
GIFT-Eval: A Benchmark for General Time Series Forecasting Model Evaluation

Taha Aksu¹, Gerald Woo^{1*}, Juncheng Liu¹, Xu Liu^{1,2*}, Chenghao Liu^{1†},
Silvio Savarese¹, Caiming Xiong¹, Doyen Sahoo¹
¹Salesforce AI Research, ²National University of Singapore
{iaksu, juncheng.liu, xu.liu, chenghao.liu}@salesforce.com
{ssavarese, cxiong, dsahoo}@salesforce.com
woogerald@yahoo.com.sg

Abstract

The development of time series foundation models has been constrained by the absence of comprehensive benchmarks. This paper introduces the **General Time Series Forecasting Model Evaluation**, GIFT-Eval, a pioneering benchmark specifically designed to address this gap. GIFT-Eval encompasses 23 datasets with over 144,000 time series and 157 million observations, spanning seven domains and featuring a variety of frequencies, number of variates and prediction lengths from short to long-term forecasts. Our benchmark facilitates the effective pretraining and evaluation of foundation models. We present a detailed analysis of 20 baseline models, including statistical, deep learning, and foundation models. We further provide a fine-grained analysis for each model across different characteristics of our benchmark. We hope that insights gleaned from this analysis along with the access to this new standard zero-shot time series forecasting benchmark shall guide future developments in time series forecasting foundation models. The code, data, and leaderboard are available at <https://github.com/SalesforceAIRResearch/gift-eval>.

1 Introduction

The success of foundation model pretraining in language and vision modalities has catalyzed similar progress in time series forecasting³. By pretraining on extensive time series datasets, a universal forecasting model can be developed, equipped to address varied downstream forecasting tasks across multiple domains, frequencies, prediction lengths, and variates in a zero-shot manner [42, 33, 4].

A critical aspect of foundation model research is creating a high quality benchmark that includes a large pretraining data and a diverse evaluation data to provide a fair evaluation ground and help identify weaknesses of models. Research in Natural Language Processing (NLP) has produced key benchmarks such as GLUE, MMLU, *etc.* [38, 15, 36, 5], which are crucial for developing high-quality assessments that evaluate model capabilities and highlight areas needing improvement.

Unlike NLP, time series foundation models lack a unified, diverse benchmark for fair comparison. For instance, Woo et al. [42] introduces LOTSA, which remains the largest collection of time series forecasting pre-training data to date. However, the proposed architecture *Moirai* is evaluated on existing benchmarks that are tailored to specific forecasting task, such as the LSF [46] dataset for long-term forecast, and the Monash [14] dataset for global-univariate forecasts. Both of them lack sufficient variety in terms of time series data characteristics and forecasting tasks. This issue is

*Work done during industrial PhD/internship at Salesforce AI Research.

†Corresponding author. Email: chenghao.liu@salesforce.com

³Please find the extended version of this work on <https://arxiv.org/abs/2410.10393>

also apparent in other benchmarks like TimesFM, Chronos, and Lag-Llama [8, 4, 33]. Moreover, inconsistent pretraining, train, and test splits across different models complicate comparison and risk data leakage. To effectively train and evaluate foundation models, it’s essential to establish a high-quality, balanced, and diverse set of pretraining, train, and test data that supports both in-distribution and out-of-distribution forecasting evaluations.

To fill identified gaps, we introduce the **General Time Series Forecasting Model Evaluation** (GIFT-Eval), consisting of distinct pretraining and test components. Details on the pretraining component are available in Appendix D. The test component features 23 datasets encompassing 144,000 time series and 157 million observations across 7 domains and 10 frequencies, with prediction lengths ranging from short to long-term. We also establish clear train/test splits for in-distribution model evaluation. Unlike previous benchmarks focused mainly on univariate forecasting, GIFT-Eval offers extensive multivariate data with 8 dedicated datasets. This benchmark not only facilitates new foundation work in time series by standardizing train and test splits but also eases model evaluation. Prior to our work, Qiu et al. [31] introduced TFB, a comprehensive dataset for time series forecasting. While it offered diversity in the number of variates and domains, it lacks pretraining data and foundation model evaluation. Our benchmark not only fills these gaps but it also includes a broader range of frequencies and a wider span of prediction lengths. Our contributions are 3 fold:

- **GIFT-Eval:** We introduce a new time series forecasting benchmark with pretrain/train/test splits maintaining diversity and balance across multiple characteristics.
- **Comprehensive Benchmarking:** We evaluate 20 baseline models spanning statistical, deep learning, and foundational approaches on GIFT-Eval.
- **Detailed Analysis:** We provide insights into the strengths of different models on all aspects of GIFT-Eval including domains, frequencies, prediction lengths, and the number of variates.

2 GIFT-Eval

In this section, we detail the design choices that underpin the development of GIFT-Eval. Initially, we describe our methods for selecting and processing the datasets, along with some key statistics from the final distribution of these datasets. Subsequently, we discuss the range of models employed to report results on GIFT-Eval and outline the configuration of our experimental environment. Finally, we describe the evaluation setup and highlight the user-friendly nature of our framework.

Datasets We curated test portion of GIFT-Eval with 15 univariate and 8 multivariate datasets, covering 7 domains and 10 frequencies, totaling 144,000 time series and 157 million observations. We adhere to established prediction lengths for well-known datasets like M4 [23], and for others, we establish three prediction settings—short, medium, and long—based on frequency and domain, with medium and long settings extending the short-term length by factors of 10 and 15, respectively. To support models without multivariate forecasting, our framework flattens multivariate datasets for broader compatibility. Data is stored using the Arrow format [34], ensuring efficient integration into deep learning pipelines. Detailed dataset statistics and characteristics, such as domain, frequency, and prediction lengths, are available in Appendix C and Table 7. Our benchmark features 97 unique triplets of dataset, frequency, and length, with aggregated results for each model reported across these configurations. For pretraining portion of our benchmark please check Appendix D.

Models We utilize 20 models with varied methodologies including traditional statistical models, various deep learning models, and the more recent foundation models. For statistical models, we incorporate Naive, Seasonal Naive [17], and Auto_Arima [12] methods. Representing deep learning, we select DeepAR [11], TFT [19], TiDE [7], N-BEATS [29], PatchTST [28], and iTransformer [21]. Additionally, we evaluate three foundation models zero-shot on our benchmark: TimesFM [8], Chronos [4] available in tiny, small, and base sizes, and Moirai [42] available in small, base, and large sizes. More details with full list of models and model-specific hyperparameters can be found in Appendix B.

Evaluation setting We structure the evaluation component of our benchmark by dedicating the final 10% of each dataset to testing, with the rest allocated for training. A non-overlapping rolling evaluation method is employed, setting a predetermined number of windows in the test split, each

Table 1: Results on GIFT-Eval aggregated by domain. The best results across each row are **bolded**, while the second best results are underlined.

