BITS: RETHINKING FAIR AND COMPREHENSIVE EVALUATION FOR IRREGULAR TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite recent progress in irregular time series forecasting, the field still lacks a unified benchmark for fair and comprehensive evaluation. Existing evaluations are often conducted on a limited set of datasets with inconsistent experimental protocols and predominantly error-based metrics, rendering it difficult to compare and assess methods fairly and comprehensively across diverse settings. To eliminate these limitations and accelerate progress, we propose BITS, a standardized, reproducible, and extensible benchmark for advancing research on irregular time series forecasting. BITS covers eleven datasets from nine domains with diverse irregularity characteristics, and it characterizes the datasets according to their missing rate, missing pattern complexity, sampling irregularity, and skewness. Further, it offers a unified pipeline for data preprocessing, model integration and evaluation, and reporting. It accommodates regular and irregular time series forecasting methods, including time series foundation models, under consistent settings, incorporating both error-based and non-error-based evaluation metrics. Findings include that method performance varies substantially across irregularity characteristics, with no single modeling strategy consistently dominating. We also find that using error-based or non-error-based metrics can yield different model rankings, highlighting the need for multi-dimensional evaluation. The code can be found at `https://anonymous.4open.science/r/BITS-8F2E/`.

1 INTRODUCTION

Irregular time series are widely encountered in real-world scenarios, such as healthcare monitoring and environmental sensing. In recent years, a variety of methods have been proposed for irregular time series forecasting, such as temporal segmentation and aggregation methods (Luo et al., 2025; Zhang et al., 2024; Qiu et al., 2026; Zhou et al., 2026), differential-equation-based models (Mercatali et al., 2024) and data-format transformation methods (Yalavarthi et al., 2024; Li et al., 2025a; Liu et al., 2026). These methods have advanced the modeling of irregular time series and achieved promising performance in irregular time series forecasting. However, despite this progress, the community still lacks a systematic, standardized, and representative benchmark for irregular time series forecasting, making it difficult to fairly compare different methods and comprehensively evaluate their effectiveness under diverse irregular settings. We identify two key limitations in existing evaluations of irregular time series forecasting.

First, existing studies provide limited dataset coverage and splitting strategies. (i) Many existing studies only rely on a few datasets such as PhysioNet, MIMIC, Human Activity and USHCN (Luo et al., 2025; Zhang et al., 2024; Li et al., 2025a), which cover a relatively narrow range of application domains and irregularity characteristics. In addition, some datasets exhibit extremely high missing rates and complex missing patterns (Silva et al., 2012), failing to fully reflect the missing mechanisms commonly observed in real-world environments. (ii) Moreover, many studies adopt a sample-aware splitting strategy without considering temporal order (Zhou et al., 2026; Yalavarthi et al., 2024), as shown in Figure 1 (a). Although such splitting can be necessary for sparsely observed datasets, where temporal splitting may leave insufficient observations for meaningful forecasting windows, it is also widely used for datasets with a sufficient number of observations. In these cases, ignoring temporal order may allow future observations to appear in the training set while earlier observations are used for validation or testing, potentially introducing information leakage.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

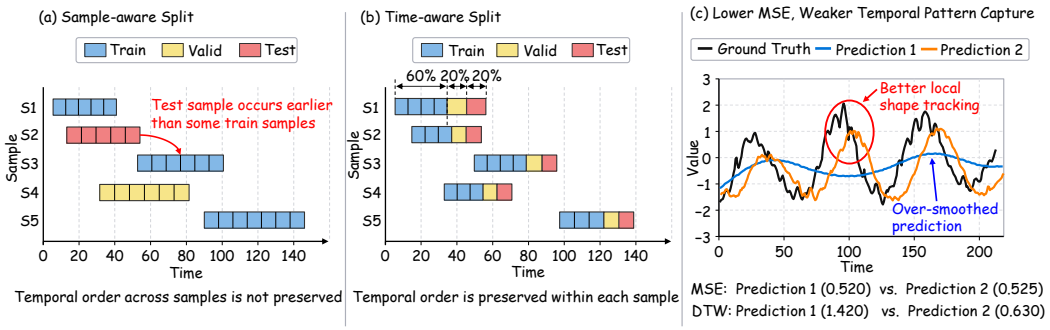


Figure 1: (a) Sample-aware splitting strategy. (b) Time-aware splitting strategy. (c) The limitation of error-based metrics.

Second, existing evaluation protocols are neither standardized nor comprehensive. (i) Different studies adopt different normalization, making reported performance dependent not only on model capability but also on experimental configuration (Zhang et al., 2024; 2023). In particular, different normalization strategies can lead to different data scales for the same dataset, making regression metrics such as MSE difficult to compare fairly across methods. (ii) Existing studies also employ inconsistent input-output window settings, with different combinations of lookback and prediction lengths across datasets and methods (Li et al., 2025a; Liu et al., 2025a). Some settings, such as using only 3 hours of historical observations to forecast the following 45 hours (Zhang et al., 2024), may be difficult to interpret in practical forecasting scenarios. (iii) In addition, most evaluations mainly rely on error-based metrics such as MSE (Shukla & Marlin, 2021; Bilos et al., 2021; Schirmer et al., 2022), which measure point-wise prediction errors and may fail to reflect whether a model captures the temporal dynamics. For example, a model may achieve relatively low MSE while producing nearly constant predictions across the forecast horizon, as shown in Figure 1 (c). Such predictions may obtain favorable regression errors despite failing to capture important temporal patterns. Therefore, low point-wise errors alone do not necessarily indicate effective forecasting of temporal dynamics.

This study addresses the above challenges by providing BITS, a fair and comprehensive **B**enchmark for **I**rregular **T**ime **S**eries forecasting. BITS is designed to support fair, reproducible, and extensible evaluation across datasets, models, and metrics, as shown in Table 1. **To address the first limitation, BITS expands the dataset coverage and splitting strategy.** (i) We collect eleven datasets from nine domains with diverse irregularity characteristics (e.g., missing rate), providing broader coverage of real-world irregular scenarios. (ii) We adopt time-aware splitting strategies for datasets with sufficient observations for temporal splitting to preserve temporal order, as shown in Figure 1 (b), while sample-aware splitting is retained for sparsely observed datasets. Under time-aware dataset splitting strategies, each sample is divided according to temporal order and predefined proportions. The corresponding segments from all instances are then merged to form the final training, validation and test sets, reducing the risk of temporal information leakage.

To address the second limitation, BITS establishes a standardized and comprehensive evaluation pipeline. (i) We standardize key procedures, including normalization, model integration, evaluation, and reporting, and apply consistent normalization strategies (e.g., min-max) to reduce scale discrepancies and enable fair comparisons across methods. (ii) We define customized lookback and prediction lengths for each dataset, which are selected according to the temporal characteristics and practical forecasting requirements of each dataset. (iii) Beyond error-based metrics, BITS also incorporates non-error-based metrics to evaluate trend consistency, shape similarity and structural changes, which enables a more comprehensive understanding of whether a model captures the temporal patterns rather than merely achieving low point-wise errors. In addition, BITS supports irregular forecasting models, regular forecasting models and time series foundation models under the same benchmarking framework, allowing systematic comparison across different modeling paradigms.

