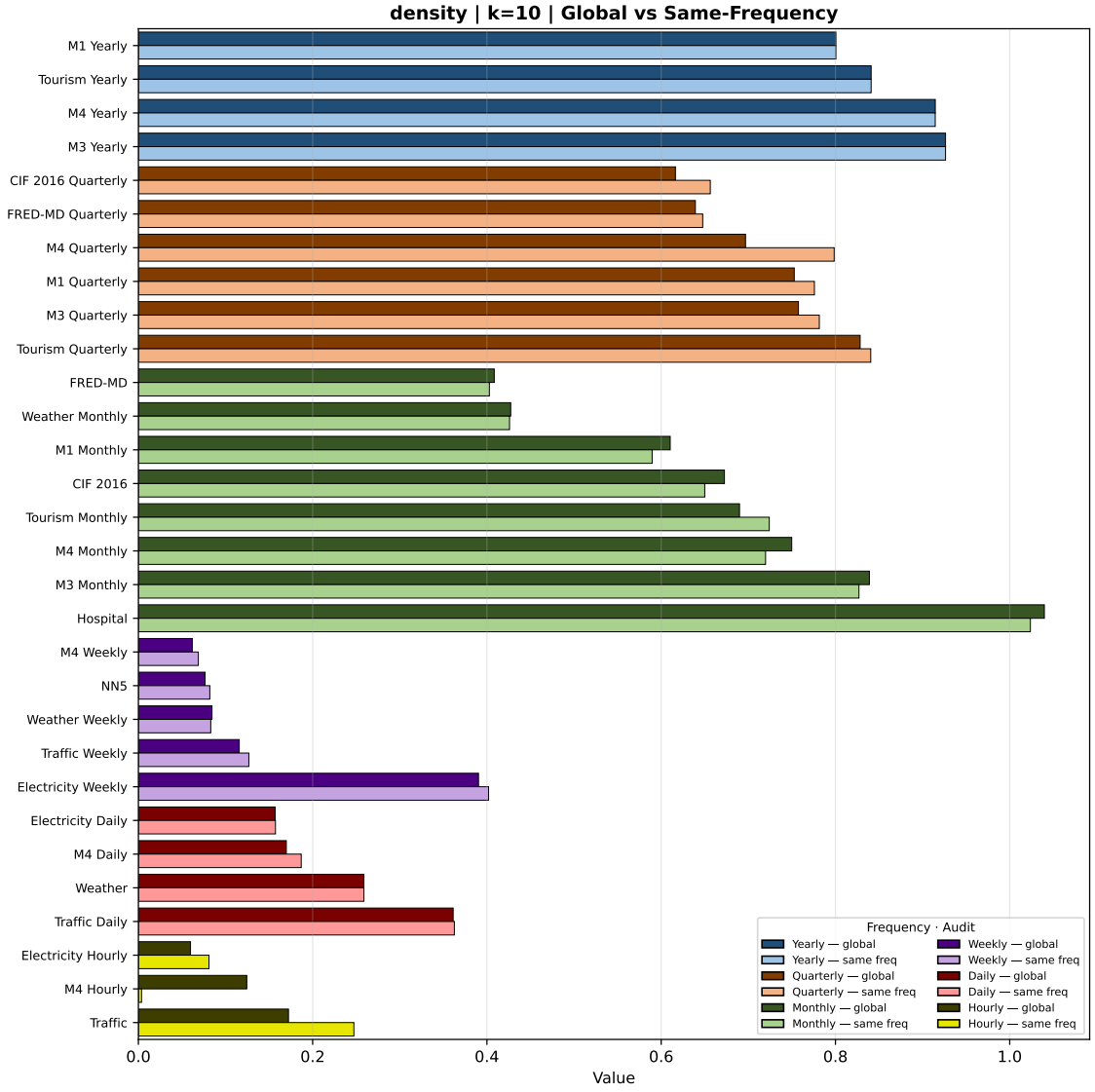


Figure B.32: Density scores per dataset at $k = 10$, ordered by frequency group.

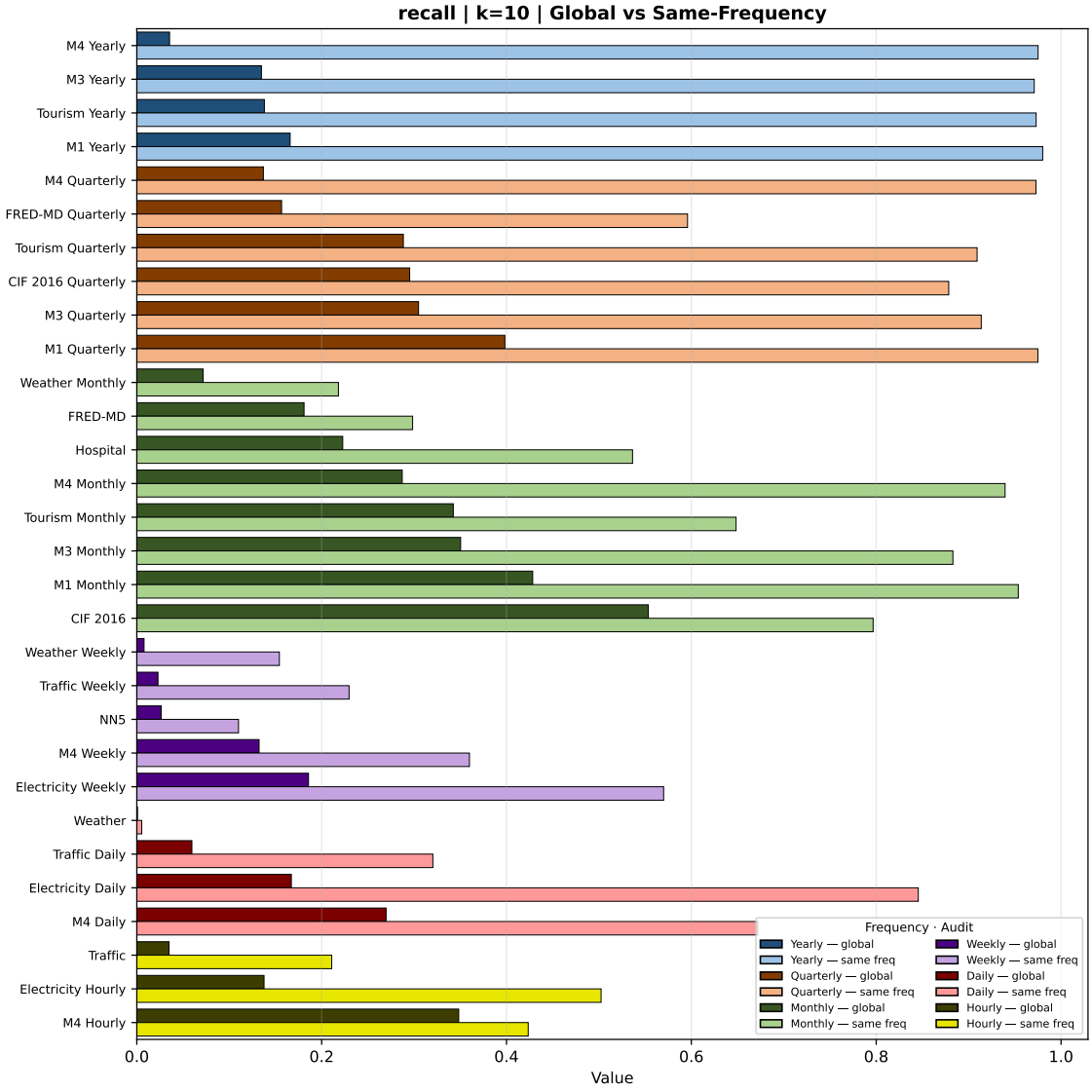


Figure B.33: Recall scores per dataset at $k = 10$, ordered by frequency group.