| Domain | Metric | Nv. | S.Nv. | A.Ar. | A.Th. | D.AR | TFT | TiDE | N-B. | P.TST | iTr. | T.FM | V.TS | Chr.s | Chr.B | Chr.L | Moi.s | Moi.B | Moi.L | Best |
|--------------|--------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------------|---------------------|---------------------------|---------------------|---------------------------|---------------------------|---------------------------|---------------------|---------------------------|---------------------------|-------|
| EconFin | MASE | 1.43 | 1.00 | 8.66e ⁻¹ | 9.83e ⁻¹ | 1.54 | 1.03 | 1.51 | 8.61e ⁻¹ | 9.08e ⁻¹ | 9.80e ⁻¹ | 8.24e ⁻¹ | 9.31e ⁻¹ | 7.97e ⁻¹ | 7.83e⁻¹ | 7.83e ⁻¹ | 1.04 | 9.27e ⁻¹ | 9.63e ⁻¹ | Chr.L |
| | CRPS | 1.17 | 1.00 | 8.21e ⁻¹ | 8.41e ⁻¹ | 1.22 | 8.41e ⁻¹ | 1.08 | 9.67e ⁻¹ | 8.03e ⁻¹ | 8.48e ⁻¹ | 7.16e⁻¹ | 1.05 | 7.63e ⁻¹ | <u>7.51e⁻¹</u> | 7.58e ⁻¹ | 7.96e ⁻¹ | 8.16e ⁻¹ | 8.47e ⁻¹ | T.FM |
| | Rank | 1.90e ⁻¹ | 1.88e ⁻¹ | 9.83 | 1.07e ⁻¹ | 1.88e ⁻¹ | 1.10e ⁻¹ | 2.12e ⁻¹ | 1.62e ⁻¹ | 9.17 | 1.15e ⁻¹ | 6.67 | 2.03e ⁻¹ | 9.50 | 8.67 | 9.00 | 1.00e ⁻¹ | 7.00 | 6.50 | 7.00 |
| Energy | MASE | 1.56 | 1.00 | 1.01 | 1.36 | 1.78 | 1.01 | 1.17 | 1.18 | 9.83e ⁻¹ | 1.11 | 1.02 | 9.93e ⁻¹ | 9.47e ⁻¹ | <u>9.24e⁻¹</u> | 9.19e⁻¹ | 1.04 | 9.87e ⁻¹ | 1.03 | Chr.L |
| | CRPS | 1.53 | 1.00 | 8.33e ⁻¹ | 1.70 | 1.07 | 6.30e ⁻¹ | 7.51e ⁻¹ | 9.35e ⁻¹ | 6.12e⁻¹ | 6.95e ⁻¹ | 6.73e ⁻¹ | 7.82e ⁻¹ | 6.48e ⁻¹ | 6.31e ⁻¹ | 6.28e ⁻¹ | 6.68e ⁻¹ | <u>6.15e⁻¹</u> | 6.27e ⁻¹ | P.TST |
| | Rank | 2.51e ⁻¹ | 2.13e ⁻¹ | 1.67e ⁻¹ | 2.32e ⁻¹ | 2.00e ⁻¹ | 9.56 | 1.41e ⁻¹ | 2.01e ⁻¹ | 7.69 | 9.44 | 1.00e ⁻¹ | 1.75e ⁻¹ | 1.12e ⁻¹ | 9.28 | 9.19 | 9.71 | 6.66 | 7.56 | Moi.B |
| Healthcare | MASE | 1.16 | 1.00 | 7.84e ⁻¹ | 9.51e ⁻¹ | 7.65e ⁻¹ | 6.60e ⁻¹ | 8.03e ⁻¹ | 6.91e ⁻¹ | 6.86e ⁻¹ | 7.74e ⁻¹ | 6.98e ⁻¹ | 7.49e ⁻¹ | <u>6.07e⁻¹</u> | 6.45e ⁻¹ | 5.99e⁻¹ | 9.51e ⁻¹ | 6.75e ⁻¹ | 6.91e ⁻¹ | Chr.L |
| | CRPS | 1.19 | 1.00 | 5.70e ⁻¹ | 8.03e ⁻¹ | 7.23e ⁻¹ | 5.12e ⁻¹ | 9.12e ⁻¹ | 7.13e ⁻¹ | 5.76e ⁻¹ | 6.28e ⁻¹ | 6.52e ⁻¹ | 6.81e ⁻¹ | 4.96e ⁻¹ | 4.85e ⁻¹ | 4.46e⁻¹ | 7.72e ⁻¹ | 5.14e ⁻¹ | 5.28e ⁻¹ | Chr.L |
| | Rank | 2.26e ⁻¹ | 1.98e ⁻¹ | 9.60 | 1.52e ⁻¹ | 1.26e ⁻¹ | 9.40 | 1.74e ⁻¹ | 1.72e ⁻¹ | 1.06e ⁻¹ | 1.26e ⁻¹ | 9.60 | 1.60e ⁻¹ | 7.00 | 6.00 | 6.00 | 1.63e ⁻¹ | 5.80 | 7.20 | Chr.L |
| Nature | MASE | 9.92e ⁻¹ | 1.00 | 1.02 | 1.06 | 1.64 | 8.71e ⁻¹ | 1.37 | 9.33e ⁻¹ | 9.16e ⁻¹ | 8.51e ⁻¹ | 8.80e ⁻¹ | 8.60e ⁻¹ | 8.51e ⁻¹ | 8.23e ⁻¹ | 8.13e ⁻¹ | 7.97e ⁻¹ | <u>7.83e⁻¹</u> | 7.56e ⁻¹ | Moi.L |
| | CRPS | 1.33 | 1.00 | 6.58e ⁻¹ | 9.10e ⁻¹ | 5.35e ⁻¹ | 3.48e ⁻¹ | 5.61e ⁻¹ | 5.32e ⁻¹ | 3.47e ⁻¹ | 3.42e ⁻¹ | 3.33e ⁻¹ | 4.06e ⁻¹ | 3.83e ⁻¹ | 3.66e ⁻¹ | 3.64e ⁻¹ | 3.73e ⁻¹ | 3.15e ⁻¹ | 3.11e ⁻¹ | Moi.L |
| | Rank | 2.75e ⁻¹ | 2.67e ⁻¹ | 2.11e ⁻¹ | 2.37e ⁻¹ | 1.73e ⁻¹ | 1.05e ⁻¹ | 1.93e ⁻¹ | 2.13e ⁻¹ | 1.03e ⁻¹ | 8.93 | 8.13 | 1.65e ⁻¹ | 1.40e ⁻¹ | 1.29e ⁻¹ | 1.23e ⁻¹ | 9.21 | 5.20 | 5.27 | Moi.B |
| Sales | MASE | 1.00 | 1.00 | 8.13e ⁻¹ | 8.73e ⁻¹ | 7.07e ⁻¹ | 7.16e ⁻¹ | 9.81e ⁻¹ | 7.04e ⁻¹ | 6.90e⁻¹ | 6.99e ⁻¹ | 7.00e ⁻¹ | 8.17e ⁻¹ | 7.33e ⁻¹ | 7.26e ⁻¹ | 7.24e ⁻¹ | 7.31e ⁻¹ | <u>6.95e⁻¹</u> | 7.10e ⁻¹ | P.TST |
| | CRPS | 8.96e ⁻¹ | 1.00 | 4.58e ⁻¹ | 4.80e ⁻¹ | 3.52e ⁻¹ | 3.52e ⁻¹ | 4.84e ⁻¹ | 4.14e ⁻¹ | 3.48e ⁻¹ | 3.51e ⁻¹ | 3.44e⁻¹ | 4.92e ⁻¹ | 3.66e ⁻¹ | 3.63e ⁻¹ | 3.62e ⁻¹ | 3.61e ⁻¹ | 3.47e ⁻¹ | 3.63e ⁻¹ | T.FM |
| | Rank | 2.80e ⁻¹ | 2.80e ⁻¹ | 1.98e ⁻¹ | 2.10e ⁻¹ | 8.75 | 1.10e ⁻¹ | 2.05e ⁻¹ | 1.45e ⁻¹ | 5.00 | 7.00 | 3.00 | 2.15e ⁻¹ | 1.22e ⁻¹ | 1.05e ⁻¹ | 1.00e ⁻¹ | 1.00e ⁻¹ | 3.25 | 6.75 | T.FM |
| Transport | MASE | 1.26 | 1.00 | 9.74e ⁻¹ | 1.08 | 7.45e ⁻¹ | 6.79e ⁻¹ | 7.90e ⁻¹ | 7.31e ⁻¹ | 7.09e ⁻¹ | 7.07e ⁻¹ | 7.41e ⁻¹ | 7.39e ⁻¹ | 7.37e ⁻¹ | 7.12e ⁻¹ | 7.14e ⁻¹ | 7.26e ⁻¹ | <u>6.34e⁻¹</u> | 6.07e⁻¹ | Moi.L |
| | CRPS | 2.07 | 1.00 | 7.63e ⁻¹ | 1.33 | 4.84e ⁻¹ | 4.43e ⁻¹ | 5.31e ⁻¹ | 5.93e ⁻¹ | 4.61e ⁻¹ | 4.60e ⁻¹ | 5.10e ⁻¹ | 6.01e ⁻¹ | 5.30e ⁻¹ | 5.12e ⁻¹ | 5.12e ⁻¹ | 4.98e ⁻¹ | <u>4.12e⁻¹</u> | 3.93e⁻¹ | Moi.L |
| | Rank | 2.84e ⁻¹ | 2.43e ⁻¹ | 2.18e ⁻¹ | 2.61e ⁻¹ | 8.73 | 6.60 | 1.39e ⁻¹ | 1.74e ⁻¹ | 8.07 | 7.93 | 1.06e ⁻¹ | 1.81e ⁻¹ | 1.39e ⁻¹ | 1.08e ⁻¹ | 1.11e ⁻¹ | 1.07e ⁻¹ | 5.40 | 5.67 | Moi.B |
| Web/CloudOps | MASE | 1.13 | 1.00 | 9.57e ⁻¹ | 5.21e ⁻¹ | 8.50e ⁻¹ | 6.62e ⁻¹ | 6.23e ⁻¹ | 5.43e ⁻¹ | 4.62e⁻¹ | 4.88e ⁻¹ | 1.42 | 4.72e ⁻¹ | 6.78e ⁻¹ | 6.76e ⁻¹ | 6.75e ⁻¹ | 7.73e ⁻¹ | 7.02e ⁻¹ | 6.79e ⁻¹ | P.TST |
| | CRPS | 1.07 | 1.00 | 9.04e ⁻¹ | 6.08e ⁻¹ | 6.33e ⁻¹ | 5.03e ⁻¹ | 5.68e ⁻¹ | 5.70e ⁻¹ | 4.37e⁻¹ | 4.54e ⁻¹ | 7.39e ⁻¹ | 6.03e ⁻¹ | 6.29e ⁻¹ | 6.51e ⁻¹ | 6.47e ⁻¹ | 6.49e ⁻¹ | 6.28e ⁻¹ | 6.19e ⁻¹ | P.TST |
| | Rank | 2.19e ⁻¹ | 2.18e ⁻¹ | 1.99e ⁻¹ | 1.66e ⁻¹ | 1.48e ⁻¹ | 6.95 | 1.22e ⁻¹ | 1.29e ⁻¹ | 4.75 | 5.85 | 1.84e ⁻¹ | 1.35e ⁻¹ | 1.29e ⁻¹ | 1.45e ⁻¹ | 1.48e ⁻¹ | 1.35e ⁻¹ | 1.22e ⁻¹ | 1.13e ⁻¹ | P.TST |

Table 2: Results on GIFT-Eval aggregated by Prediction Length. The best results across each row are **bolded**, while the second best results are underlined.

| Pred. Len. | Metric | Nv. | S.Nv. | A.Ar. | A.Th. | D.AR | TFT | TiDE | N-B. | P.TST | iTr. | T.FM | V.TS | Chr.s | Chr.B | Chr.L | Moi.s | Moi.B | Moi.L | Best |
|------------|--------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------------|---------------------|---------------------|---------------------------|---------------------|---------------------|---------------------------|---------------------|---------------------------|---------------------------|---------------------|---------------------|---------------------|-------|
| Long | MASE | 1.40 | 1.00 | 9.85e ⁻¹ | 8.69e ⁻¹ | 1.10 | 5.89e ⁻¹ | 6.55e ⁻¹ | 6.44e ⁻¹ | <u>5.37e⁻¹</u> | 5.66e ⁻¹ | 9.90e ⁻¹ | 5.22e⁻¹ | 6.58e ⁻¹ | 6.34e ⁻¹ | 6.32e ⁻¹ | 6.44e ⁻¹ | 6.25e ⁻¹ | 6.04e ⁻¹ | V.TS |
| | CRPS | 1.89 | 1.00 | 8.05e ⁻¹ | 1.40 | 6.28e ⁻¹ | 3.79e⁻¹ | 4.48e ⁻¹ | 5.05e ⁻¹ | 3.91e ⁻¹ | 5.18e ⁻¹ | 4.56e ⁻¹ | 5.22e ⁻¹ | 5.04e ⁻¹ | 5.02e ⁻¹ | 4.45e ⁻¹ | 4.23e ⁻¹ | 4.22e ⁻¹ | 4.22e ⁻¹ | P.TST |
| | Rank | 2.72e ⁻¹ | 2.31e ⁻¹ | 2.09e ⁻¹ | 2.43e ⁻¹ | 1.72e ⁻¹ | 6.48 | 1.16e ⁻¹ | 1.61e ⁻¹ | 6.00 | 7.19 | 1.51e ⁻¹ | 1.26e ⁻¹ | 1.36e ⁻¹ | 1.40e ⁻¹ | 1.44e ⁻¹ | 9.29 | 8.24 | 8.19 | P.TST |
| Medium | MASE | 1.46 | 1.00 | 1.02 | 1.17 | 1.33 | 9.49e ⁻¹ | 9.86e ⁻¹ | 1.03 | <u>8.20e⁻¹</u> | 8.97e ⁻¹ | 1.44 | 8.47e⁻¹ | 1.04 | 1.04 | 1.03 | 1.03 | 1.03 | 9.72e ⁻¹ | V.TS |
| | CRPS | 1.87 | 1.00 | 8.33e ⁻¹ | 1.53 | 6.40e ⁻¹ | 4.68e ⁻¹ | 5.69e ⁻¹ | 6.78e ⁻¹ | 4.61e ⁻¹ | 4.70e ⁻¹ | 6.30e ⁻¹ | 5.89e ⁻¹ | 6.25e ⁻¹ | 6.30e ⁻¹ | 6.22e ⁻¹ | 5.55e ⁻¹ | 5.35e ⁻¹ | 5.23e ⁻¹ | P.TST |
| | Rank | 2.62e ⁻¹ | 2.16e ⁻¹ | 1.99e ⁻¹ | 2.43e ⁻¹ | 1.36e ⁻¹ | 5.90 | 1.24e ⁻¹ | 1.70e ⁻¹ | 5.14 | 5.71 | 1.41e ⁻¹ | 1.41e ⁻¹ | 1.50e ⁻¹ | 1.50e ⁻¹ | 1.42e ⁻¹ | 1.00e ⁻¹ | 8.86 | 8.62 | P.TST |
| Short | MASE | 1.14 | 1.00 | 9.35e ⁻¹ | 9.55e ⁻¹ | 1.20 | 8.83e ⁻¹ | 1.14 | 8.62e ⁻¹ | 8.32e ⁻¹ | 8.89e ⁻¹ | 8.23e ⁻¹ | 8.71e ⁻¹ | 7.79e ⁻¹ | <u>7.68e⁻¹</u> | 7.61e⁻¹ | 8.97e ⁻¹ | 8.19e ⁻¹ | 8.21e ⁻¹ | Chr.L |
| | CRPS | 1.09 | 1.00 | 7.35e ⁻¹ | 8.16e ⁻¹ | 7.95e ⁻¹ | 5.92e ⁻¹ | 7.95e ⁻¹ | 7.48e ⁻¹ | 5.71e ⁻¹ | 6.11e ⁻¹ | 5.77e ⁻¹ | 7.51e ⁻¹ | 5.52e ⁻¹ | <u>5.42e⁻¹</u> | 5.38e⁻¹ | 6.09e ⁻¹ | 5.48e ⁻¹ | 5.53e ⁻¹ | Chr.L |
| | Rank | 2.36e ⁻¹ | 2.31e ⁻¹ | 1.64e ⁻¹ | 1.86e ⁻¹ | 1.62e ⁻¹ | 1.09e ⁻¹ | 1.79e ⁻¹ | 1.87e ⁻¹ | 9.27 | 1.02e ⁻¹ | 8.80 | 1.96e ⁻¹ | 9.65 | 8.33 | 8.33 | 1.14e ⁻¹ | 6.18 | <u>6.93</u> | Moi.B |