In summary, our main contributions are as follows:

- We propose BITS, a comprehensive benchmark for irregular time series forecasting. BITS supports fair and reproducible evaluation of specialized irregular forecasting models, regular forecasting models and time series foundation models across diverse irregular datasets.

Table 1: Comparison of BITS with existing time series forecasting benchmarks. ✓ indicates present, × indicates absent, and ○ indicates incomplete.

	Benchmark	Multi Domain	Irregularity Analysis	Irregular TS	Irregular Forecasting	Regular Model	Irregular Model	TS Foundation Model	Scalable Pipeline	Non-error-based Metric
Regular	BasicTS (Liang et al., 2022)	✓	×	×	×	✓	×	×	○	×
	BasicTS+ (Shao et al., 2025)	✓	×	×	×	✓	×	×	○	×
	Monash (Godahewa et al., 2021)	✓	×	×	×	✓	×	×	○	×
	TSLib (Wang et al., 2026)	✓	×	×	×	✓	×	×	○	×
	TFB (Qiu et al., 2024)	✓	×	×	×	✓	×	×	✓	×
	TSFM-Bench (Li et al., 2025c)	✓	×	×	×	✓	×	✓	✓	×
Irregular	MIMIC-III (Harutyunyan et al., 2019)	×	×	✓	×	✓	×	×	○	×
	MIMIC-IV (Bui et al., 2024)	×	×	✓	×	✓	×	×	○	×
	HyperIMTS (Li et al., 2025a)	○	×	✓	✓	✓	✓	×	○	×
	tPatchGNN (Zhang et al., 2024)	○	○	✓	✓	✓	✓	×	○	×
	Time-IMM (Chang et al., 2025)	✓	✓	✓	✓	✓	✓	×	○	×
	BITS (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓

- We construct a diverse dataset collection with systematic irregularity analysis. BITS includes eleven datasets from nine domains and characterizes them from multiple dimensions, including missing rate, missing pattern complexity, sampling irregularity, and skewness.
- We design a unified and extensible benchmarking pipeline. BITS standardizes preprocessing, normalization, dataset splitting, window construction, model integration, evaluation, and reporting, reducing experimental bias and improving reproducibility. In addition to error-based metrics, BITS also incorporates non-error-based metrics to evaluate whether models capture temporal dynamics.
- We conduct extensive experiments and derive empirical insights into irregular forecasting. The results reveal substantial performance variation across irregularity characteristics and modeling paradigms, while different evaluation metrics can lead to different assessments of model performance.

2 Related Work

Irregular Time Series Forecasting. Existing irregular time series forecasting methods adopt diverse strategies to handle irregular observations. Imputation-based methods reconstruct or aggregate sparse observations into regular representations, such as mTAN (Shukla & Marlin, 2021). Temporal segmentation and aggregation methods, such as tPatchGNN (Zhang et al., 2024), Hi-Patch (Luo et al., 2025) and TiWeaver (Li et al., 2026), organize irregular observations into adaptive temporal units for representation learning. Multi-scale methods such as Warpformer (Zhang et al., 2023) capture temporal patterns at different resolutions. Continuous-time approaches, such as NeuralFlows (Bilos et al., 2021), CRU (Schirmer et al., 2022), and GNeuralFlow (Mercatali et al., 2024), model dynamics directly in continuous time. Other methods transform observations into structured representations such as sets or graphs, including SeFT (Horn et al., 2020), GraFITi (Yalavarthi et al., 2024), HyperIMTS (Li et al., 2025a), or explicitly incorporate sampling context, as in GRU-D (Che et al., 2016).

Benchmarks for Time Series Forecasting. Existing time series benchmarks mainly focus on regularly sampled forecasting. BasicTS (Liang et al., 2022), BasicTS+ (Shao et al., 2025), Monash (Godahewa et al., 2021), TSLib (Wang et al., 2026), and TFB (Qiu et al., 2024) provide standardized datasets and evaluation pipelines for regular forecasting. TSFM-Bench (Li et al., 2025c) further extends evaluation to time series foundation models. However, these benchmarks provide limited support for irregular time series forecasting. Several studies have explored evaluation on irregular time series. MIMIC-III (Harutyunyan et al., 2019) and MIMIC-IV (Bui et al., 2024) primarily focus on clinical applications. HyperIMTS (Li et al., 2025a) and tPatchGNN (Zhang et al., 2024) evaluate irregular forecasting models on a limited number of datasets. More recently, Time-IMM (Chang et al., 2025) extends irregular evaluation toward multimodal representation learning. As summarized in Table 1, existing benchmarks still lack a unified framework that jointly covers diverse irregular datasets, irregularity analysis, regular and irregular forecasting models, TSFMs, scalable evaluation pipelines, and multi-dimensional metrics. BITS is designed to bridge these gaps through systematic and unified evaluation of irregular time series forecasting.

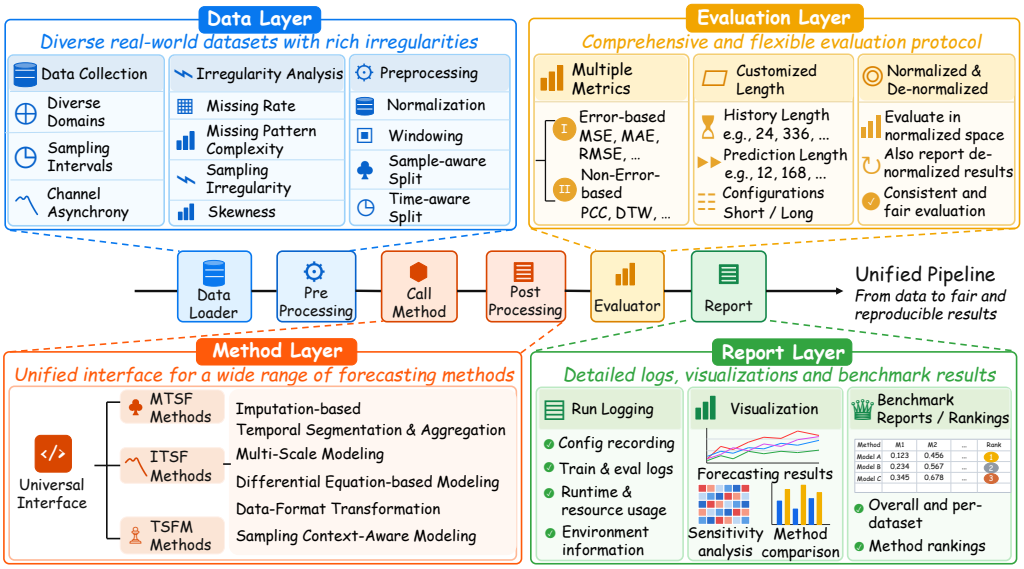


Figure 2: Overview of the BITS pipeline.