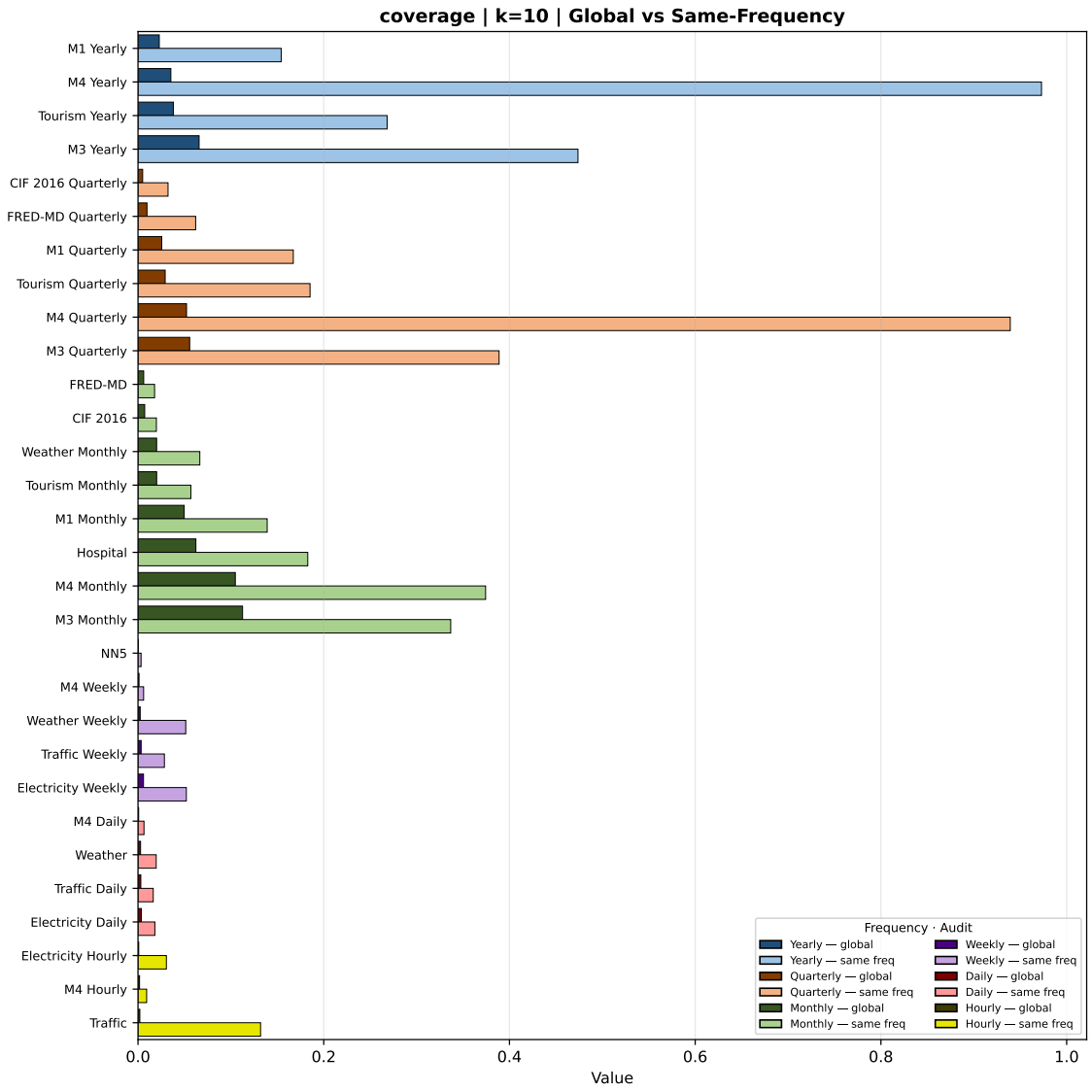


Figure B.34: Coverage scores per dataset at $k = 10$, ordered by frequency group.