equal to the dataset’s prediction length. The final window of the training data serves as validation for tuning deep learning model hyperparameters.

Performance is assessed using two primary metrics: the median Mean Absolute Percentage Error (MAPE) for point forecasts and the Continuous Ranked Probability Score (CRPS) [13] for probabilistic forecasts. To standardize comparison across benchmarks, both MAPE and CRPS are normalized against the seasonal naive model as a baseline. To avoid skew from any single dataset, we employ a ‘Rank’ metric that assigns a numerical ranking to each model across all 97 configurations. The average of these ranks is then reported as the final result for each model.

3 Results

The first four parts in results section aggregate the results by the key characteristics that guided the development of our benchmark: domain, prediction length, frequency, and number of variates, then conclude the section with aggregation of results across all configurations. For results on all dataset, frequency and prediction length combinations with more metrics see Appendix E.

Domain | Table 1 The results across various domains, illustrate that foundation models consistently outperform both statistical and deep learning models. In particular foundation models exhibit top performance with the exception of Web/CloudOps domain. Deep learning models like PatchTST and iTransformer show stronger results in this domain. This suggests a potential lack of representative data for foundation models in this specific domain.

Prediction length | Table 2 Prediction length reveals a distinct performance variance across settings. For short-term forecasts, most foundation models, specifically Moirai variants, outperform other models. However, as the prediction length increases, deep learning models like PatchTST and iTransformer begin to surpass foundation models, indicating fine-tuning’s effect in handling longer-term dependencies within the data.

Frequency | Table 3 Deep learning and statistical models demonstrate strong performances at higher frequencies (secondly and minutely granularities), achieving the best results. Conversely, foundation models, particularly the Moirai variants, consistently outperform others at lower frequencies, dominating in 6 different settings. This robustness in less granular time series may suggest that such frequencies exhibit more common patterns that are easier to learn, thus amplifying the benefits derived from pretraining objectives.

References

- [1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.
- [2] Alexander Alexandrov, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C. Maddix, Syama Rangapuram, David Salinas, Jasper Schulz, Lorenzo Stella, Ali Caner Türkmen, and Yuyang Wang. Gluonts: Probabilistic and neural time series modeling in python. *Journal of Machine Learning Research*, 21(116): 1–6, 2020. URL <http://jmlr.org/papers/v21/19-820.html>.
- [3] Alexander Alexandrov, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C. Maddix, Syama Rangapuram, David Salinas, Jasper Schulz, Lorenzo Stella, Ali Caner Türkmen, and Yuyang Wang. GluonTS: Probabilistic and Neural Time Series Modeling in Python. *Journal of Machine Learning Research*, 21(116): 1–6, 2020. URL <http://jmlr.org/papers/v21/19-820.html>.
- [4] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.
- [5] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé, Jared Kaplan, Harrison Edwards, Yura Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, David W. Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William H. Guss, Alex Nichol, Igor Babuschkin, Suchir Balaji, Shantanu Jain, Andrew Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew M. Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *ArXiv*, abs/2107.03374, 2021. URL <https://api.semanticscholar.org/CorpusID:235755472>.
- [6] Mouxiang Chen, Lefei Shen, Zhuo Li, Xiaoyun Joy Wang, Jianling Sun, and Chenghao Liu. Visions: Visual masked autoencoders are free-lunch zero-shot time series forecasters. 2024. URL <https://api.semanticscholar.org/CorpusID:272310529>.
- [7] Abhimanyu Das, Weihao Kong, Andrew B. Leach, Shaan Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with tide: Time-series dense encoder. *ArXiv*, abs/2304.08424, 2023. URL <https://api.semanticscholar.org/CorpusID:258180439>.
- [8] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *ArXiv*, abs/2310.10688, 2023. URL <https://api.semanticscholar.org/CorpusID:264172792>.
- [9] Vijay Ekambaram, Arindam Jati, Nam H. Nguyen, Pankaj Dayama, Chandra Reddy, Wesley M. Gifford, and Jayant Kalagnanam. Tiny time mixers (ttms): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. *ArXiv*, abs/2401.03955, 2024. URL <https://api.semanticscholar.org/CorpusID:266844130>.
- [10] Patrick Emami, Abhijeet Sahu, and Peter Graf. Buildingsbench: A large-scale dataset of 900k buildings and benchmark for short-term load forecasting. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL <https://openreview.net/forum?id=c5rqd6PZn6>.
- [11] Valentin Flunkert, David Salinas, and Jan Gasthaus. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *ArXiv*, abs/1704.04110, 2017. URL <https://api.semanticscholar.org/CorpusID:12199225>.
- [12] F. Garza, M. M. Canseco, C. Challu, and K. G. Olivares. Statsforecast: Lightning fast forecasting with statistical and econometric models. Presented at PyCon Salt Lake City, Utah, US, 2022. URL: <https://github.com/Nixtla/statsforecast>.

- [13] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.
- [14] Rakshitha Wathsadini Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL <https://openreview.net/forum?id=wEc1mgAjU->.
- [15] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *ArXiv*, abs/2009.03300, 2020. URL <https://api.semanticscholar.org/CorpusID:221516475>.
- [16] Addison Howard, Haruka Yui, Mark McDonald, and Will Cukierski. Recruit restaurant visitor forecasting. <https://kaggle.com/competitions/recruit-restaurant-visitor-forecasting>, 2017. Kaggle.
- [17] Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 2018.
- [18] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2017. URL <https://api.semanticscholar.org/CorpusID:4922476>.
- [19] Bryan Lim, Sercan Ö. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *ArXiv*, abs/1912.09363, 2019. URL <https://api.semanticscholar.org/CorpusID:209414891>.
- [20] Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai, Chao Huang, Zhenguang Liu, Bryan Hooi, and Roger Zimmermann. Largest: A benchmark dataset for large-scale traffic forecasting. *arXiv preprint arXiv:2306.08259*, 2023.
- [21] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *ArXiv*, abs/2310.06625, 2023. URL <https://api.semanticscholar.org/CorpusID:263830644>.
- [22] Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *International Conference on Machine Learning*, 2024. URL <https://api.semanticscholar.org/CorpusID:267412273>.
- [23] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 2018. URL <https://api.semanticscholar.org/CorpusID:158696437>.
- [24] Paolo Mancuso, Veronica Piccialli, and Antonio Maria Sudoso. A machine learning approach for forecasting hierarchical time series. *Expert Syst. Appl.*, 182:115102, 2020. URL <https://api.semanticscholar.org/CorpusID:219177009>.
- [25] Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, William Paul, Michael I. Jordan, and Ion Stoica. Ray: A distributed framework for emerging ai applications. *ArXiv*, abs/1712.05889, 2017. URL <https://api.semanticscholar.org/CorpusID:34552495>.
- [26] Soukayna Mouatadid, Paulo Orenstein, Genevieve Elaine Flaspohler, Miruna Oprescu, Judah Cohen, Franklyn Wang, Sean Edward Knight, Maria Geogdzhayeva, Samuel James Levang, Ernest Fraenkel, and Lester Mackey. SubseasonalclimateUSA: A dataset for subseasonal forecasting and benchmarking. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL <https://openreview.net/forum?id=pWkrU6raMt>.