3 BITS Benchmark Framework

3.1 Datasets Details

We collect eleven datasets spanning nine domains, such as activity and climate. For datasets containing samples with substantially different durations or numbers of observations, we select a subset of samples and variables so that the retained samples have comparable temporal spans and observation counts, facilitating consistent window construction across samples. Then we process them into a unified standard format. Detailed descriptions and statistical information about the datasets can be found in Appendix A.2. We adopt different splitting strategies according to the observation counts of each dataset. For sparsely observed datasets (e.g., PhysioNet, USHCN, and APTC), we adopt sample-aware splitting which randomly divides all samples in each dataset, because time-aware splitting would leave insufficient observations for meaningful forecasting window construction. For the remaining datasets which provide sufficient observations, we adopt time-aware splitting by dividing each sample chronologically and merging the corresponding segments across samples. This overcomes the shortcoming of existing irregular time series forecasting methods that do not consider the time order in dataset splitting even when the observation counts of the dataset are sufficient to support time-aware splitting.

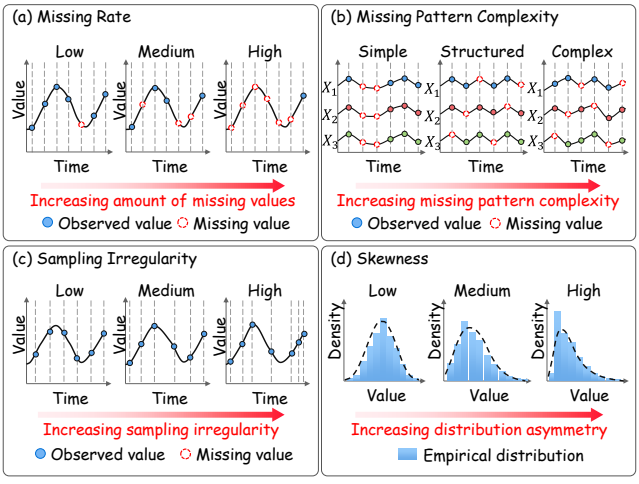


Figure 3: Illustration of four irregularity characteristics.

We characterize dataset irregularity from four dimensions, including missing rate, missing pattern complexity, sampling irregularity and skewness, as shown in Figure 3. Detailed definitions and formulas are provided in Appendix A.4. To further illustrate dataset diversity, we visualize the distributions of the four irregularity characteristics in a shared PCA space in Figure 9 in Appendix A.4. Compared with commonly used datasets such as PhysioNet, Human Activity, and USHCN, BITS covers a broader region of the characteristic space, highlighting the rich diversity of irregularity characteristics across its eleven datasets.

3.2 COMPARED BASELINES

To investigate the advantages and limitations of different methods, we evaluate six categories of irregular time series forecasting methods, including imputation-based, temporal segmentation and aggregation, multi-scale modeling, differential equation-based modeling, data-format transformation and sampling context-aware modeling methods. For imputation-based methods, we choose mTAN (Shukla & Marlin, 2021). In terms of temporal segmentation and aggregation methods, we include tPatchGNN (Zhang et al., 2024), Hi-Patch (Luo et al., 2025), APN (Liu et al., 2025a), TFMixer (Qiu et al., 2026), KAFNet (Zhou et al., 2026) and TiWeaver (Li et al., 2026). For multi-scale modeling methods, we include Warpformer (Zhang et al., 2023). For differential equation-based modeling methods, we choose NeuralFlows (Bilos et al., 2021), GNeuralFlow (Mercatali et al., 2024) and CRU (Schirmer et al., 2022). In terms of data-format transformation methods, we include GraFITi (Yalavarthi et al., 2024), SeFT (Horn et al., 2020), HyperIMTS (Li et al., 2025a) and ASTGI (Liu et al., 2026). For sampling context-aware modeling methods, we choose GRU-D (Che et al., 2016). For a more comprehensive and fair comparison, we also include multivariate time series forecasting methods: DLinear (Zeng et al., 2023), TimesNet (Wu et al., 2023), iTransformer (Liu et al., 2024a), PatchTST (Nie et al., 2023), Crossformer (Zhang & Yan, 2023), DUET (Qiu et al., 2025), SRSNet (Wu et al., 2025), and time series foundation models: MIRA (Li et al., 2025b), Sundial (Liu et al., 2025b) and Timer (Liu et al., 2024b).

3.3 EVALUATION METRICS

Existing methods primarily rely on error-based metrics such as MSE and MAE, and report prediction performance based on normalized data. However, the error-based metrics can be heavily influenced by the data range, making them less intuitive for interpretation. Also, assessing prediction performance on normalized data may result in low error values, which may obscure the practical interpretation of forecasting errors. Furthermore, these metrics are based on distance measures, comparing the predicted values with the ground truth on a per-time-step basis. In scenarios where the model outputs near-constant predictions for the entire forecast horizon, the error-based metrics may fail to provide a meaningful assessment of model performance, as shown in Figure 1 (c). Therefore, the evaluation metrics should go beyond simple point-wise distance measures and also consider pattern alignment and temporal structure.

In order to conduct an effective, comprehensive, and consistent evaluation of the prediction results, we evaluate on both normalized and denormalized data and incorporate additional metrics, including Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (SMAPE), Root Mean Squared Error (RMSE), Weighted Absolute Percent Error (WAPE), and Modified Symmetric Mean Absolute Percentage Error (MSMAPE). In addition, we propose incorporating multi-dimensional metrics that evaluate both value accuracy and pattern similarity, including Pearson Correlation (PCC), Coefficient of Determination (R^2), Directional Accuracy (DA), Dynamic Time Warping (DTW), Change Point Number Difference (CPN) and Information Entropy Ratio (IER). This allows for a more holistic assessment of a model's ability to learn and forecast complex temporal structures inherent in irregular time series data. The formulations and detailed descriptions of these metrics are provided in Appendix A.3.

3.4 UNIFIED PIPELINE

Different data preprocessing procedures and evaluation criteria lead to differences in model performance. To ensure fair and consistent evaluation, we designed a unified pipeline comprising a data layer, a method layer, an evaluation layer, and a reporting layer. A detailed description of each layer is provided below.

Data Layer stores comprehensive irregular time series from different domains, encompassing diverse sampling intervals, channel asynchrony and irregularity characteristics. For data preprocessing, we apply a normalization procedure, with min-max normalization as the default choice. The loader selects different dataset splitting strategies, including sample-aware or time-aware splitting, according to the observation counts of the dataset. This layer ensures that model performance is evaluated under consistent conditions by maintaining a consistent data pattern and structure.