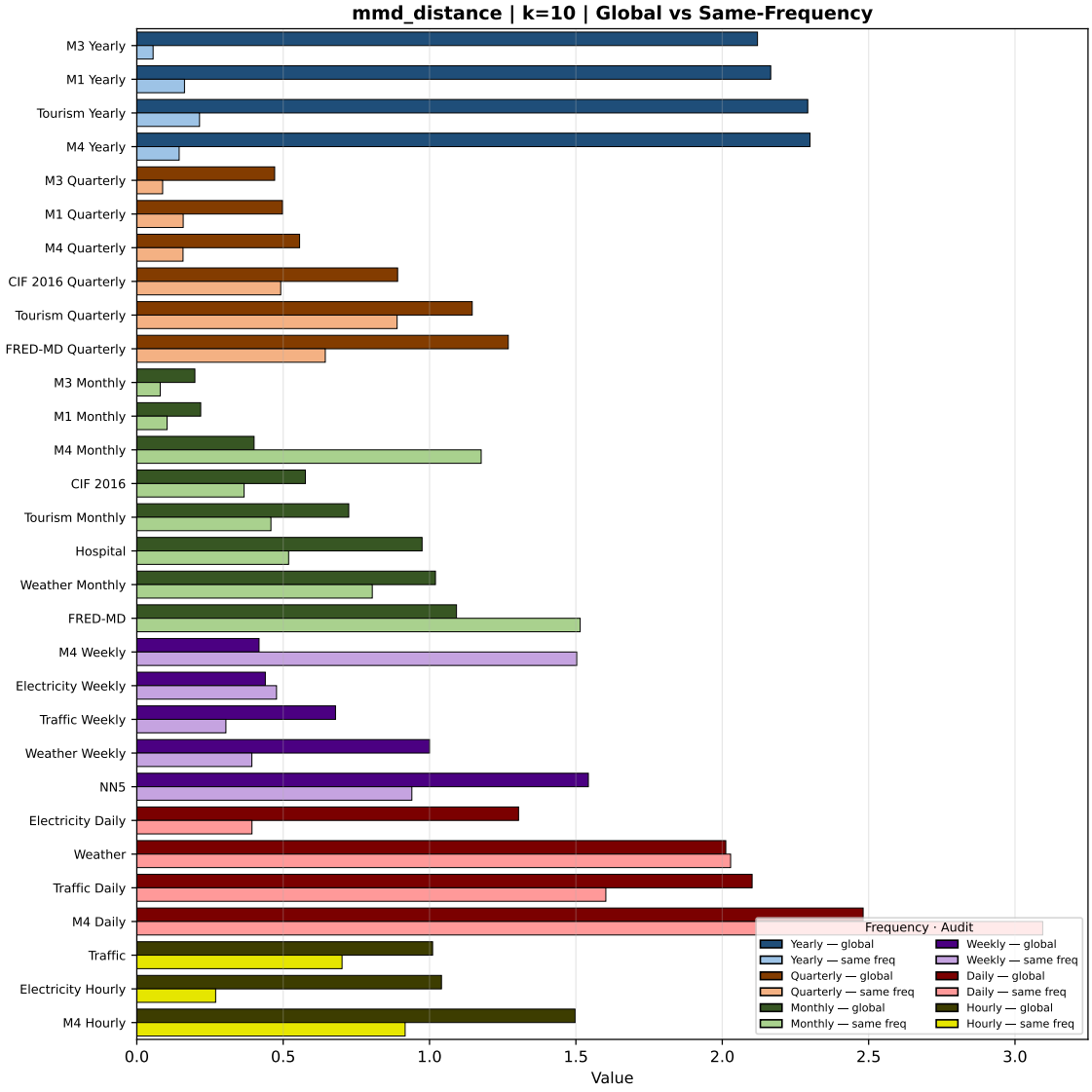


Figure B.35: MMD distance per dataset at $k = 10$, ordered by frequency group.

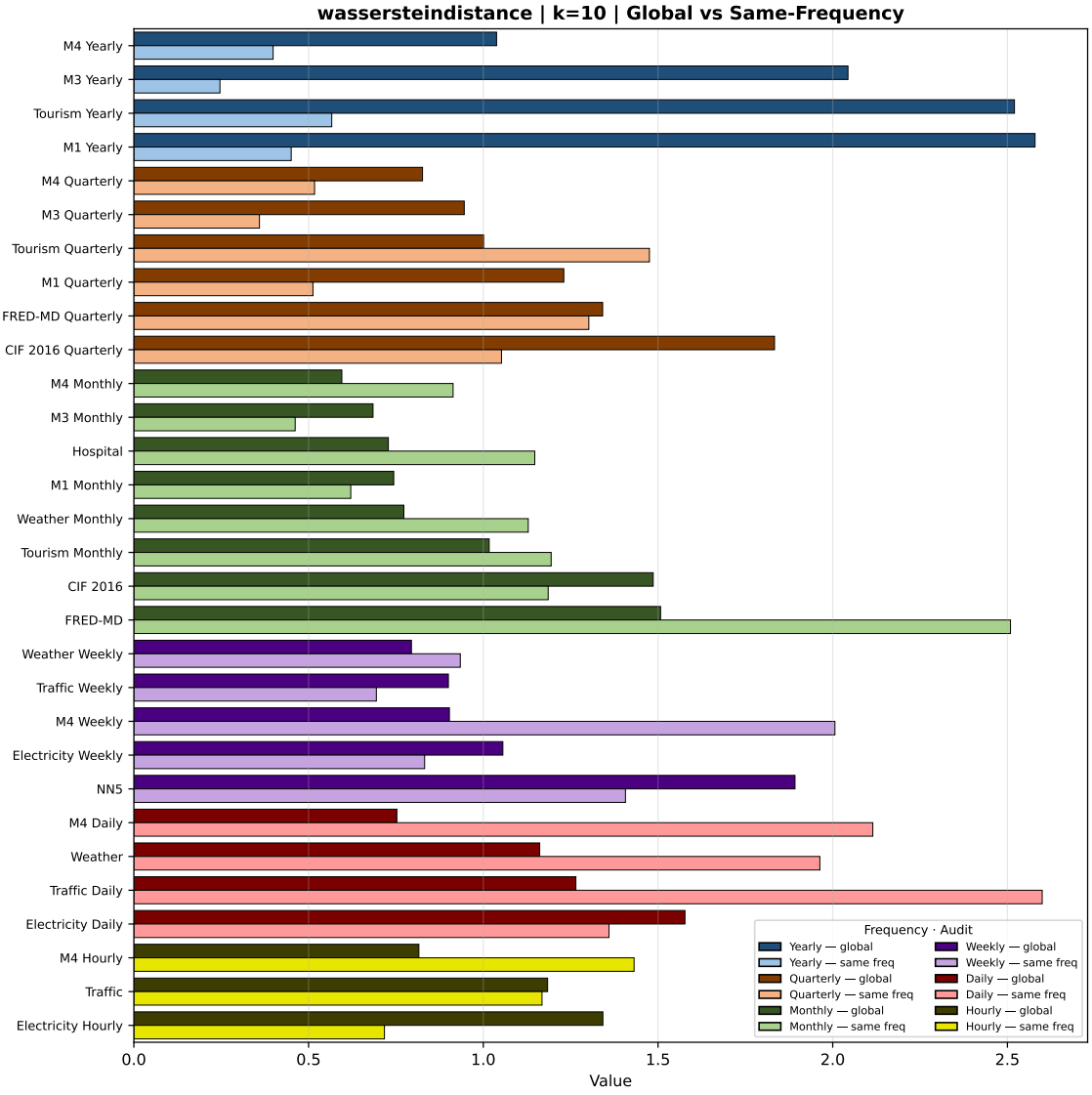


Figure B.36: Wasserstein distance per dataset at $k = 10$, ordered by frequency group.

B.3 Representativeness–Stability Scatter Plots

This appendix presents the scatter plots of representativeness metrics against rank stability metrics at $k = 10$, for both the global and same-frequency audit variants. Each figure corresponds to one representativeness metric and shows its relationship with all five stability metrics. Datasets are coloured by temporal frequency. These plots support the correlation analysis presented in Section 4.5.

Global Audit

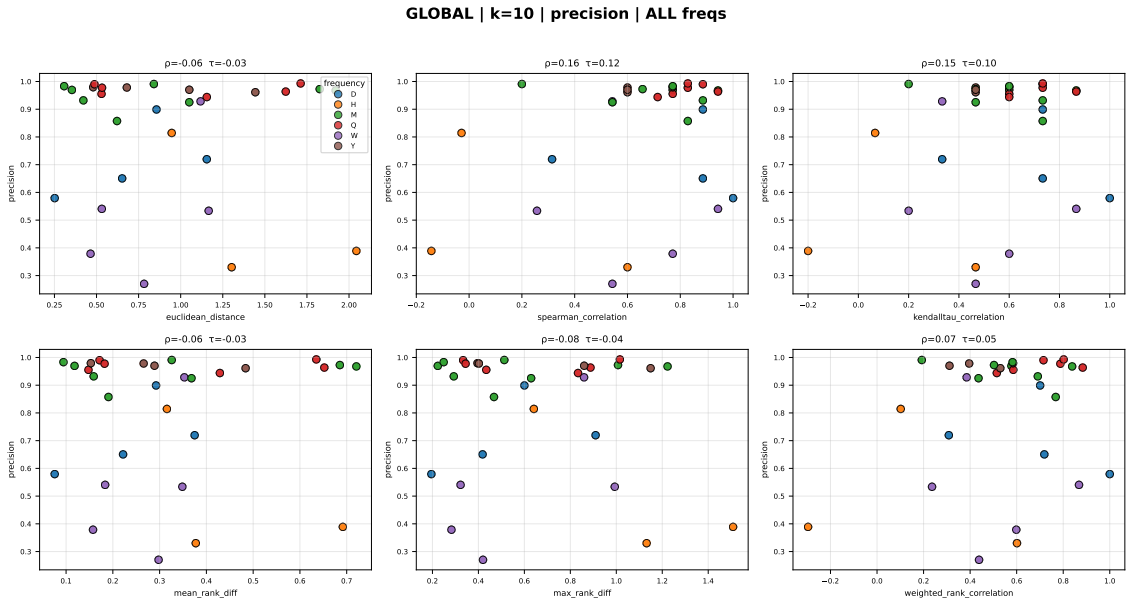


Figure B.37: Precision vs rank stability metrics — global audit, $k = 10$.