- [27] Tung Nguyen, Jason Kyle Jewik, Hritik Bansal, Prakhar Sharma, and Aditya Grover. Climate-learn: Benchmarking machine learning for weather and climate modeling. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL <https://openreview.net/forum?id=RZJEkLF1Px>.
- [28] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *ArXiv*, abs/2211.14730, 2022. URL <https://api.semanticscholar.org/CorpusID:254044221>.
- [29] Boris N. Oreshkin, Dmitri Carпов, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. *ArXiv*, abs/1905.10437, 2019. URL <https://api.semanticscholar.org/CorpusID:166228758>.
- [30] Santosh Palaskar, Vijay Ekambaram, Arindam Jati, Neelamadhav Gantayat, Avirup Saha, Seema Nagar, Nam Nguyen, Pankaj Dayama, Renuka Sindhgatta, Prateeti Mohapatra, Harshit Kumar, Jayant Kalagnanam, Nandyala Hemachandra, and Narayan Rangaraj. Automixer for improved multivariate time-series forecasting on business and it observability data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:22962–22968, 03 2024. doi: 10.1609/aaai.v38i21.30336.
- [31] 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. *Proc. VLDB Endow.*, 17:2363–2377, 2024. URL <https://api.semanticscholar.org/CorpusID:268793935>.
- [32] Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Bilovs, Hena Ghonia, Nadhir Vincent Hassen, Anderson Schneider, Sahil Garg, Alexandre Drouin, Nicolas Chapados, Yuriy Nevmyvaka, and Irina Rish. Lag-llama: Towards foundation models for probabilistic time series forecasting. 2023. URL <https://api.semanticscholar.org/CorpusID:263909560>.
- [33] Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Bilovs, Hena Ghonia, Nadhir Vincent Hassen, Anderson Schneider, Sahil Garg, Alexandre Drouin, Nicolas Chapados, Yuriy Nevmyvaka, and Irina Rish. Lag-llama: Towards foundation models for probabilistic time series forecasting. 2023. URL <https://api.semanticscholar.org/CorpusID:263909560>.
- [34] Neal Richardson, Ian Cook, Neal Crane, Dewey Dunnington, Romain François, Jonathan Keane, Diana Moldovan-Grunfeld, Jeroen Ooms, Julia Wujciak-Jens, and Apache Arrow. arrow: Integration to 'apache' 'arrow', 2023. URL <https://github.com/apache/arrow/>. R package version 14.0.2, <https://arrow.apache.org/docs/r/>.
- [35] Siqi Shen, Vincent Beek, and Alexandru Iosup. Statistical characterization of business-critical workloads hosted in cloud datacenters. *Proceedings - 2015 IEEE/ACM 15th International Symposium on Cluster, Cloud, and Grid Computing, CCGrid 2015*, pages 465–474, 07 2015. doi: 10.1109/CCGrid.2015.60.
- [36] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askeell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmuller, Andrew M. Dai, Andrew La, Andrew Kyle Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakacs, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Ozyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Stephen Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, C’esar Ferri Ram’irez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts,

Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Daniel H Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Mosegu'i Gonz'alez, Danielle R. Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth P. Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan J. Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Mart'inez-Plumed, Francesca Happ'e, François Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-L'opez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schutze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, John Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Koco'n, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Narain Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Oluwadara Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Jane W Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jorg Frohberg, Jos Rozen, José Hernández-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kautubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Luca Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Col'on, Luke Metz, Lutfi Kerem cSenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqi, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ram'irez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michal Swkedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimeo Xu, Mirac Suzgun, Mitch Walker, Monica Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, T MukundVarma, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipourmolas, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, P Hwang, P. Milkowski, Piyush S. Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphael Milliere, Rhythm Garg, Richard Barnes, Rif A. Sauros, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi S. Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Bradley Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Dovic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T Piantadosi, Stuart M. Shieber, Summer Mishnerghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg,

- Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yu Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv*, abs/2206.04615, 2022. URL <https://api.semanticscholar.org/CorpusID:263625818>.
- [37] Artur Trindade. ElectricityLoadDiagrams20112014. UCI Machine Learning Repository, 2015. DOI: <https://doi.org/10.24432/C58C86>.
- [38] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *BlackboxNLP@EMNLP*, 2018. URL <https://api.semanticscholar.org/CorpusID:5034059>.
- [39] Jingyuan Wang, Jiawei Jiang, Wenjun Jiang, Chengkai Han, and Wayne Xin Zhao. Towards efficient and comprehensive urban spatial-temporal prediction: A unified library and performance benchmark. *arXiv preprint arXiv:2304.14343*, 2023.
- [40] Zhixian Wang, Qingsong Wen, Chaoli Zhang, Liang Sun, Leandro Von Krannichfeldt, and Yi Wang. Benchmarks and custom package for electrical load forecasting. *arXiv preprint arXiv:2307.07191*, 2023.
- [41] Gerald Woo, Chenghao Liu, Akshat Kumar, and Doyen Sahoo. Pushing the limits of pre-training for time series forecasting in the cloudops domain. *arXiv preprint arXiv:2310.05063*, 2023.
- [42] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. *ArXiv*, abs/2402.02592, 2024. URL <https://api.semanticscholar.org/CorpusID:267411817>.
- [43] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Neural Information Processing Systems*, 2021. URL <https://api.semanticscholar.org/CorpusID:235623791>.
- [44] Ailing Zeng, Mu-Hwa Chen, L. Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *AAAI Conference on Artificial Intelligence*, 2022. URL <https://api.semanticscholar.org/CorpusID:249097444>.
- [45] Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=vSVLM2j9eie>.
- [46] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wan Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. *ArXiv*, abs/2012.07436, 2020. URL <https://api.semanticscholar.org/CorpusID:229156802>.
- [47] 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.

A Related work

Our benchmark, GIFT-Eval, builds upon and seeks to address gaps identified in existing time series forecasting benchmarks. This section outlines how our work relates to and expands upon previous efforts.

Woo et al. [42] introduced LOTSA, which holds the title for the largest collection of open time series datasets, encompassing 27 billion observations across nine domains. Despite its vast size, the evaluation datasets still lack sufficient variety in terms of time series data characteristics and forecasting tasks, which our benchmark aims to augment. Ansari et al. [4] developed a dataset specifically structured for pretraining, in-domain evaluation, and zero-shot evaluation splits. However, their work is constrained by a limited range in prediction lengths (from 6 to 56), missing out on long-term forecasts, and restricts the data to univariate forecasting, whereas our benchmark includes extensive multivariate scenarios. The corpus by Rasul et al. [33] presents a diverse array of domains, yet it comprises only univariate datasets totaling 8,000 time series. In contrast, GIFT-Eval dramatically expands this scope with 144,000 time series, enhancing the breadth and depth of the dataset. The benchmark by Qiu et al. [31] is closely aligned with our work in its aim to curate a diverse and comprehensive set of data. However, it lacks pretraining data, does not evaluate foundational models, and limits its scope to point forecasts. Our benchmark not only includes pretraining data (with zero-shot evaluation support) but also provides evaluations for foundational models and offers both point and probabilistic forecasts, filling a critical gap in the existing landscape.

B Experimental setup details

Time series forecasting training and inference may take different forms for different families of models. Statistical models make predictions by directly analyzing patterns in the historical data without a separate training phase. We incorporate five statistical models in our benchmark: Naive, Seasonal Naive [17], Auto_Arima, Auto_ETS, and Auto_Theta [12] methods. Deep learning models require training a specific model instance for each dataset. Representing deep learning, we select 8 models: DeepAR [11], TFT [19], TiDE [7], N-BEATS [29], PatchTST [28], DLinear [44], Crossformer [45] and iTransformer [21]. To obtain both point and probabilistic forecasts, we either adapt models using gluonts [3] with a small probabilistic head or implement our own modifications. We conduct an extensive hyperparameter search for each deep learning model, see Appendix B for details. We evaluate four foundation models on our benchmark: TimesFM [8], Chronos [4] available in tiny, small, and base sizes, Moirai [42] available in small, base, and large sizes and, Lag-Llama [32], Timer [22], TTM [9], VisionTS [6]. These models all provide publicly accessible model parameters for direct use. However, it is important to note that pre-training datasets of TimesFM, Chronos, and Moirai exhibit partial data leakage issues for GIFT-Eval.

Statistical models We utilize the statsforecast [12] library to implement all three statistical baselines: Naive, Seasonal Naive, Auto_ETS, Auto_Theta, and Auto_Arima. Inference is performed on a CPU server equipped with 96 cores. For each dataset, a time limit of one day is set for the statistical model to complete its run, with any model that times out being halted and its results replaced with those from the Seasonal Naive model as a fallback. Given that some datasets in our benchmark are particularly long, we impose a maximum size constraint on each statistical baseline (set to 1000 with our time constraints), truncating the time series to this max_size.

Table 6: Hyperparameter search range for deep learning baselines.

| | | | | | | | | |
|--------------|--------------------------|------------------------|------------------------|---------------------|--------------------|-----------------|--------------------|-------------|
| | TiDE | | | | | | | |
| Parameters | num_layers_encoder | num_layers_decoder | hidden_dim | temporal_hidden_dim | decoder_output_dim | dropout_rate | lr | |
| Search Range | [1,2] | [1,2] | [256,512,1024] | [64,128] | [8,16,32] | [0.0, 0.5] | [1e-5:1e-1] | |
| | N-BEATS | | | | | | | |
| Parameters | loss_function | hidden_layer_units | share_weights_in_stack | nb_blocks_per_stack | lr | PatchTST | num_encoder_layers | lr |
| Search Range | ["mse", "mape", "smape"] | [256, 512, 1024, 2048] | [True, False] | [3, 4] | [1e-5:1e-1] | [128, 256, 512] | [2, 3, 4] | [1e-5:1e-1] |
| | iTransformer | | | | | | | |
| Parameters | d_model | num_encoder_layers | lr | hidden_size | num_layers | lr | DLinear | |
| Search Range | [128, 256, 512] | [2, 3, 4] | [1e-5:1e-1] | [20,25,...80] | [1,2,3,4] | [1e-5:1e-1] | [1e-5:1e-1] | |
| | Crossformer | | | | | | | |
| Parameters | d_model | n_heads | lr | num_heads | hidden_dim | lr | | |
| Search Range | [64,128,256] | [2,4,8] | [1e-5:5e-3] | [2,4,8] | [16,32,64] | [1e-5:1e-1] | | |

Deep learning models For all deeplearning models we either used models readily available in gluonts library [3] or we write our own wrappers. Where feasible we also add a probabilistic forecasting head to the models. Where direct probabilistic outputs are not feasible, we generate probabilistic evaluations by converting point forecasts into sample forecasts using a single sample. To identify the optimal hyperparameters, we conducted a comprehensive search across all 97 runs included in GIFT-Eval. We employed the ray library [25] to parallelize the search on a single GPU and used the optuna [1] library to extend this parallelization across multiple GPU servers. We search for 15 trials for each deep learning model per each of the 97 runs. Table 6 lists the range

of parameters we search for each model. On top of the listed parameters for each model, we also search for weight decay on all runs in the range: $[1e - 8 : 1e - 2]$, and for context length in range $[1, 2, 4, 8] \times prediction_length$. For the Crossformer model on the long term setting of *Jena Weather* dataset with both ten-minutely and hourly frequencies, we had to limit the search for `d_model` and `n_heads`, fixing them at 32 and 1, respectively. This adjustment was necessary because the model’s attention mechanism operates across multiple variates, leading to an OOM (Out of Memory) error due to the high number of variates present in this dataset.