Method Layer integrates dozens of different types of irregular time series forecasting methods, as well as multiple multivariate time series forecasting methods and time series foundation models. This layer

features a unified model interface and scalability, ensuring consistency in evaluation and compatibility with different models. Subsequent work can build upon this foundation by writing simple and general-purpose interfaces to easily integrate their new models, thus enabling fair comparisons. This layer ensures the versatility and flexibility of BITS by providing support for a variety of models.

Evaluation Layer includes various evaluation metrics, including commonly used error-based metrics and complementary non-error-based metrics, to enable more comprehensive evaluation. We also address the unreasonable lookback and prediction lengths in existing studies by setting customized and uniform lookback and prediction lengths for fair and effective evaluation. This layer comprehensively and effectively evaluates the performance of various models through multidimensional assessment across multiple metrics and more reasonable history and prediction lengths.

Reporting Layer records experimental configurations, training and evaluation logs, runtime and resource usage, and environment information to ensure reproducibility and traceability. It also provides visualization tools for intuitive analysis of forecasting results. Moreover, BITS summarizes benchmark results across datasets and metrics and generates model rankings and leaderboards, enabling users to conveniently compare different methods and identify competitive models under diverse irregular time series forecasting scenarios.

Users only need to deploy their method at the method layer and configure the configuration file. Then BITS can automatically run the pipeline in Figure 2, enabling users to better understand, compare, and select irregular time series forecasting methods for specific irregular scenarios.

4 Experiments

4.1 Experimental Setup

All experiments are conducted using PyTorch on an NVIDIA 3090 GPU. The models are trained with the MSE loss using the Adam optimizer. To ensure a consistent training objective across trainable baselines, we use MSE as the default loss function, while all other metrics are used only for evaluation and are not directly optimized during training, allowing us to assess model behavior under a unified optimization setting. For dataset splitting, we set it to 6:2:2 for training, validation, and test. The training, validation, and test sets are kept consistent for all models. We set customized lookback and prediction lengths for different datasets. The lookback horizons are set to 24 for PhysioNet and USHCN, 6000 for Human Activity and Pamap2, 336 for EPA-Air, 168 for ClusterTrace, 672 for CESNET, 400 for APTC, 730 for FNSPID, 1440 for Seabirds, 365 for GDELT. The prediction lengths are set to {1, 6, 12} for PhysioNet and USHCN, {200, 1200, 3000} for Human Activity, {200, 1200, 3000} for Pamap2, {24, 72, 168} for EPA-Air, {24, 48, 72} for ClusterTrace, {24, 168, 336} for CESNET, {50, 100, 200} for APTC, {30, 180, 365} for FNSPID, {60, 360, 720} for Seabirds, {30, 90, 180} for GDELT. For each method, we perform hyperparameter searches across multiple sets and report the best prediction performance. For splitting strategy, we adopt sample-aware splitting for PhysioNet, USHCN, APTC, and time-aware splitting for the remaining datasets.

4.2 Experimental Results

4.2.1 Performance Comparison on Irregular Time Series Forecasting

We evaluate a wide range of models on 11 irregular time series datasets, reporting the average MSE over three prediction lengths in Table 2 and Table 3, and ranking irregular models using two criteria in Figure 4. Overall, irregular forecasting methods generally outperform regular forecasting models, demonstrating the benefits of explicitly handling irregular observations. Among them, TiWeaver, TFMixer and ASTGI demonstrate the strongest overall performance. TiWeaver and ASTGI achieve the lowest average MSE rank, and TFMixer ranks first in DTW. These results highlight the strong overall performance of temporal segmentation and aggregation methods and data-format transformation methods, suggesting that reorganizing irregular observations into informative temporal segments (e.g., patches) or structured representations (e.g., graphs and hypergraphs) is a promising direction for irregular forecasting.

However, performance varies considerably within these categories. APN and tPatchGNN excel on the highly sparse PhysioNet, but tPatchGNN performs poorly on USHCN and EPA-Air, where TFMixer remains competitive, suggesting that temporal segmentation (e.g., patching) alone may be insufficient and that how information is integrated across temporal representations also matters.

Table 2: Average performance comparison of irregular time series forecasting models in terms of MSE ($\times 10^{-2}$) over three prediction lengths. Full results are provided in Appendix A.5.2.

Dataset	TFM	TiW	AST	APN	KAF	HiP	HIMTS	GNF	tPGNN	GFT	Warp	CRU	NF	mTAN	SeFT	GRU-D
PhysioNet	0.558	0.567	0.518	0.471	0.567	0.551	0.588	1.040	**0.468**	0.567	1.125	1.167	0.996	0.890	1.177	0.957
Human Activity	**0.363**	0.402	0.390	0.366	0.383	0.395	0.391	1.841	0.383	0.378	1.393	0.897	1.004	0.587	1.402	0.561
USHCN	**50.48**	56.08	52.36	54.86	64.47	58.32	68.96	59.13	69.21	59.58	53.19	63.49	57.26	68.15	76.41	68.96
CESNET	2.327	**2.156**	2.160	2.319	2.351	2.435	2.657	2.339	2.334	2.281	3.049	2.338	2.331	2.334	4.396	2.424
Pamap2	**0.599**	0.628	0.646	0.964	1.005	0.861	0.621	1.804	0.852	0.688	-	1.297	1.550	0.872	1.874	1.775
EPA-Air	0.791	0.813	**0.778**	1.731	1.539	1.059	1.009	2.236	1.579	0.838	2.268	1.505	1.733	1.995	2.694	2.050
ClusterTrace	4.199	4.332	**4.177**	4.317	4.492	4.301	4.406	4.576	4.446	4.498	4.627	4.937	4.517	4.766	4.910	4.504
APTC	**0.381**	0.381	0.383	0.382	0.381	0.382	0.381	0.383	0.602	0.381	0.535	0.381	0.384	0.382	0.426	0.381
FNSPID	0.332	**0.300**	1.069	0.785	0.373	0.662	0.641	0.938	0.893	0.885	1.801	1.132	0.727	0.660	1.722	1.175
GDELT	1.859	1.829	**1.826**	1.841	1.837	1.955	1.849	1.865	1.859	1.953	2.307	1.864	1.843	1.918	2.720	1.837
Seabirds	0.032	0.026	**0.022**	0.081	0.097	0.098	0.049	0.134	0.117	0.075	0.117	0.114	0.170	0.116	0.141	0.164

Note. "-" indicates out-of-memory (OOM) even with the batch size set to 1. *Abbreviations.* TFM: TFMixer; TiW: TiWeaver; AST: ASTGI; KAF: KAFNet; HiP: Hi-Patch; HIMTS: HyperIMTS; GNF: GNeuralFlow; tPGNN: tPatchGNN; GFT: GraFITi; Warp: Warpformer; NF: NeuralFlows.

A similar disparity is observed among data-format transformation methods: ASTGI and GraFITi perform well overall, whereas SeFT falls behind. These observations indicate that transforming irregular observations into structured representations is beneficial, but the effectiveness of the resulting representation and subsequent dependency modeling remains crucial. Furthermore, the differences between the MSE and DTW rankings highlight the importance of incorporating non-error-based metrics for comprehensive evaluation.