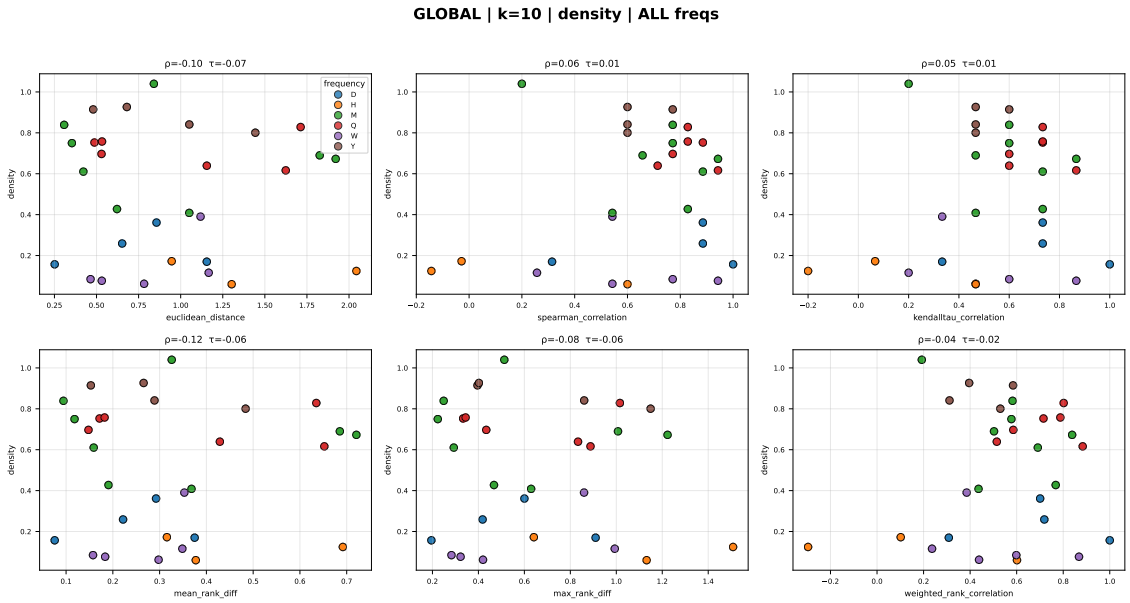


Figure B.38: Density vs rank stability metrics — global audit, $k = 10$.

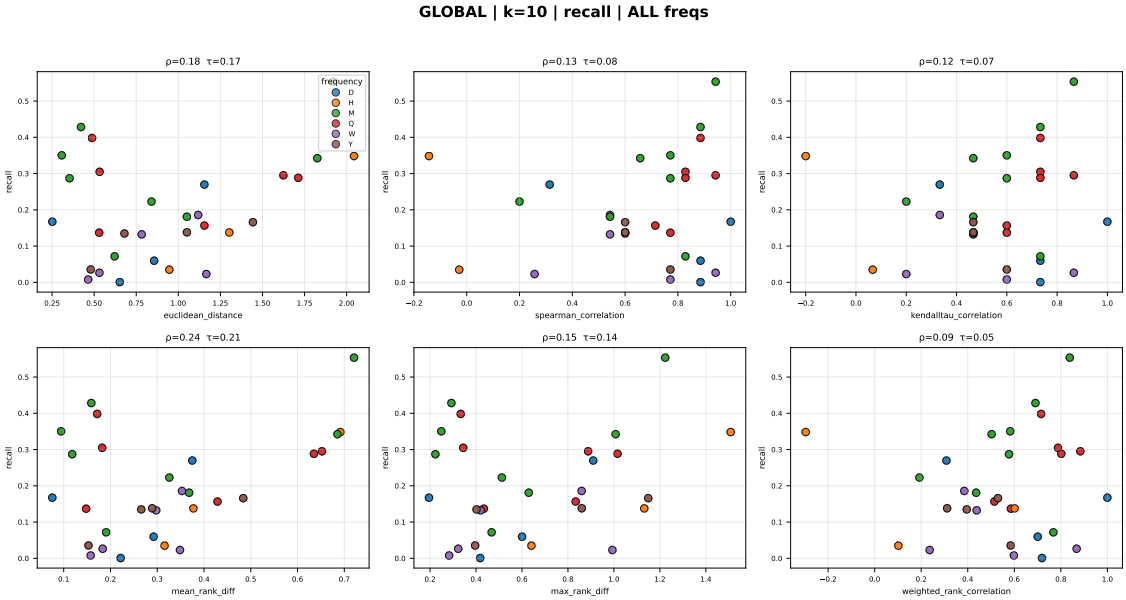


Figure B.39: Recall vs rank stability metrics — global audit, $k = 10$.

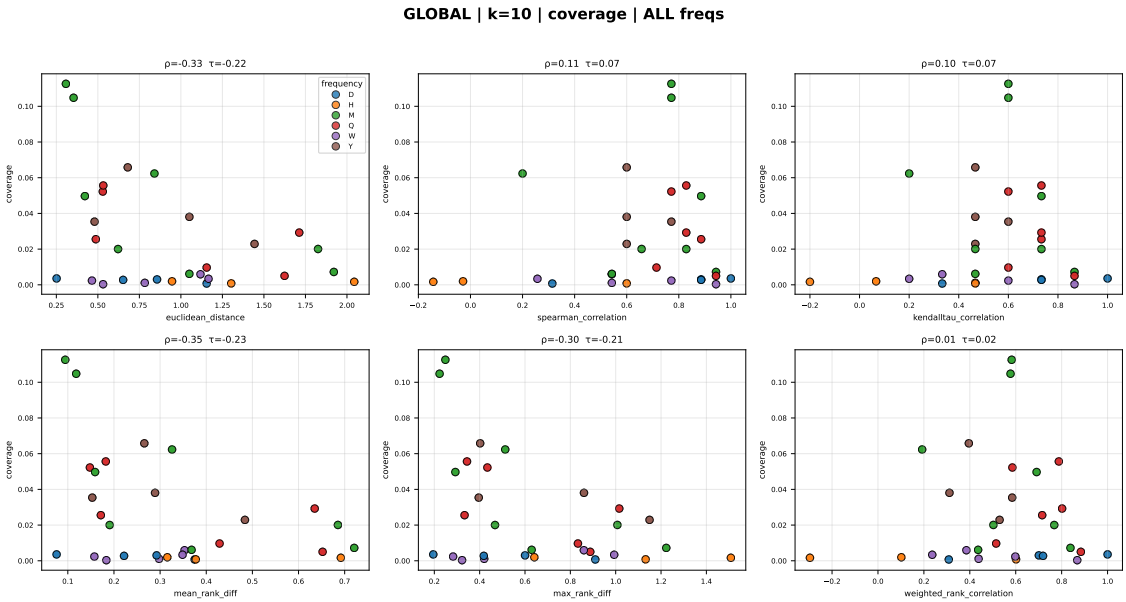


Figure B.40: Coverage vs rank stability metrics — global audit, $k = 10$.