Foundation models For all foundation models we use their public versions available online and conduct zero-shot evaluation on our benchmark’s test-split. Since Moirai [42] provides multi-patch size projections and varying context lengths. We adopt the similar approach by defining a frequency-to-patch size mapping as follows:

- Yearly, Quarterly: 8
- Monthly: 8
- Weekly, Daily: 16
- Hourly: 32
- Minute-level: 32
- Second-level: 64

We set context length to 4000. We used the public available Moirai models from the corresponding HuggingFace repos, i.e., `MoiraiSmall` - <https://huggingface.co/Salesforce/moirai-1.1-R-small>, `MoiraiBase` - <https://huggingface.co/Salesforce/moirai-1.1-R-base>, `MoiraiLarge` - <https://huggingface.co/Salesforce/moirai-1.1-R-large>.

For Chronos, we mainly follow their official implementation⁴ for evaluation: with the number of samples as 20. The models are loaded from the corresponding HuggingFace repos, e.g., `ChronosTiny` - <https://huggingface.co/amazon/chronos-t5-tiny>, `ChronosSmall` - <https://huggingface.co/amazon/chronos-t5-small>, `ChronosBase` - <https://huggingface.co/amazon/chronos-t5-base>.

For TimesFM, we follow their official implementation⁵ for evaluation. We set the context length for evaluation as 512 as mentioned in their paper since the maximum context length in training is 512. Following their default setting in their example, we keep the input patch length as 32, the output patch length as 128, the number of layers as 20, and the model dimension as 1280. TimesFM comes with only one model size, i.e., `timesfm-1.0-200m`, and we load the model from <https://huggingface.co/google/timesfm-1.0-200m>.

For VisionTS, we follow their official implementation⁶ for evaluation. We set the context length as 2000, the norm constant as 0.4, the alignment constant as 0.4 according to their default settings. We use their implementation for seasonality detection to generate a candidate list and search an optimal seasonality parameter with the validation data.

Additional parameters and computational resources. All experiments are conducted on eight NVIDIA A100 GPUs. For models that has gone through training the loss function and optimizer are set following their original implementation. Additionally we set the batch size to 128 and, number of batches per epoch to 100, and finally number of epochs to 50.

C GIFT-Eval test datasets

In this section we provide comprehensive list of datasets used in test portion of GIFT-Eval along with original sources and statistics, for details regarding the pretraining portion see Appendix D. Table 7 lists all datasets, along with their source, frequency, prediction length and number of variates setup

⁴<https://github.com/amazon-science/chronos-forecasting/blob/main/scripts/evaluation/evaluate.py>

⁵https://github.com/google-research/timesfm/blob/master/experiments/long_horizon_benchmarks/run_eval.py

⁶https://github.com/Keytoyze/VisionTS/blob/main/eval_gluonts/run.py

Table 7: Individual statistics of GIFT-Eval benchmark across all datasets.

| Dataset | Source | Domain | Frequency | # Series | Series Length | | | | Short-term | | | Med-term | | | Long-term | |
|--------------------------|-----------------------------|-------------|-----------|----------|---------------|---------|---------|------------|-----------------|----------------|---------|----------------|---------|----------------|-----------|--|
| | | | | | Avg | Min | Max | # Obs | Target Variates | Pred Length(S) | Windows | Pred Length(M) | Windows | Pred Length(L) | Windows | |
| Jena Weather | Autoformer [43] | Nature | 10T | 1 | 52,704 | 52,704 | 52,704 | 52,704 | 21 | 48 | 20 | 480 | 11 | 720 | 8 | |
| Jena Weather | Autoformer [43] | Nature | H | 1 | 8,784 | 8,784 | 8,784 | 8,784 | 21 | 48 | 19 | 480 | 2 | 720 | 2 | |
| Jena Weather | Autoformer [43] | Nature | D | 1 | 366 | 366 | 366 | 366 | 21 | 30 | 2 | 30 | 2 | 30 | 2 | |
| BizITObs - Application | AutoMixer [30] | WebCloudOps | 10S | 1 | 8,834 | 8,834 | 8,834 | 8,834 | 2 | 60 | 15 | 600 | 2 | 900 | 1 | |
| BizITObs - Service | AutoMixer [30] | WebCloudOps | 10S | 1 | 8,835 | 8,835 | 8,835 | 183,535 | 2 | 60 | 15 | 600 | 2 | 900 | 1 | |
| BizITObs - L2C | AutoMixer [30] | WebCloudOps | 5T | 1 | 31,968 | 31,968 | 31,968 | 31,968 | 7 | 48 | 7 | 480 | 2 | 720 | 5 | |
| BizITObs - L2C | AutoMixer [30] | WebCloudOps | H | 1 | 2,664 | 2,664 | 2,664 | 2,664 | 7 | 48 | 6 | 480 | 1 | 720 | 1 | |
| Bitbrains - Fast Storage | Grid Workloads Archive [35] | WebCloudOps | ST | 1,250 | 8,640 | 8,640 | 8,640 | 10,000,000 | 2 | 48 | 18 | 480 | 2 | 720 | 2 | |
| Bitbrains - Fast Storage | Grid Workloads Archive [35] | WebCloudOps | H | 1,250 | 721 | 721 | 721 | 901,250 | 2 | 48 | 2 | 480 | 2 | 720 | 2 | |
| Bitbrains - rnd | Grid Workloads Archive [35] | WebCloudOps | 5T | 500 | 8,640 | 8,640 | 8,640 | 4,320,000 | 2 | 48 | 18 | 480 | 2 | 720 | 2 | |
| Bitbrains - rnd | Grid Workloads Archive [35] | WebCloudOps | H | 500 | 720 | 720 | 720 | 360,000 | 2 | 48 | 2 | 480 | 2 | 720 | 2 | |
| Restaurant | Recruit Rest. Comp. [16] | Sales | D | 807 | 358 | 67 | 478 | 289,303 | 1 | 30 | 1 | 30 | 1 | 30 | 1 | |
| ETT1 | Informers [46] | Energy | 15T | 1 | 69,680 | 69,680 | 69,680 | 69,680 | 7 | 48 | 20 | 480 | 15 | 720 | 10 | |
| ETT1 | Informers [46] | Energy | H | 1 | 17,420 | 17,420 | 17,420 | 17,420 | 7 | 48 | 20 | 480 | 4 | 720 | 3 | |
| ETT1 | Informers [46] | Energy | D | 1 | 725 | 725 | 725 | 725 | 7 | 30 | 3 | 30 | 3 | 30 | 3 | |
| ETT1 | Informers [46] | Energy | W-THU | 1 | 103 | 103 | 103 | 103 | 7 | 8 | 2 | 80 | 15 | 720 | 10 | |
| ETT2 | Informers [46] | Energy | 15T | 1 | 69,680 | 69,680 | 69,680 | 69,680 | 7 | 48 | 20 | 480 | 4 | 720 | 3 | |
| ETT2 | Informers [46] | Energy | H | 1 | 17,420 | 17,420 | 17,420 | 17,420 | 7 | 48 | 20 | 480 | 4 | 720 | 3 | |
| ETT2 | Informers [46] | Energy | D | 1 | 725 | 725 | 725 | 725 | 7 | 30 | 3 | 30 | 3 | 30 | 3 | |
| ETT2 | Informers [46] | Energy | W-THU | 1 | 103 | 103 | 103 | 103 | 7 | 8 | 2 | 80 | 15 | 720 | 10 | |
| Loop Seattle | LibCity [39] | Transport | 5T | 323 | 105,120 | 105,120 | 105,120 | 33,953,760 | 1 | 48 | 20 | 480 | 20 | 720 | 15 | |
| Loop Seattle | LibCity [39] | Transport | H | 323 | 8,760 | 8,760 | 8,760 | 2,829,480 | 1 | 48 | 19 | 480 | 2 | 720 | 2 | |
| Loop Seattle | LibCity [39] | Transport | D | 323 | 365 | 365 | 365 | 117,895 | 1 | 30 | 2 | 30 | 2 | 30 | 2 | |
| SZ-Taxi | LibCity [39] | Transport | 15T | 156 | 2,976 | 2,976 | 2,976 | 464,256 | 1 | 48 | 7 | 480 | 1 | 720 | 1 | |
| SZ-Taxi | LibCity [39] | Transport | H | 156 | 744 | 744 | 744 | 116,064 | 1 | 48 | 2 | 480 | 1 | 720 | 1 | |
| M_DENSE | LibCity [39] | Transport | H | 30 | 17,520 | 17,520 | 17,520 | 525,600 | 1 | 48 | 20 | 480 | 4 | 720 | 3 | |
| M_DENSE | LibCity [39] | Transport | D | 30 | 730 | 730 | 730 | 21,900 | 1 | 30 | 3 | 30 | 3 | 30 | 3 | |
| Solar | LSTNet [18] | Energy | 10T | 137 | 52,560 | 52,560 | 52,560 | 7,200,720 | 1 | 48 | 20 | 480 | 11 | 720 | 8 | |
| Solar | LSTNet [18] | Energy | H | 137 | 8,760 | 8,760 | 8,760 | 1,200,120 | 1 | 48 | 19 | 480 | 2 | 720 | 2 | |
| Solar | LSTNet [18] | Energy | D | 137 | 365 | 365 | 365 | 50,005 | 1 | 30 | 2 | 30 | 2 | 30 | 2 | |
| Solar | LSTNet [18] | Energy | W-FRI | 137 | 52 | 52 | 52 | 7,124 | 1 | 8 | 1 | 80 | 15 | 720 | 10 | |
| Hierarchical Sales | Mancuso et al. [24] | Sales | D | 118 | 1,825 | 1,825 | 1,825 | 213,350 | 1 | 30 | 7 | 30 | 7 | 30 | 7 | |
| Hierarchical Sales | Mancuso et al. [24] | Sales | W-WED | 118 | 260 | 260 | 260 | 30,680 | 1 | 8 | 4 | 80 | 15 | 720 | 10 | |
| M4 Yearly | Monash [14] | EconFin | A-DEC | 22,974 | 37 | 19 | 284 | 845,109 | 1 | 6 | 1 | 60 | 1 | 60 | 1 | |
| M4 Quarterly | Monash [14] | EconFin | Q-DEC | 24,000 | 100 | 24 | 874 | 2,406,108 | 1 | 8 | 1 | 80 | 1 | 80 | 1 | |
| M4 Monthly | Monash [14] | EconFin | M | 48,000 | 234 | 60 | 2,812 | 11,246,411 | 1 | 18 | 1 | 180 | 1 | 180 | 1 | |
| M4 Weekly | Monash [14] | EconFin | W-SUN | 359 | 1,035 | 93 | 2,610 | 371,579 | 1 | 13 | 1 | 130 | 1 | 130 | 1 | |
| M4 Daily | Monash [14] | EconFin | D | 4,227 | 2,371 | 107 | 9,933 | 10,023,836 | 1 | 14 | 1 | 140 | 1 | 140 | 1 | |
| M4 Hourly | Monash [14] | EconFin | H | 414 | 902 | 748 | 1,008 | 373,372 | 1 | 48 | 2 | 480 | 2 | 480 | 2 | |
| Hospital | Monash [14] | Healthcare | M | 767 | 84 | 84 | 84 | 64,428 | 1 | 12 | 1 | 120 | 1 | 120 | 1 | |
| COVID Deaths | Monash [14] | Healthcare | D | 266 | 212 | 212 | 212 | 56,392 | 1 | 30 | 1 | 30 | 1 | 30 | 1 | |
| US Births | Monash [14] | Healthcare | D | 1 | 7,305 | 7,305 | 7,305 | 7,305 | 1 | 30 | 20 | 30 | 20 | 30 | 20 | |
| US Births | Monash [14] | Healthcare | W-TUE | 1 | 1,043 | 1,043 | 1,043 | 1,043 | 1 | 8 | 14 | 80 | 14 | 80 | 14 | |
| US Births | Monash [14] | Healthcare | M | 1 | 240 | 240 | 240 | 240 | 1 | 12 | 2 | 120 | 2 | 120 | 2 | |
| Saugen | Monash [14] | Nature | D | 1 | 23,741 | 23,741 | 23,741 | 23,741 | 1 | 30 | 20 | 30 | 20 | 30 | 20 | |
| Saugen | Monash [14] | Nature | W-THU | 1 | 3,391 | 3,391 | 3,391 | 3,391 | 1 | 8 | 20 | 80 | 20 | 80 | 20 | |
| Saugen | Monash [14] | Nature | M | 1 | 780 | 780 | 780 | 780 | 1 | 12 | 7 | 120 | 7 | 120 | 7 | |
| Temperature Rain | Monash [14] | Nature | D | 32,072 | 725 | 725 | 725 | 780 | 1 | 30 | 30 | 30 | 30 | 30 | 30 | |
| KDD Cup 2018 | Monash [14] | Nature | H | 270 | 10,898 | 9,504 | 10,920 | 2,942,364 | 1 | 48 | 20 | 480 | 2 | 720 | 2 | |
| KDD Cup 2018 | Monash [14] | Nature | D | 270 | 455 | 396 | 455 | 122,791 | 1 | 30 | 2 | 30 | 2 | 30 | 2 | |
| Car Parts | Monash [14] | Sales | M | 2,674 | 51 | 51 | 51 | 136,374 | 1 | 12 | 1 | 120 | 1 | 120 | 1 | |
| Electricity | UCI ML Archive [37] | Energy | 15T | 370 | 140,256 | 140,256 | 140,256 | 51,894,720 | 1 | 48 | 20 | 480 | 20 | 720 | 20 | |
| Electricity | UCI ML Archive [37] | Energy | H | 370 | 35,064 | 35,064 | 35,064 | 12,973,680 | 1 | 48 | 20 | 480 | 8 | 720 | 5 | |
| Electricity | UCI ML Archive [37] | Energy | D | 370 | 1,461 | 1,461 | 1,461 | 540,570 | 1 | 30 | 5 | 30 | 5 | 30 | 5 | |
| Electricity | UCI ML Archive [37] | Energy | W-FRI | 370 | 208 | 208 | 208 | 76,960 | 1 | 8 | 3 | 80 | 3 | 80 | 3 | |