Moreover, regular models occasionally outperform specialized irregular methods. DUET and SRSNet achieve the lowest average MSE on USHCN and EPA-Air, respectively, despite their high missing rates. This challenges the assumption that explicit irregularity modeling necessarily leads to better forecasting accuracy, which suggests that capturing predictive patterns remains essential, and that current irregular forecasting methods may not always translate their specialized designs into effective representations. Together, these findings highlight the importance of developing models that combine effective irregularity handling with robust temporal and cross-variable representation learning, rather than relying only on increasingly sophisticated irregularity-specific architectures. We further analyze the trade-off between forecasting performance and computational efficiency in terms of inference time and memory usage, as detailed in Appendix A.5.1.

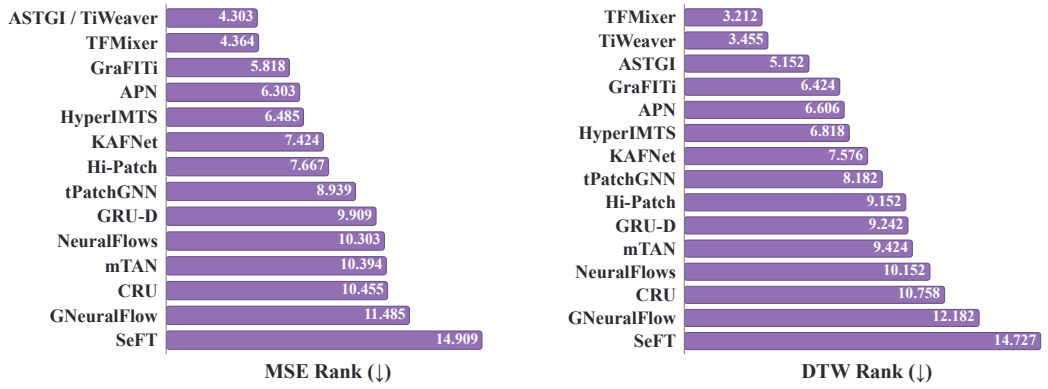


Figure 4: Overall model rankings on BITS based on average per-task MSE and DTW ranks.

4.2.2 Irregular Time Series Forecasting with Time Series Foundation Model

We further evaluate the generalization of time series foundation models (TSFMs) to irregular time series forecasting. As shown in Table 3, we compare three representative TSFMs under the zero-shot setting, reporting the average MSE over three prediction lengths. Detailed results are provided in Appendix A.5.2. Overall, TSFMs generally underperform specialized irregular forecasting methods, particularly on highly sparse datasets such as PhysioNet and USHCN. This suggests that large-scale pretraining alone may not be sufficient to effectively handle missing values, irregular sampling, and asynchronous variables. Among the evaluated TSFMs, MIRA remains competitive on CESNET, ClusterTrace, FNSPID, and Seabirds. Notably, ClusterTrace exhibits the highest sampling irregularity

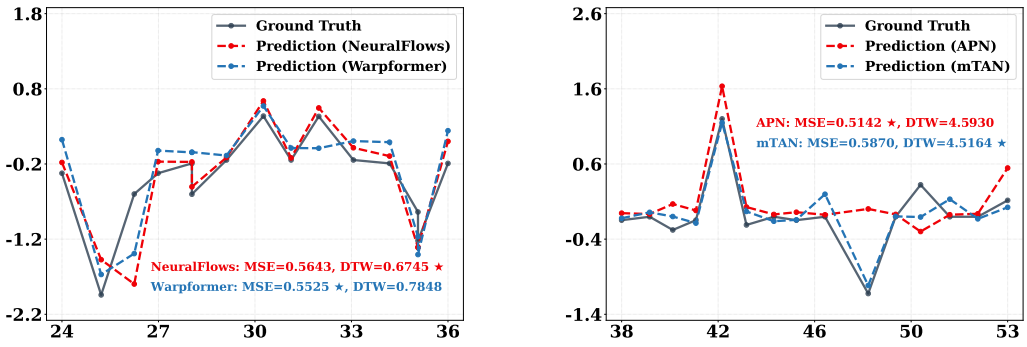
Table 3: Average performance comparison of time series forecasting models and TSFMs in terms of MSE ($\times 10^{-2}$) over three prediction lengths. Full results are provided in Appendix A.5.2.

Dataset	Crossformer	DLinear	PatchTST	TimesNet	iTransformer	DUET	SRSNet	MIRA	Sundial	Timer
PhysioNet	**1.006**	5.819	3.844	4.537	3.847	1.387	1.413	8.740	17.292	12.601
Human Activity	1.260	1.705	0.781	23.338	1.565	**0.390**	0.397	9.425	29.647	23.550
USHCN	73.40	76.55	76.22	75.55	76.03	**49.01**	55.05	73.85	79.39	77.40
CESNET	2.566	5.332	3.500	7.377	5.389	2.276	3.722	**2.231**	5.132	5.133
Pamap2	0.772	1.565	2.431	0.922	2.860	0.559	**0.553**	5.711	11.063	11.620
EPA-Air	1.955	2.808	1.082	1.293	1.163	0.718	**0.672**	6.915	8.487	8.160
ClusterTrace	6.355	8.142	7.100	9.309	7.697	4.303	**4.269**	4.337	10.311	10.372
APTC	0.380	**0.375**	0.377	0.375	0.375	0.387	0.390	0.413	0.379	0.399
FNSPID	3.213	4.813	0.881	6.872	2.811	0.539	5.626	**0.439**	7.597	7.290
GDELT	**1.894**	15.587	4.760	9.921	7.743	1.969	1.921	3.884	16.282	15.502
Seabirds	2.756	0.084	0.070	0.042	0.038	0.027	**0.019**	0.038	0.045	0.116

but a relatively low missing rate, whereas PhysioNet and USHCN suffer from severe missingness and complex missing patterns. This contrast suggests that irregular sampling and missing-data structures pose distinct challenges for zero-shot transfer, with the latter potentially creating a greater obstacle. These findings motivate irregularity-aware pretraining and adaptation strategies that explicitly account for missing observations and asynchronous variables.

4.2.3 Case Study on the Limitation of Error-based Metrics

To further illustrate the limitation of error-based metrics, we visualize two representative cases on USHCN in Figure 5. In Figure 5a, Warpformer achieves a lower MSE than NeuralFlows, while NeuralFlows obtains a lower DTW and better follows the local rising and falling patterns of the ground truth. A similar inconsistency is observed in Figure 5b: APN achieves a lower MSE than mTAN, while mTAN has a lower DTW and better captures several sharp temporal variations. These examples show that lower point-wise errors do not necessarily imply better temporal shape modeling. Therefore, non-error-based metrics provide complementary information for evaluating whether predictions preserve the underlying temporal dynamics.



(a) NeuralFlows vs. Warpformer

(b) APN vs. mTAN

Figure 5: Prediction visualization on USHCN.