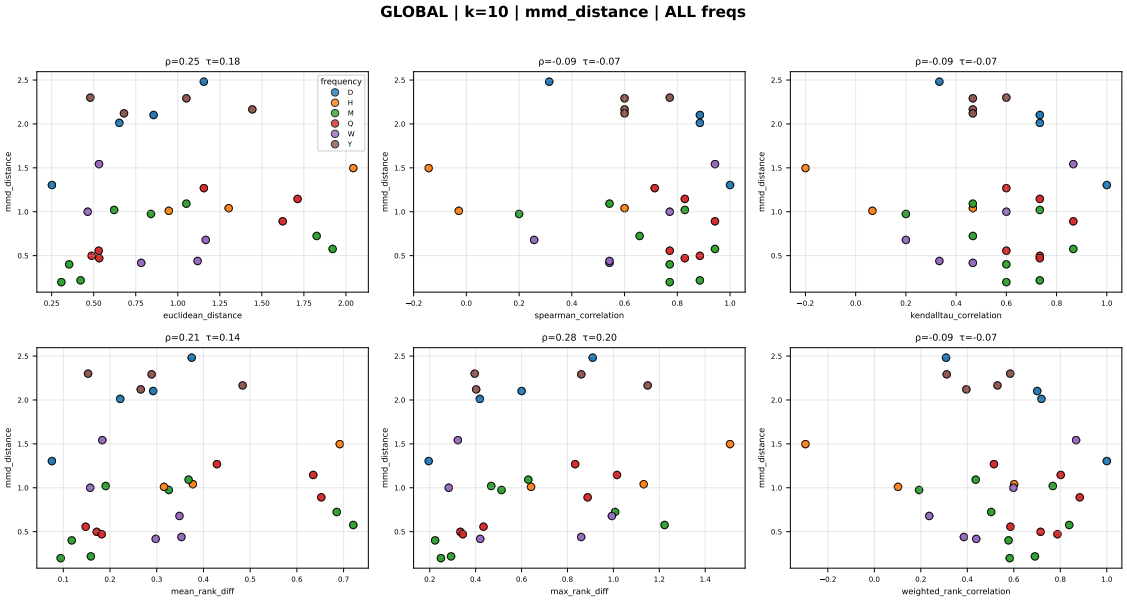


Figure B.41: MMD distance vs rank stability metrics — global audit, $k = 10$.

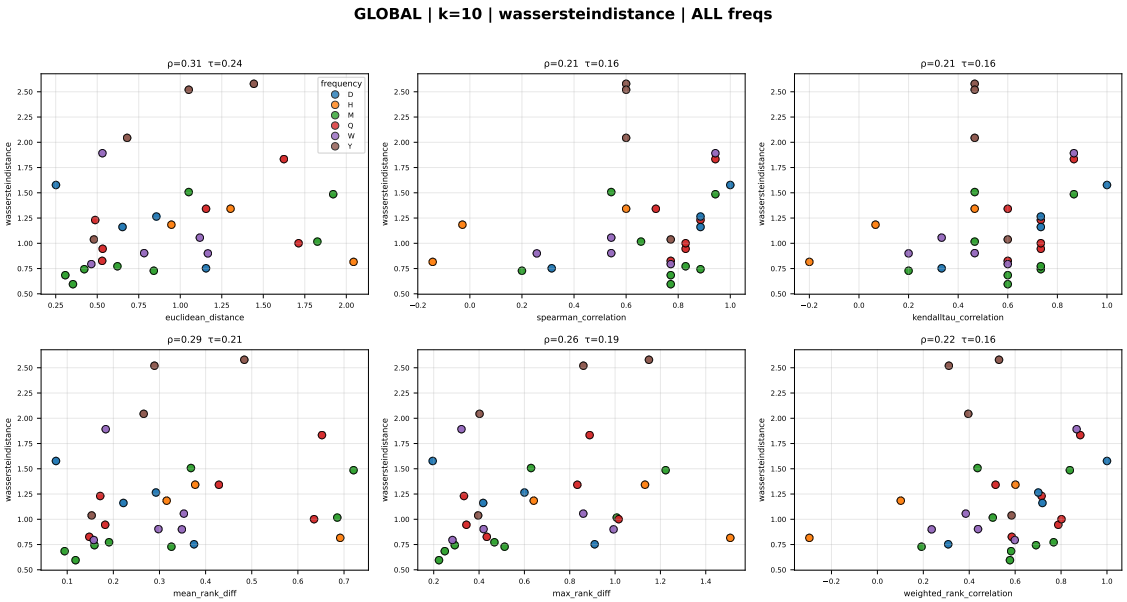


Figure B.42: Wasserstein distance vs rank stability metrics — global audit, $k = 10$.

Same-Frequency Audit

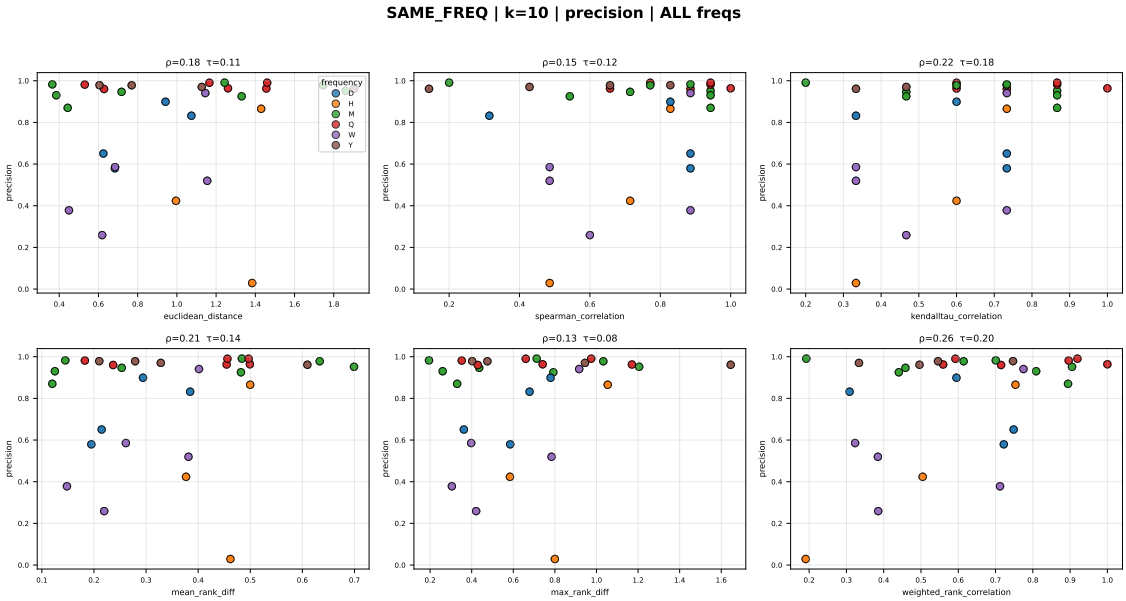


Figure B.43: Precision vs rank stability metrics — same-frequency audit, $k = 10$.