and presents various statistics from number of series, to series length, and also number of observations. We use last 10% of each timeseries in the test portion of our data for testing and keep the rest for training.

In order to help reader digest these tables easily we have also curated tables Tables 8 to 11 which aggregates number of time series and observations for prediction length, domain, frequency and number of variates respectively.

In curating the test portion of our benchmark we made use of 10 open domain sources. We use Jena Weather⁷ dataset following **Autoformer** [43]. It is recorded every 10 minutes a year, which contains 21 variates like temperature, humidity and so on. We process BizITObs Application, Service, and L2C⁸ following the pipeline in **AutoMixer** [30]. These are a series of business and IT observability data and they fuse both business KPIs and IT event channels together as multivariate time series data. Within the same domain we also process Bitbrains datasets from **Grid Workloads Archive** [35]. The Restaurant data is borrowed from **Recruit Restaurant Forecasting Competition** [16], the task in this dataset is to use reservation and visitation data to predict the total number of visitors to a restaurant for future dates. From **Informers** [47] we utilize ETT1 and ETT2 datasets which denote electricity transformer temperature and is an indicator used in the electric power long-term deployment. Dataset for Transport domain are extracted from **LibCity** [39], which provides a collection of urban time series datasets. We utilize the solar dataset from **LSTNet** [18] where the task is to predict solar plant energy outputs. The second and last dataset for Sales data is by Mancuso et al. [24]. **Monash** [14] is a large collection of diverse time series datasets across many domains, we choose a subset of these datasets making sure there is no leak from pretrain to test split. Finally from **UCI ML Archive** [37] we use the electricity dataset which contains electricity consumption of 370 individual clients.

Table 8: Statistics aggregated by prediction length.

| Pred. Length | 6 | 8 | 12 | 13 | 14 | 18 | 30 | 48 | 60 | 480 | 600 | 720 | 900 |
|--------------|---------|-----------|---------|---------|------------|------------|-----------|-------------|---------|-------------|---------|-------------|---------|
| # Series | 22,974 | 24,629 | 3,443 | 359 | 4,227 | 48,000 | 34,398 | 6,194 | 22 | 3,874 | 22 | 3,874 | 22 |
| # Obs | 845,109 | 2,525,512 | 201,042 | 371,579 | 10,023,836 | 11,246,411 | 1,447,848 | 131,125,706 | 194,369 | 129,375,020 | 194,369 | 129,375,020 | 194,369 |

⁷<https://www.bgc-jena.mpg.de/wetter/>

⁸<https://github.com/BizITObs/BizITObservabilityData/tree/main>

Table 9: Statistics aggregated by domain.

| Domain | Econ/Fin | Energy | Healthcare | Nature | Sales | Transport | Web/CloudOps | Grand Total |
|----------|------------|------------|------------|-----------|---------|------------|--------------|-------------|
| # Series | 99,974 | 2,036 | 1,036 | 32,618 | 3,717 | 1,341 | 3,524 | 144,246 |
| # Obs | 25,266,415 | 74,119,755 | 129,408 | 3,154,921 | 671,707 | 38,028,955 | 16,610,251 | 157,981,412 |

Table 10: Statistics aggregated by frequency.

| Frequency | 10S | 10T | 15T | 5T | A | D | H | M | Q | W | Grand Total |
|-----------|---------|-----------|------------|------------|---------|------------|------------|------------|-----------|---------|-------------|
| # Series | 22 | 138 | 528 | 2,074 | 22,974 | 38,625 | 3,454 | 51,443 | 24,000 | 988 | 144,246 |
| # Obs | 194,369 | 7,253,424 | 52,498,336 | 49,105,728 | 845,109 | 11,471,684 | 22,268,218 | 11,447,453 | 2,406,108 | 490,983 | 157,981,412 |

Table 11: Statistics aggregated by number of variates.

| # Variates | 1 | 2 | 7 | 21 | Grand Total |
|------------|-------------|------------|---------|--------|-------------|
| # Series | 140,711 | 3,522 | 10 | 3 | 144,246 |
| # Obs | 141,133,451 | 16,575,619 | 210,488 | 61,854 | 157,981,412 |

D GIFT-Eval pre-training datasets

The pre-training split of GIFT-Eval is constructed based on LOTSA [42], and we excluded certain datasets from it to form part of the evaluation set, making it more diverse and balanced. The complete list of pre-training datasets and their respective sources, key properties are provided in Table 12.

BuildingsBench [10] compiles datasets on residential and commercial building energy consumption. **ClimateLearn** [27] offers time series of various climate-related variables, including temperature, humidity, and multiple pressure levels. **CloudOps TSF** [41] introduces large-scale CloudOps time series datasets that capture key variables such as CPU and memory utilization. **GluonTS** [2] provides a variety of datasets commonly used in time series forecasting. **LargeST** [20] sourced from the California Department of Transportation Performance Measurement System (PeMS) is one of the largest collections to date, which is widely used for traffic forecasting. **LibCity** [39] provides a collection urban spatio-temporal datasets. **SubseasonalClimateUSA** [26] provides climate time series data at daily level. **ProEnFo** [40] introduces a range of datasets for load forecasting which include various covariates such as temperature, humidity, and wind speed. **Monash** [14] is a large collection of diverse time series datasets, the most popular source for building time series foundation models. **LOTSAs_Others** [42] are a complementary datasets collected by LOTSA to enhance the diversity.

E Results with all models

In this section, we present results for all models with more metrics, including those omitted from the main paper due to space constraints. The results are displayed in the same aggregated form through Tables 13 to 17. Furthermore, we provide non-aggregated results across all dataset, term and frequency combinations in Tables 18 to 20.