4.2.4 Comparison of Six Categories of Irregular Forecasting Models

We compare six categories of irregular forecasting methods across different missing rates, as shown in Figure 6. Temporal segmentation and aggregation methods achieve the best performance under most missing rates, and data-format transformation methods also perform competitively under high missing rates. Their advantages may stem from their ability to effectively utilize limited observations. For example, temporal aggregation consolidates sparse observations into informative local representations, while graph- and hypergraph-based transformations exploit complementary information across variables. Imputation-based and differential equation-based methods exhibit higher errors across different missing rates, potentially due to the difficulty of reconstructing missing values or estimating continuous dynamics from limited observations. These findings suggest that effectively utilizing available observations may be more beneficial than attempting to reconstruct

fine-grained temporal information under severe missingness. In addition, the sampling context-aware modeling method achieves the best performance on GDELT. Sampling context-aware modeling (e.g., GRU-D) explicitly incorporates observation masks and time gaps, potentially allowing it to capture useful temporal information without complex representation transformations when observations are relatively abundant. This suggests that explicit sampling-context modeling can provide an effective alternative to temporal aggregation methods under certain irregularity conditions.

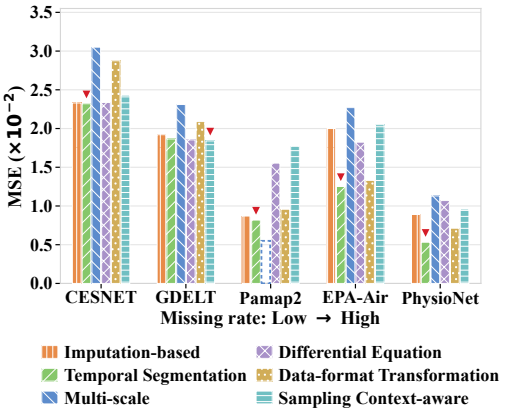


Figure 6: Comparison of six model categories across datasets with varying missing rates. Red triangles indicate the best-performing category. The missing bar on Pamap2 indicates OOM.

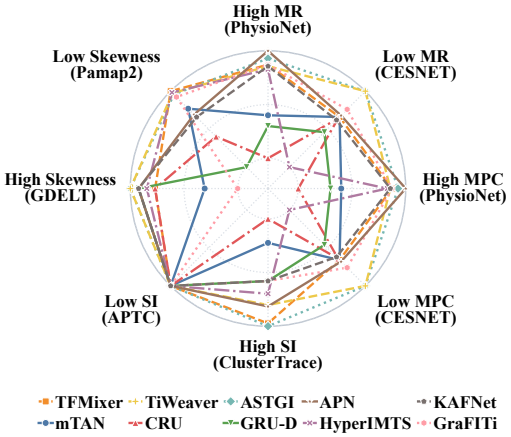


Figure 7: The MSE results for different irregularity characteristics. MR, MPC, and SI denote missing rate, missing pattern complexity, and sampling irregularity, respectively.

4.2.5 Performance on different irregularity characteristics

We further evaluate ten representative models across different irregularity characteristics, as shown in Figure 7. No model consistently performs best across all irregularity settings. APN achieves the best performance on PhysioNet with high missing rate and missing pattern complexity, whereas its advantage does not extend to datasets with other characteristics. This may benefit from its adaptive patching mechanism, which flexibly aggregates informative observations under severe sparsity. ASTGI performs strongly on CESNET, ClusterTrace, and GDELT, spanning low missingness, high sampling irregularity and high skewness. Its adaptive spatio-temporal graph construction may provide greater flexibility in capturing dependencies among irregular observations under these heterogeneous conditions. TiWeaver achieves the lowest MSE on CESNET and remains competitive on GDELT, suggesting the adaptability of its contextual patching strategy. Meanwhile, TFMixer achieves the best performance on Pamap2 with low skewness, despite ranking strongly but not uniformly best in the other settings. These complementary strengths indicate that different modeling strategies respond differently to irregularity characteristics, and strong overall performance does not imply universal superiority. These results highlight the importance of evaluating models across diverse irregular characteristics rather than relying on a limited number of datasets and irregularity settings.

5 Conclusion

In this paper, we present BITS, a fair and comprehensive benchmark for irregular time series forecasting. BITS provides eleven multi-domain datasets with diverse irregularity characteristics, together with a standardized and scalable pipeline covering preprocessing, splitting, modeling, evaluation, and reporting. It further characterizes datasets from multiple irregularity dimensions and incorporates both error-based and non-error-based metrics for more comprehensive evaluation. Extensive experiments reveal that TiWeaver and TFMixer achieve consistently strong overall performance across multiple evaluation criteria, with ASTGI also demonstrating competitive results. Nevertheless, model performance varies across irregularity characteristics, and regular forecasting models can outperform specialized irregular methods in certain settings. TSFMs also show limitations under severe missingness and complex missing patterns, motivating irregularity-aware pretraining and adaptation. Our analysis also shows that error-based metrics alone may not fully reflect forecasting performance. Overall, BITS provides a reproducible and extensible evaluation platform for systematically comparing irregular time series forecasting methods.

AI USE STATEMENT

In this work, we did not use generative AI tools for tasks requiring mandatory disclosure, and the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to assist with language polishing, improving the readability and organization of the manuscript, and identifying potentially relevant literature. We have reviewed all AI-assisted work and manually verified the resulting revisions and literature suggestions. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide an anonymous implementation of BITS to facilitate reproducibility. The benchmark includes standardized data preprocessing, model integration, evaluation, and reporting procedures. Detailed dataset descriptions and preprocessing settings are provided in Appendix A.2, evaluation metrics are described in Appendix A.3, irregularity characteristics are defined in Appendix A.4, and complete experimental results are reported in Appendix A.5.2. Experimental settings, including dataset splitting strategies, lookback lengths, prediction lengths, and training configurations, are described in Sections 3 and 4.1. The source code is available at the anonymous repository provided in the paper.

REFERENCES

Marin Bilos, Johanna Sommer, Syama Sundar Rangapuram, Tim Januschowski, and Stephan Günnemann. Neural flows: Efficient alternative to neural odes. In *NeurIPS*, pp. 21325–21337, 2021.

Ella Browning, Mark Bolton, Ellie Owen, Akiko Shoji, Tim Guilford, and Robin Freeman. Predicting animal behaviour using deep learning: Gps data alone accurately predict diving in seabirds. *Methods in Ecology and Evolution*, 9(3):681–692, 2018.

Hung Bui, Harikrishna Warrier, and Yogesh Gupta. Benchmarking with mimic-iv, an irregular, spare clinical time series dataset. *arXiv preprint arXiv:2401.15290*, 2024.

Ching Chang, Jeehyun Hwang, Yidan Shi, Haixin Wang, Wei Wang, Wen-Chih Peng, and Tien-Fu Chen. Time-imm: A dataset and benchmark for irregular multimodal multivariate time series. In *NeurIPS*, 2025.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David A. Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *CoRR*, abs/1606.01865, 2016.