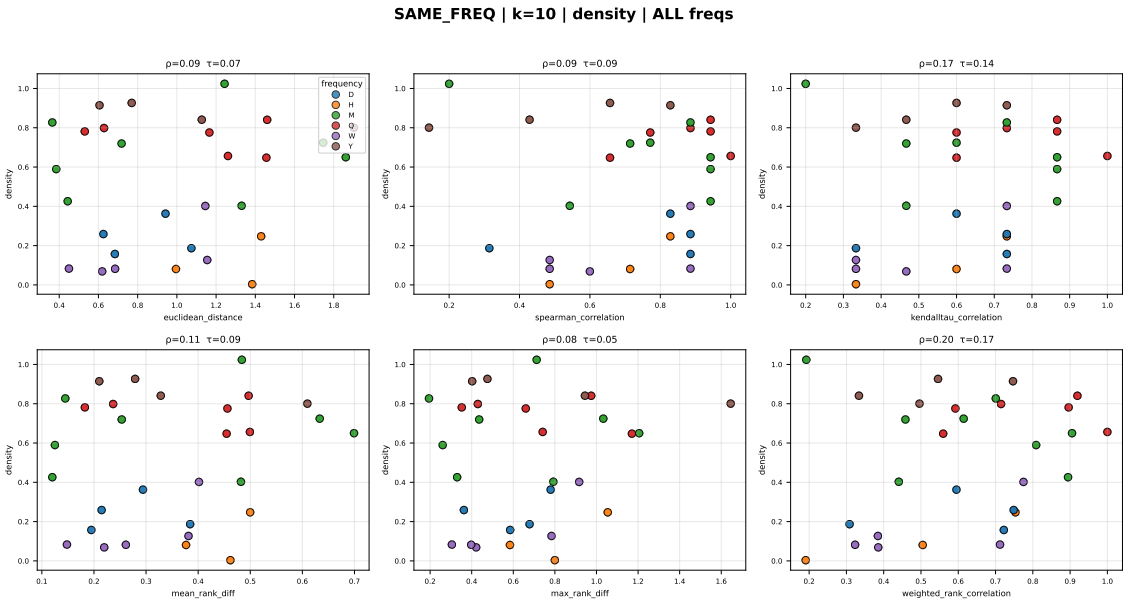


Figure B.44: Density vs rank stability metrics — same-frequency audit, $k = 10$.

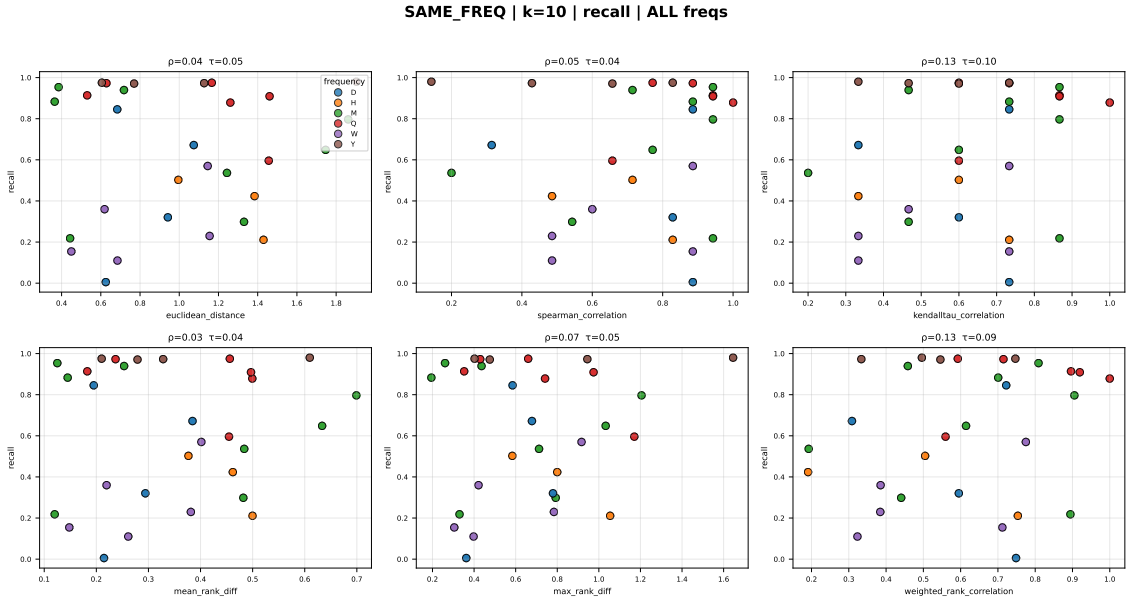


Figure B.45: Recall vs rank stability metrics — same-frequency audit, $k = 10$.

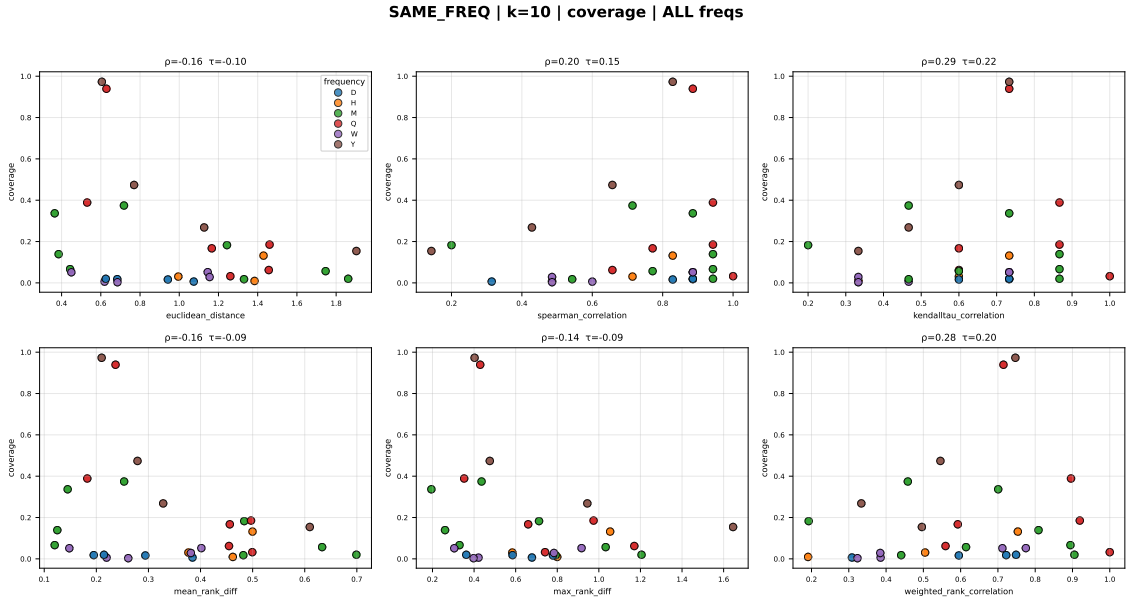


Figure B.46: Coverage vs rank stability metrics — same-frequency audit, $k = 10$.