Table 12: Pretraining datasets and their key properties.

| Dataset | Source | Domain | Frequency | # Time Series | # Targets | # Covariates | # Obs. |
|-------------------------------|----------------------------|------------|-----------|---------------|-----------|--------------|----------------|
| BDG-2 Panther | BuildingsBench [10] | Energy | H | 105 | 1 | 0 | 919,800 |
| BDG-2 Fox | BuildingsBench [10] | Energy | H | 135 | 1 | 0 | 2,324,568 |
| BDG-2 Rat | BuildingsBench [10] | Energy | H | 280 | 1 | 0 | 4,728,288 |
| BDG-2 Bear | BuildingsBench [10] | Energy | H | 91 | 1 | 0 | 1,482,312 |
| Low Carbon London | BuildingsBench [10] | Energy | H | 713 | 1 | 0 | 9,543,348 |
| SMART | BuildingsBench [10] | Energy | H | 5 | 1 | 0 | 95,709 |
| IDEAL | BuildingsBench [10] | Energy | H | 219 | 1 | 0 | 1,265,672 |
| Sceaux | BuildingsBench [10] | Energy | H | 1 | 1 | 0 | 34,223 |
| Borealis | BuildingsBench [10] | Energy | H | 15 | 1 | 0 | 83,269 |
| Buildings900K | BuildingsBench [10] | Energy | H | 1,792,328 | 1 | 0 | 15,702,590,000 |
| CMIP6 | ClimateLearn [27] | Climate | 6H | 1,351,680 | 53 | 0 | 1,973,453,000 |
| ERA5 | ClimateLearn [27] | Climate | H | 245,760 | 45 | 0 | 2,146,959,000 |
| Azure VM Traces 2017 | CloudOpsTSF [41] | CloudOps | 5T | 159,472 | 1 | 2 | 885,522,908 |
| Borg Cluster Data 2011 | CloudOpsTSF [41] | CloudOps | 5T | 143,386 | 2 | 5 | 537,552,854 |
| Alibaba Cluster Trace 2018 | CloudOpsTSF [41] | CloudOps | 5T | 58,409 | 2 | 6 | 95,192,530 |
| Taxi | GluonTS [2] | Transport | 30T | 67,984 | 1 | 0 | 54,999,060 |
| Uber TLC Daily | GluonTS [2] | Transport | D | 262 | 1 | 0 | 47,087 |
| Uber TLC Hourly | GluonTS [2] | Transport | H | 262 | 1 | 0 | 1,129,444 |
| Wiki-Rolling | GluonTS [2] | Web | D | 47,675 | 1 | 0 | 40,619,100 |
| M5 | GluonTS [2] | Sales | D | 30,490 | 1 | 0 | 58,327,370 |
| LargeST | LargeST [20] | Transport | 5T | 42,333 | 1 | 0 | 4,452,510,528 |
| PEMS03 | LibCity [39] | Transport | 5T | 358 | 1 | 0 | 9,382,464 |
| PEMS04 | LibCity [39] | Transport | 5T | 307 | 3 | 0 | 5,216,544 |
| PEMS07 | LibCity [39] | Transport | 5T | 883 | 1 | 0 | 24,921,792 |
| PEMS08 | LibCity [39] | Transport | 5T | 170 | 3 | 0 | 3,035,520 |
| PEMS Bay | LibCity [39] | Transport | 5T | 325 | 1 | 0 | 16,937,700 |
| Los-Loop | LibCity [39] | Transport | 5T | 207 | 1 | 0 | 7,094,304 |
| Beijing Subway | LibCity [39] | Transport | 30T | 276 | 2 | 11 | 248,400 |
| SHMetro | LibCity [39] | Transport | 15T | 288 | 2 | 0 | 1,934,208 |
| HZMetro | LibCity [39] | Transport | 15T | 80 | 2 | 0 | 146,000 |
| Q-Traffic | LibCity [39] | Transport | 15T | 45,148 | 1 | 0 | 264,386,688 |
| Subseasonal | SubseasonalClimateUSA [26] | Climate | D | 862 | 4 | 0 | 14,097,148 |
| Subseasonal Precipitation | SubseasonalClimateUSA [26] | Climate | D | 862 | 1 | 0 | 9,760,426 |
| Covid19 Energy | ProEnFo [40] | Energy | H | 1 | 6 | 31,912 | |
| GEF12 | ProEnFo [40] | Energy | H | 20 | 1 | 788,280 | |
| GEF14 | ProEnFo [40] | Energy | H | 1 | 1 | 17,520 | |
| GEF17 | ProEnFo [40] | Energy | H | 8 | 1 | 140,352 | |
| PDB | ProEnFo [40] | Energy | H | 1 | 1 | 17,520 | |
| Spanish | ProEnFo [40] | Energy | H | 1 | 1 | 35,064 | |
| BDG-2 Hog | ProEnFo [40] | Energy | H | 24 | 1 | 5 | 421,056 |
| BDG-2 Bull | ProEnFo [40] | Energy | H | 41 | 1 | 3 | 719,304 |
| BDG-2 Cockatoo | ProEnFo [40] | Energy | H | 1 | 1 | 5 | 17,544 |
| ELF | ProEnFo [40] | Energy | H | 1 | 1 | 0 | 21,792 |
| London Smart Meters | Monash [14] | Energy | 30T | 5,520 | 1 | 0 | 166,238,880 |
| Wind Farms | Monash [14] | Energy | T | 337 | 1 | 0 | 172,165,370 |
| Wind Power | Monash [14] | Energy | 4S | 1 | 1 | 0 | 7,397,147 |
| Solar Power | Monash [14] | Energy | 4S | 1 | 1 | 0 | 7,397,222 |
| Oikolab Weather | Monash [14] | Climate | H | 8 | 1 | 0 | 800,456 |
| ElecDemand | Monash [14] | Energy | 30T | 1 | 1 | 0 | 17,520 |
| Covid Mobility | Monash [14] | Transport | D | 362 | 1 | 0 | 148,602 |
| Kaggle Web Traffic Weekly | Monash [14] | Web | W | 145,063 | 1 | 0 | 16,537,182 |
| Extended Web Traffic | Monash [14] | Web | D | 145,063 | 1 | 0 | 370,926,091 |
| M1 Yearly | Monash [14] | Econ/Fin | Y | 106 | 1 | 0 | 3,136 |
| M1 Quarterly | Monash [14] | Econ/Fin | Q | 198 | 1 | 0 | 9,854 |
| M1 Monthly | Monash [14] | Econ/Fin | M | 617 | 1 | 0 | 44,892 |
| M3 Yearly | Monash [14] | Econ/Fin | Y | 645 | 1 | 0 | 18,319 |
| M3 Quarterly | Monash [14] | Econ/Fin | Q | 756 | 1 | 0 | 37,004 |
| M3 Monthly | Monash [14] | Econ/Fin | M | 1,428 | 1 | 0 | 141,858 |
| M3 Other | Monash [14] | Econ/Fin | Q | 174 | 1 | 0 | 11,933 |
| NN5 Daily | Monash [14] | Econ/Fin | D | 111 | 1 | 0 | 81,585 |
| NN5 Weekly | Monash [14] | Econ/Fin | W | 111 | 1 | 0 | 11,655 |
| Tourism Yearly | Monash [14] | Econ/Fin | Y | 419 | 1 | 0 | 11,198 |
| Tourism Quarterly | Monash [14] | Econ/Fin | Q | 427 | 1 | 0 | 39,128 |
| Tourism Monthly | Monash [14] | Econ/Fin | M | 366 | 1 | 0 | 100,496 |
| CIF 2016 | Monash [14] | Econ/Fin | M | 72 | 1 | 0 | 6,334 |
| Traffic Weekly | Monash [14] | Transport | W | 862 | 1 | 0 | 82,752 |
| Traffic Hourly | Monash [14] | Transport | H | 862 | 1 | 0 | 14,978,112 |
| Australian Electricity Demand | Monash [14] | Energy | 30T | 5 | 1 | 0 | 1,153,584 |
| Rideshare | Monash [14] | Transport | H | 2,304 | 1 | 0 | 859,392 |
| Sunspot | Monash [14] | Nature | D | 1 | 1 | 0 | 73,894 |
| Vehicle Trips | Monash [14] | Transport | D | 329 | 1 | 0 | 32,512 |
| Weather | Monash [14] | Climate | D | 3,010 | 1 | 0 | 42,941,700 |
| FRED MD | Monash [14] | Econ/Fin | M | 107 | 1 | 0 | 76,612 |
| Pedestrian Counts | Monash [14] | Transport | H | 66 | 1 | 0 | 3,130,762 |
| Bitcoin | Monash [14] | Econ/Fin | D | 18 | 1 | 0 | 74,824 |
| KDD Cup 2022 | LOTSAs_Others [42] | Energy | 10T | 134 | 1 | 9 | 4,727,519 |
| GoDaddy | LOTSAs_Others [42] | Econ/Fin | M | 3,135 | 2 | 0 | 128,535 |
| Favorita Sales | LOTSAs_Others [42] | Sales | D | 111,840 | 1 | 0 | 139,179,538 |
| Favorita Transactions | LOTSAs_Others [42] | Sales | D | 54 | 1 | 0 | 84,408 |
| China Air Quality | LOTSAs_Others [42] | Nature | H | 437 | 6 | 0 | 5,739,234 |
| Beijing Air Quality | LOTSAs_Others [42] | Nature | H | 12 | 11 | 0 | 420,768 |
| Residential Load Power | LOTSAs_Others [42] | Energy | T | 271 | 3 | 0 | 145,994,559 |
| Residential PV Power | LOTSAs_Others [42] | Energy | T | 233 | 3 | 0 | 125,338,950 |
| CDC Fluview ILLNet | LOTSAs_Others [42] | Healthcare | W | 75 | 5 | 0 | 63,903 |
| CDC Fluview WHO NREVSS | LOTSAs_Others [42] | Healthcare | W | 74 | 4 | 0 | 41,760 |
| Project Tycho | LOTSAs_Others [42] | Healthcare | W | 1,258 | 1 | 0 | 1,377,707 |

Table 15: Results on GIFT-Eval with all models aggregated by frequency. The best results across each row are **bolded**, while second best results are underlined.