Vinícius M. A. de Souza. Asphalt pavement classification using smartphone accelerometer and complexity invariant distance. *Eng. Appl. Artif. Intell.*, 74:198–211, 2018.

Zihan Dong, Xinyu Fan, and Zhiyuan Peng. FNSPID: A comprehensive financial news dataset in time series. In Ricardo Baeza-Yates and Francesco Bonchi (eds.), *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pp. 4918–4927. ACM, 2024. doi: 10.1145/3637528.3671629. URL https://doi.org/10.1145/3637528.3671629.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten M. Borgwardt. Set functions for time series. In *ICML*, volume 119, pp. 4353–4363, 2020.

Josef Koumar, Karel Hynek, Tomáš Čejka, and Pavel Šiška. Cesnet-timeseries24: Time series dataset for network traffic anomaly detection and forecasting. *Scientific Data*, 12(1):338, 2025.

Kalev Leetaru and Philip A Schrodt. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pp. 1–49, 2013.

Boyuan Li, Yicheng Luo, Zhen Liu, Junhao Zheng, Jianming Lv, and Qianli Ma. Hyperimts: Hypergraph neural network for irregular multivariate time series forecasting. In *ICML*, 2025a.

Hao Li, Bowen Deng, Chang Xu, Zhiyuan Feng, Viktor Schlegel, Yu-Hao Huang, Yizheng Sun, Jingyuan Sun, Kailai Yang, Yiyao Yu, and Jiang Bian. MIRA: medical time series foundation model for real-world health data. In *NeurIPS*, 2025b.

Zhe Li, Xiangfei Qiu, Peng Chen, Yihang Wang, Hanyin Cheng, Yang Shu, Jilin Hu, Chenjuan Guo, Aoying Zhou, Christian S Jensen, et al. Tsfm-bench: A comprehensive and unified benchmark of foundation models for time series forecasting. In *KDD*, pp. 5595–5606, 2025c.

Zhe Li, Jindong Tian, Hao Miao, Zhi Lei, Chenjuan Guo, and Bin Yang. Tiweaver: Unified temporal dynamics modeling via contextual patching. In *KDD*, pp. 2933–2944, 2026.

Yubo Liang, Zezhi Shao, Fei Wang, Zhao Zhang, Tao Sun, and Yongjun Xu. Basicts: An open source fair multivariate time series prediction benchmark. In *International symposium on benchmarking, measuring and optimization*, pp. 87–101, 2022.

Xvyuan Liu, Xiangfei Qiu, Xingjian Wu, Zhengyu Li, Chenjuan Guo, Jilin Hu, and Bin Yang. Rethinking irregular time series forecasting: A simple yet effective baseline. *CoRR*, abs/2505.11250, 2025a.

Xvyuan Liu, Xiangfei Qiu, Hanyin Cheng, Xingjian Wu, Guo Guo, Bin Yang, and Jilin Hu. Astgi: Adaptive spatio-temporal graph interactions for irregular multivariate time series forecasting. In *ICLR*, volume 2026, pp. 23940–23958, 2026.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *ICLR*, 2024a.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: generative pre-trained transformers are large time series models. In *ICML*, pp. 32369–32399, 2024b.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *ICML*, 2025b.

Yicheng Luo, Bowen Zhang, Zhen Liu, and Qianli Ma. Hi-patch: Hierarchical patch GNN for irregular multivariate time series. In *ICML*, 2025.

Matthew J Menne, Claude N Williams Jr, and Russell S Vose. Long-term daily and monthly climate records from stations across the contiguous united states (us historical climatology network). Technical report, Environmental System Science Data Infrastructure for a Virtual Ecosystem; CDIAC, 2015.

Giangiacomo Mercatali, André Freitas, and Jie Chen. Graph neural flows for unveiling systemic interactions among irregularly sampled time series. In *NeurIPS*, 2024.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *ICLR*, 2023.

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

Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. DUET: dual clustering enhanced multivariate time series forecasting. In *KDD*, pp. 1185–1196, 2025.

Xiangfei Qiu, Kangjia Yan, Xvyuan Liu, Xingjian Wu, and Jilin Hu. Bridging time and frequency: A joint modeling framework for irregular multivariate time series forecasting. In *ICML*, 2026.

Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *ISWC*, pp. 108–109, 2012.

Mona Schirmer, Mazin Eltayeb, Stefan Lessmann, and Maja Rudolph. Modeling irregular time series with continuous recurrent units. In *ICML*, volume 162, pp. 19388–19405, 2022.

Zezhi Shao, Fei Wang, Yongjun Xu, Wei Wei, Chengqing Yu, Zhao Zhang, Di Yao, Tao Sun, Guangyin Jin, Xin Cao, Gao Cong, Christian S. Jensen, and Xueqi Cheng. Exploring progress in multivariate time series forecasting: Comprehensive benchmarking and heterogeneity analysis. *IEEE Trans. Knowl. Data Eng.*, 37(1):291–305, 2025.

Satya Narayan Shukla and Benjamin M. Marlin. Multi-time attention networks for irregularly sampled time series. In *ICLR*, 2021.

Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting in-hospital mortality of icu patients: The physionet computing in cardiology challenge 2012. *Computing in cardiology*, 39:245, 2012.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Chen Wang, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Qizhen Weng, Wencong Xiao, Yinghao Yu, Wei Wang, Cheng Wang, Jian He, Yong Li, Liping Zhang, Wei Lin, and Yu Ding. Mlaas in the wild: Workload analysis and scheduling in large-scale heterogeneous GPU clusters. In *NSDI*, pp. 945–960, 2022.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *ICLR*, 2023.

Xingjian Wu, Xiangfei Qiu, Hanyin Cheng, Zhengyu Li, Jilin Hu, Chenjuan Guo, and Bin Yang. Enhancing time series forecasting through selective representation spaces: A patch perspective. In *NeurIPS*, 2025.

Vijaya Krishna Yalavarthi, Kiran Madhusudhanan, Randolf Scholz, Nourhan Ahmed, Johannes Burchert, Shayan Jawed, Stefan Born, and Lars Schmidt-Thieme. Grafiti: Graphs for forecasting irregularly sampled time series. In *AAAI*, pp. 16255–16263, 2024.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *AAAI*, pp. 11121–11128, 2023.

Jiawen Zhang, Shun Zheng, Wei Cao, Jiang Bian, and Jia Li. Warpformer: A multi-scale modeling approach for irregular clinical time series. In *SIGKDD*, pp. 3273–3285, 2023.

Weijia Zhang, Chenlong Yin, Hao Liu, Xiaofang Zhou, and Hui Xiong. Irregular multivariate time series forecasting: A transformable patching graph neural networks approach. In *ICML*, 2024.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *ICLR*, 2023.