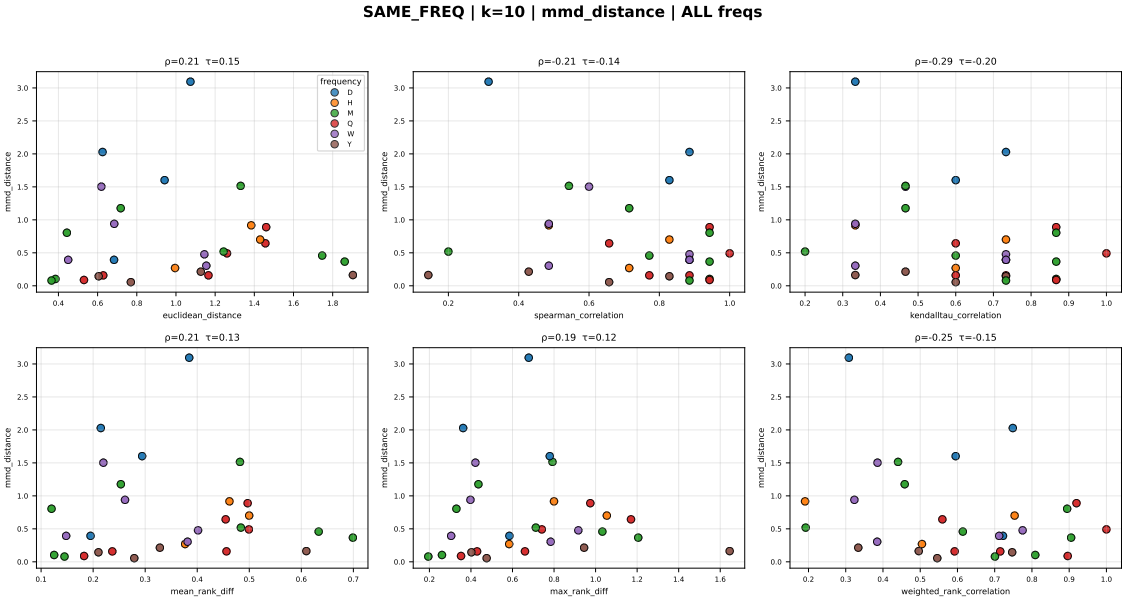


Figure B.47: MMD distance vs rank stability metrics — same-frequency audit, $k = 10$.

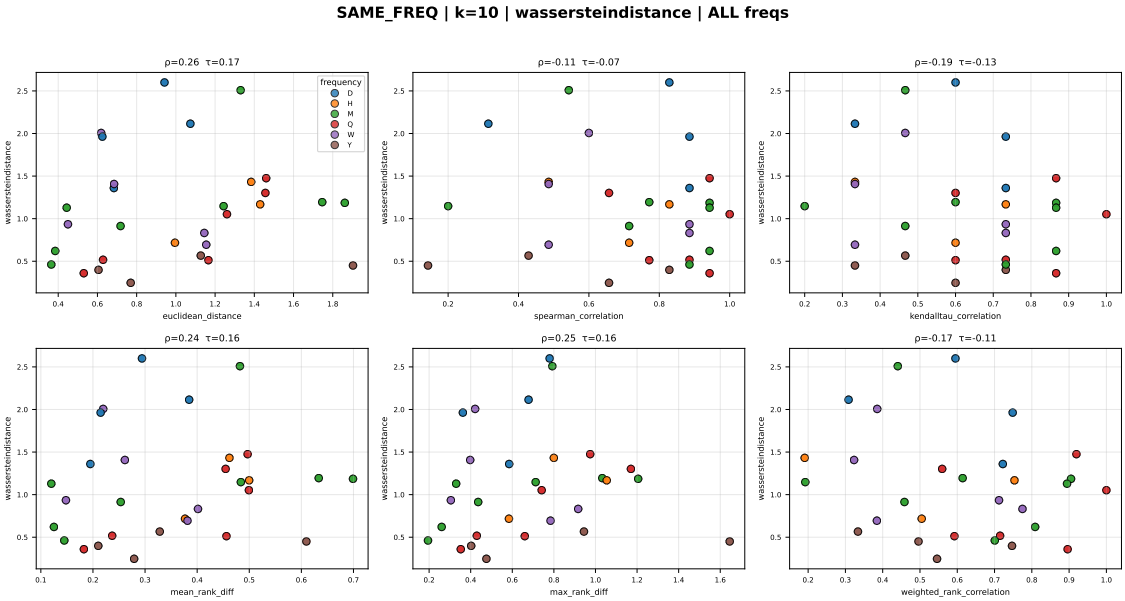


Figure B.48: Wasserstein distance vs rank stability metrics — same-frequency audit, $k = 10$.

Rank-Based Audit

This section presents the supplementary scatter plots from the rank-based feature experiment discussed in Section 4.6. As noted in the main text, because the rank-based representation forms a bounded, dense combinatorial space ($6! = 720$ possible permutations) evaluated against a massive global corpus, boundary and overlap metrics naturally saturate. Consequently, Precision, Density, and Recall yield almost flat distributions and carry minimal predictive signal in this specific

setting.

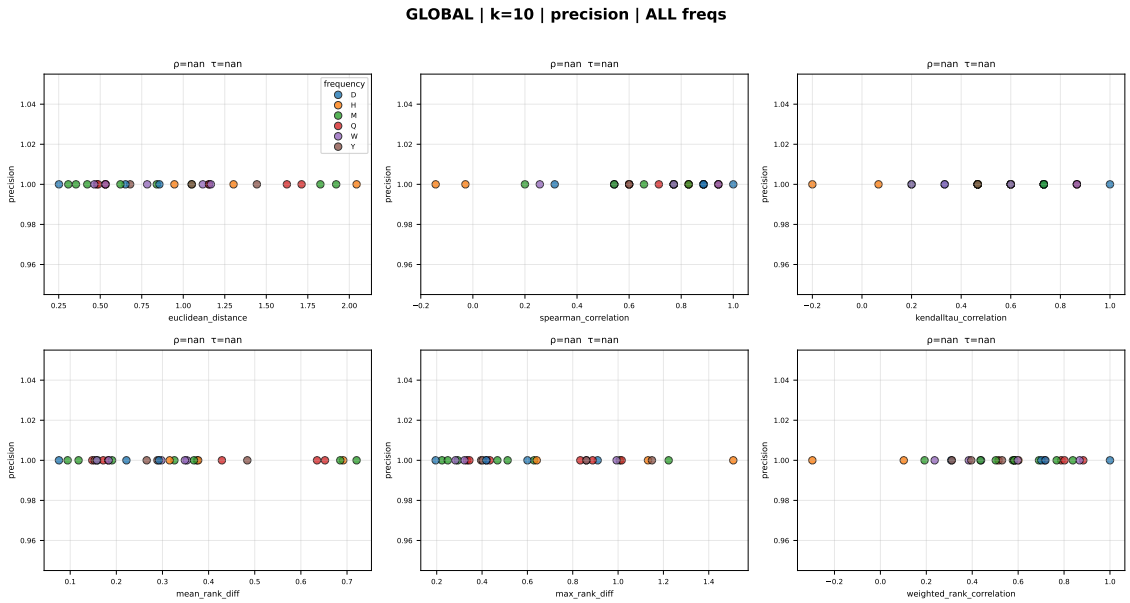


Figure B.49: Precision vs rank stability metrics — rank-based global audit, $k = 10$. The distribution is perfectly flat at 1.0, reflecting complete containment within the global combinatorial space.