| Model | Freq. | Metric | Nv. | S.Nv. | A.Ar. | A.Th. | A.ETS | D.AR | TFT | TIDE | N-B | P.TST | Tr. | D.Lin. | C-former | Timer | TTM | L-Llama | T.FM | V.TS | Chr.s | Chr.g | Chr.u | Mois.s | Mois.g | Mois.u | Best | | |
|-------|-------|--------|--------|--------|----------------|---------|---------|---------|---------|---------|---------|----------------|----------------|---------|----------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|--------|--------|--------|
| I08 | sMAPE | 1.57 | 1.00 | 1.00 | <u>4.49e-1</u> | 1.19 | 9.12e-1 | 1.46 | 8.34e-1 | 7.45e-1 | 6.59e-1 | <u>6.17e-1</u> | 1.08 | 1.53 | 1.34 | 1.67 | 1.34 | 1.80 | 1.80 | 7.21e-1 | 1.21 | 1.19 | 1.16 | 1.34 | 1.78 | 1.24 | A.Th. | | |
| | | MSE | 1.98 | 1.00 | 1.00 | 1.59e-1 | 5.51e-1 | 3.70e-1 | 5.37e-1 | 3.25e-1 | 2.71e-1 | 2.24e-1 | 2.35e-1 | 3.69e-1 | 6.15e-1 | 5.17e-1 | 7.28e-1 | 3.98e-1 | 7.87e-1 | 2.10e-1 | 3.23e-1 | 5.23e-1 | 3.09e-1 | 7.05e-1 | 8.44e-1 | 5.72e-1 | A.Th. | | |
| | | ND | 1.24 | 1.00 | 1.00 | 3.51e-1 | 1.01 | 8.02e-1 | 6.79e-1 | 7.36e-1 | 5.62e-1 | 5.83e-1 | 5.44e-1 | 7.35e-1 | 6.86e-1 | 9.00e-1 | 1.44 | 1.17 | 1.35 | 6.50e-1 | 8.03e-1 | 8.72e-1 | 8.27e-1 | 1.34 | 1.20 | 1.12 | A.Th. | | |
| | | MSE | 1.29 | 1.00 | 1.00 | 1.03e-1 | 1.06 | 5.01e-1 | 4.60e-1 | 4.31e-1 | 3.02e-1 | 3.59e-1 | <u>2.65e-1</u> | 4.13e-1 | 3.36e-1 | 6.00e-1 | 1.08 | 1.15 | 1.17 | 4.44e-1 | 5.85e-1 | 6.02e-1 | 5.63e-1 | 1.46 | 1.06 | 1.07 | A.Th. | | |
| | | MSE | 1.24 | 1.00 | 1.00 | 3.51e-1 | 1.01 | 8.02e-1 | 6.79e-1 | 7.36e-1 | 5.61e-1 | 5.83e-1 | 5.44e-1 | 7.35e-1 | 6.86e-1 | 9.00e-1 | 1.44 | 1.17 | 1.35 | 6.48e-1 | 8.03e-1 | 8.72e-1 | 8.27e-1 | 1.34 | 1.20 | 1.12 | A.Th. | | |
| | | CRPS | 1.44 | 1.00 | 1.00 | 3.15e-1 | 1.93 | 7.54e-1 | 6.72e-1 | 7.05e-1 | 5.98e-1 | 5.36e-1 | <u>5.10e-1</u> | 7.82e-1 | 7.29e-1 | 9.58e-1 | 1.53 | 1.12 | 1.30 | 6.01e-1 | 7.93e-1 | 8.59e-1 | 8.18e-1 | 1.24 | 1.06 | 1.02 | A.Th. | | |
| | | Rank | 1.96e1 | 1.13e1 | 1.03e1 | 1.00 | 2.56e1 | 1.23e1 | 8.83 | 1.12e1 | 7.17 | 5.00 | 2.50 | 1.28e1 | 1.22e1 | 1.63e1 | 2.55e1 | 1.92e1 | 2.53e1 | 1.08e1 | 1.12e1 | 1.24e1 | 1.23e1 | 1.33e1 | 1.23e1 | 2.26e1 | 1.95e1 | 1.78e1 | Mo.s.u |

Table 16: Results on GIFT-Eval aggregated by number of variates. The best results across each row are **bolded**, while second best results are underlined.

| Model | Num. Var. | Metric | Nv. | S.Nv. | A.Ar. | A.Th. | A.ETS | D.AR | TFT | TIDE | N-B | P.TST | Tr. | D.Lin. | C-former | Timer | TTM | L-Llama | T.FM | V.TS | Chr.s | Chr.g | Chr.u | Mois.s | Mois.g | Mois.u | Best | |
|-------|-----------|--------|--------|--------|--------|---------|--------|---------|---------|---------|----------------|----------------|----------------|---------|----------|---------|---------|---------|----------------|----------------|---------|---------|---------|---------|---------|---------|---------|-------|
| False | sMAPE | 1.15 | 1.00 | 1.12 | 1.14 | 1.26 | 1.14 | 1.23 | 1.29 | 1.13 | <u>1.01</u> | 1.12 | 1.24 | 1.29 | 1.18 | 1.25 | 1.24 | 1.20 | 1.08 | 1.18 | 1.16 | 1.15 | 1.19 | 1.22 | 1.20 | S.Nv. | | |
| | | MSE | 1.15 | 1.00 | 1.03 | 8.01e-1 | 1.10 | 8.40e-1 | 1.01 | 7.82e-1 | <u>7.11e-1</u> | 7.37e-1 | 6.63e-1 | 1.16 | 8.05e-1 | 9.39e-1 | 9.57e-1 | 1.17 | 6.95e-1 | 8.04e-1 | 7.94e-1 | 7.86e-1 | 8.44e-1 | 8.31e-1 | 8.11e-1 | V.TS | | |
| | | ND | 1.05 | 1.00 | 1.04 | 9.56e-1 | 1.10 | 1.30 | 8.19e-1 | 1.03 | 8.78e-1 | 7.45e-1 | <u>7.92e-1</u> | 9.08e-1 | 9.17e-1 | 9.81e-1 | 9.51e-1 | 1.06 | 9.23e-1 | 8.01e-1 | 8.45e-1 | 8.43e-1 | 8.36e-1 | 8.72e-1 | 8.74e-1 | 8.89e-1 | P.TST | |
| | | MSE | 1.08 | 1.00 | 1.06 | 7.87e-1 | 1.28 | 1.42 | 6.42e-1 | 8.52e-1 | 7.18e-1 | 5.60e-1 | 6.24e-1 | 6.38e-1 | 6.68e-1 | 7.99e-1 | 7.24e-1 | 1.04 | 7.88e-1 | <u>6.20e-1</u> | 7.55e-1 | 7.45e-1 | 7.38e-1 | 7.92e-1 | 7.80e-1 | 8.08e-1 | P.TST | |
| | | MSE | 1.05 | 1.00 | 1.04 | 9.56e-1 | 1.10 | 1.30 | 8.20e-1 | 1.03 | 8.78e-1 | 7.45e-1 | <u>7.92e-1</u> | 9.08e-1 | 9.17e-1 | 9.81e-1 | 9.51e-1 | 1.06 | 9.23e-1 | 8.01e-1 | 8.45e-1 | 8.43e-1 | 8.36e-1 | 8.72e-1 | 8.74e-1 | 8.89e-1 | P.TST | |
| | | CRPS | 1.26 | 1.00 | 1.07 | 9.26e-1 | 1.45 | 1.89 | 8.02e-1 | 1.49e-1 | 1.30e-1 | 1.41e-1 | 4.51e-1 | 1.47e-1 | 1.51e-1 | 1.76e-1 | 1.61e-1 | 1.64e-1 | 1.64e-1 | 1.64e-1 | 1.64e-1 | 1.64e-1 | 1.64e-1 | 1.64e-1 | 1.64e-1 | 1.64e-1 | 1.64e-1 | P.TST |
| | | Rank | 2.40e1 | 2.26e1 | 1.95e1 | 2.08e1 | 2.37e1 | 1.90e1 | 8.95 | 1.55e1 | 1.69e1 | 6.56 | 7.05 | 1.80e1 | 1.73e1 | 2.00e1 | 1.91e1 | 1.70e1 | 1.37e1 | 1.53e1 | 1.24e1 | 1.23e1 | 1.23e1 | 1.35e1 | 5.94e1 | 8.63 | 8.91 | P.TST |

Table 17: Results on GIFT-Eval with all models aggregated by all variates. The best results across each row are **bolded**, while second best results are underlined.

| Model | Metric | Nv. | S.Nv. | A.Ar. | A.Th. | A.ETS | D.AR | TFT | TIDE | N-B | P.TST | Tr. | D.Lin. | C-former | Timer | TTM | L-Llama | T.FM | V.TS | Chr.s | Chr.g | Chr.u | Mois.s | Mois.g | Mois.u | Best | |
|-------|--------|--------|--------|---------|---------|--------|---------|---------|---------|----------------|----------------|-------------|---------|----------|---------|---------|---------|---------|----------------|---------|----------------|----------------|---------|---------|---------|-------|--------|
| sMAPE | 1.25 | 1.00 | 1.05 | 1.16 | 1.14 | 1.37 | 1.06 | 1.16 | 1.06 | 9.83e-1 | 1.04 | 1.14 | 1.14 | 1.12 | 1.23 | 1.17 | 1.32 | 1.04 | 1.02 | 1.02 | 1.01 | 1.00 | 1.06 | 1.04 | 1.02 | P.TST | |
| | MSE | 1.26 | 1.00 | 9.64e-1 | 9.78e-1 | 1.09 | 1.21 | 8.22e-1 | 7.80e-1 | 8.42e-1 | 7.62e-1 | 8.02e-1 | 9.52e-1 | 2.31 | 1.02 | 6.69e-1 | 1.10 | 9.67e-1 | <u>7.55e-1</u> | 8.00e-1 | 7.86e-1 | 7.81e-1 | 8.74e-1 | 8.11e-1 | 7.97e-1 | P.TST | |
| | ND | 1.20 | 1.00 | 9.60e-1 | 1.06 | 1.10 | 1.11 | 8.00e-1 | 9.73e-1 | 8.86e-1 | 7.79e-1 | 8.24e-1 | 9.18e-1 | 1.78 | 1.05 | 9.68e-1 | 1.15 | 8.75e-1 | 8.20e-1 | 8.17e-1 | 8.02e-1 | <u>7.40e-1</u> | 8.73e-1 | 8.22e-1 | 8.19e-1 | P.TST | |
| | MSE | 1.36 | 1.00 | 9.11e-1 | 9.30e-1 | 1.21 | 1.10 | 8.27e-1 | 8.20e-1 | 7.59e-1 | 6.08e-1 | 6.09e-1 | 6.32e-1 | 9.22e-1 | 7.65e-1 | 1.22 | 7.65e-1 | 7.48e-1 | 6.68e-1 | 6.05e-1 | <u>6.32e-1</u> | 7.49e-1 | 6.39e-1 | 6.04e-1 | Mo.s.u | | |
| | MAE | 1.20 | 1.00 | 9.60e-1 | 1.06 | 1.10 | 1.11 | 8.01e-1 | 9.73e-1 | 8.86e-1 | 7.78e-1 | 8.24e-1 | 9.18e-1 | 1.47 | 1.05 | 9.68e-1 | 1.15 | 8.75e-1 | 8.20e-1 | 8.17e-1 | 8.02e-1 | 7.65e-1 | 8.73e-1 | 8.22e-1 | 8.18e-1 | P.TST | |
| | CRPS | 1.38 | 1.00 | 7.70e-1 | 1.15 | 1.33 | 7.21e-1 | 1.63e-1 | 6.52e-1 | 6.89e-1 | 4.96e-1 | 5.24e-1 | 7.14e-1 | 1.38 | 8.20e-1 | 7.53e-1 | 7.44e-1 | 5.75e-1 | 6.38e-1 | 5.60e-1 | 5.51e-1 | 5.47e-1 | 5.47e-1 | 5.36e-1 | 5.15e-1 | P.TST | |
| | Rank | 2.49e1 | 2.28e1 | 1.81e1 | 2.01e1 | 2.30e1 | 1.59e1 | 8.85 | 1.53e1 | 1.78e1 | 7.67 | 8.58 | 1.92e1 | 1.88e1 | 2.13e1 | 1.98e1 | 1.80e1 | 1.13e1 | 1.09e1 | 1.21e1 | 1.21e1 | 1.10e1 | 1.09e1 | 1.10e1 | 7.21 | 7.57 | Mo.s.u |