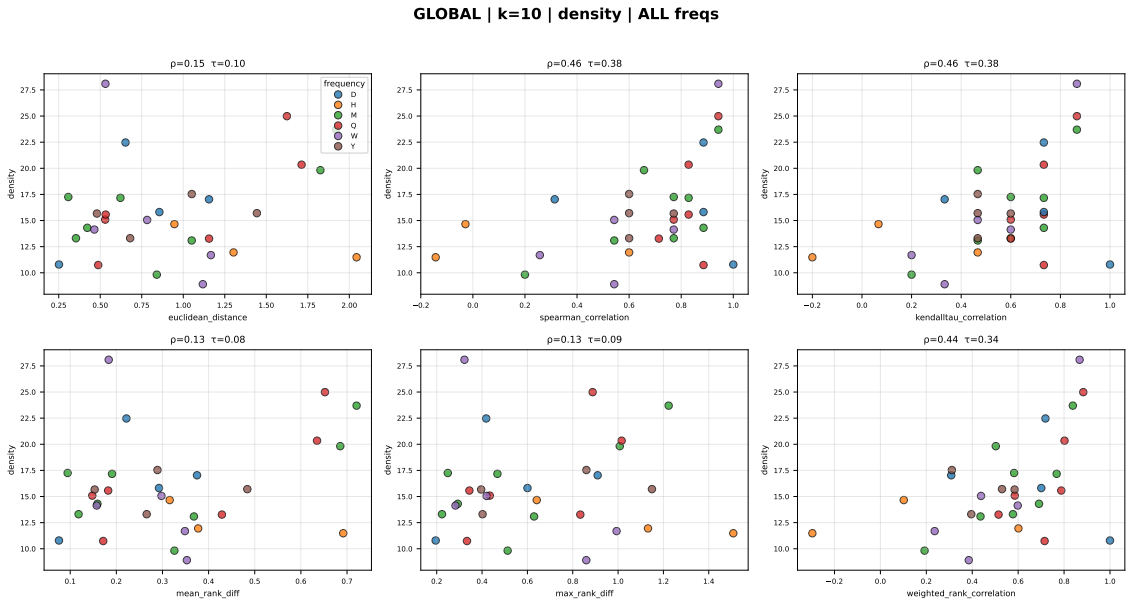


Figure B.50: Density vs rank stability metrics — rank-based global audit, $k = 10$.

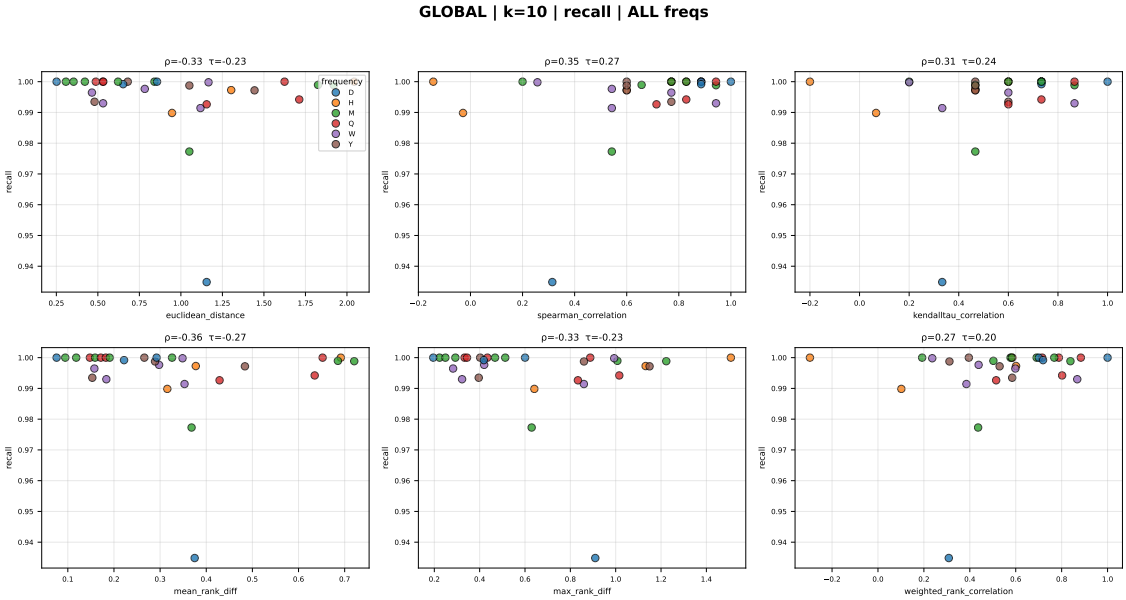


Figure B.51: Recall vs rank stability metrics — rank-based global audit, $k = 10$. Similar to Precision, Recall saturates near 1.0.

Appendix C

Transparency on the Use of AI

This appendix declares the use of Artificial Intelligence tools in the production of this dissertation, in accordance with the AI Use Transparency Policy of the University of Porto. All AI-assisted output was critically reviewed and validated, and full responsibility is assumed for the content submitted.

Tools Used

- **Claude Sonnet 4.6** (Anthropic)
- **Grammarly** (Grammarly Inc.)
- **ChatGPT** (OpenAI)
- **GitHub Copilot** (Microsoft/OpenAI)

Writing Support

Claude Sonnet 4.6 — Purpose: Grammar and style corrections, and targeted reformulation of the written text. In this work, the tool was used to rewrite selected paragraphs for clarity, to make ideas more precise and easier to follow, to improve sentence flow and word choice, and to reduce repetition and awkward phrasing. It was also used to cross-check citation keys against the `.bib` file and to flag missing or mismatched references. Each suggestion was reviewed and accepted, modified, or rejected on a case-by-case basis, and all reformulations were directed rather than delegated in substance.
Sections: All sections.

Grammarly — Purpose: Spelling correction, grammar checking, and occasional sentence rephrasing throughout the writing process. It was used as a passive writing aid integrated into the text editor, with all suggested changes reviewed and accepted or rejected individually.

Sections: All chapters.

Code Assistance

ChatGPT and GitHub Copilot — Purpose: Code generation, debugging, and documentation for the implementation of the methodology.
Sections: Chapter 3 implementation (corpus construction, feature extraction, representativeness auditing, empirical validation, data pipelines, and visualisation).
Prompts used: Functional descriptions of specific components, error messages for debugging, and requests for inline documentation.