Ziyu Zhou, Yiming Huang, Yanyun Wang, Yuankai Wu, James Kwok, and Yuxuan Liang. Revitalizing canonical pre-alignment for irregular multivariate time series forecasting. In *AAAI*, pp. 29115–29123, 2026.

A Appendix

A.1 Preliminaries

A.1.1 Irregular Time Series

Let a multivariate time series be defined as $\mathcal{X} = \{(t_i, \mathbf{x}_i)\}_{i=1}^{N}$, where $t_i \in \mathbb{R}^+$ denotes the timestamp of the i-th observation and $\mathbf{x}_i \in \mathbb{R}^D$ is a D-dimensional measurement vector. A time series is regular

if timestamps are uniformly spaced, (i.e., $t_{i+1} - t_i = \Delta t, \forall i$). In contrast, a time series is irregular when the intervals are non-uniform: $t_{i+1} - t_i \neq \Delta t$, or $t_{i+1} - t_i \in \mathbb{R}^+$ varies significantly.

Irregularity also manifests in two common forms:

Variable-level asynchronous sampling. Each variable $x^{(d)}$ may have its own observation times:

$$\mathcal{X}^{(d)} = \{(t_i^{(d)}, x_i^{(d)})\}_{i=1}^{N_d}, t_i^{(d)} \neq t_j^{(d')}. \tag{1}$$

Instance-level missingness and sparsity. For a shared timeline $t_1, ..., t_T$, some variables may be missing at certain time steps. A masking matrix $\mathbf{M} \in \{0, 1\}^{T \times D}$ is typically defined as

$$M_{t,d} = \begin{cases} 1, & \text{if } x_{t,d} \text{ is observed,} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Irregular time series arise naturally in domains such as medicine, geoscience, and industrial monitoring due to asynchronous sensors, event-driven sampling, and device failures.

A.1.2 Irregular Time Series Forecasting

Given an irregular multivariate time series $\mathcal{X}_{1:T} = \{(t_i, \mathbf{x_i})\}_{i=1}^{T}$, the goal of irregular time series forecasting (ITSF) is to predict future values over a horizon H:

$$\mathcal{X}_{T+1:T+H} = f_\theta(\mathcal{X}_{1:T}, \mathbf{M}_{1:T}, \Delta_{t_{1:T}}), \tag{3}$$

where f_θ is the forecasting model, $\mathbf{M}_{1:T}$ contains observation masks, and $\Delta t_{1:T}$ captures time gaps between observations: $\Delta t_i = t_i - t_{i-1}$.

Different from regular forecasting, ITSF models must explicitly handle non-uniform temporal intervals. Two main challenges arise:

Temporal Encoding under Irregularity. Models need to account for non-uniform temporal intervals, for example, by feeding time gaps into the model to avoid misleading assumptions of uniform sampling: $\mathbf{z}_i = g(\mathbf{x}_i, \Delta t_i)$, where $g(\cdot)$ is a temporal encoding function (e.g., exponential decay, continuous-time embeddings, or neural ODE dynamics).

Alignment Across Variables. When variables are asynchronously sampled, the model must align their representations: $\mathbf{h}_t = A(\{(t_i^{(d)}, x_i^{(d)}) : t_i^{(d)} \leq t\})$, where $A(\cdot)$ aggregates information with respect to time.

A.2 Datasets

We collect a series of open-source irregular time series datasets that cover a wide range of domains and characteristics for irregular time series forecasting. Detailed descriptions of these datasets are provided below.

PhysioNet (Silva et al., 2012). This dataset includes 12000 IMTS from different patients, each with 41 clinical signals collected irregularly during the first 48 hours of ICU admission. We use the first 24 hours as the observed data to predict the queried values in the subsequent 1/6/12 hours.

Human Activity. This dataset consists of 12 irregularly measured 3D positional variables from sensors worn on the ankles, belts, and chests of five individuals performing various activities. We chunk the original time series into 25 IMTS. The first 6000 milliseconds are used as observed data to predict the sensor positions for the next 200/1200/3000 milliseconds.

USHCN (Menne et al., 2015). The USHCN dataset includes over 150 years of climate data from multiple U.S. stations, covering 5 climate variables. We focus on data from 1114 stations between 1996 and 2000, resulting in 1114 IMTS. Each instance uses data from the previous 24 months to predict the next 1/6/12 month's climate conditions.

CESNET (Koumar et al., 2025). This dataset comprises network traffic data collected from the CESNET academic network, representing a large-scale collection of flow-level statistics. It contains 11 IMTS derived from millions of network flows, where each instance includes 5 features such as packet counts, byte counts, and inter-arrival times recorded over varying observation windows.

Following standard protocol for network traffic analysis, the flows are processed into discrete intervals. We use the first 672 hours as the observed sequence to predict the traffic characteristics and flow types in the subsequent 24/168/336 hours.

Pamap2 (Reiss & Stricker, 2012). The Pamap2 dataset contains data from multiple participants performing different physical activities (such as walking, cycling, and Nordic walking) while wearing three inertial measurement units and a heart rate monitor. The raw sensory data includes multiple dimensions of 3D-acceleration, gyroscope, magnetometer, and temperature readings, along with heart rate. To adapt it for forecasting, the continuous recordings are segmented and 2 IMTS are used, each representing a specific activity window. We select 4 variables for irregular time series forecasting. The first 6000 milliseconds are utilized as observed data to predict the sensor readings for the following 200/1200/3000 milliseconds.

EPA-Air. This dataset is derived from the Air Quality System (AQS), the U.S. Environmental Protection Agency's repository of ambient air quality data. It consists of hourly and daily measurements from multiple monitoring stations across the United States, covering several criteria air pollutants (such as PM2.5, Ozone, and CO) and various meteorological variables. Due to sensor maintenance and varying sampling frequencies, the records form a large-scale collection of IMTS. To simulate real-world environmental monitoring and forecasting, the data is typically organized into thousands of time series instances based on station locations. We select 8 IMTS and 4 variables for irregular time series forecasting. We use the measurements from the previous 336 hours as the observed data to predict the pollutant concentrations and Air Quality Index (AQI) for the following 24/72/168 hours.

ClusterTrace (Weng et al., 2022). The ClusterTrace dataset provides a detailed trace of a production cluster containing multiple machines over a period of time. It captures complex workloads and resource utilization metrics, including CPU, memory, and disk usage for various tasks and containers. Due to the dynamic scheduling and heterogeneous nature of cloud workloads, the resource metrics are recorded as IMTS across thousands of unique job instances. Following common practices in cluster management research, the data is processed to focus on temporal resource demands. We select 3 IMTS and 3 variables for irregular time series forecasting. We use the resource usage data from the first 168 hours as the observed sequence to predict the workload intensity and resource requirements for the subsequent 24/48/72 hours.

APTC (de Souza, 2018). This dataset contains three-axis acceleration measurements collected by smartphone sensors during vehicle travel over different road surfaces. Originally developed for pavement type classification, we adapt it for irregular time series forecasting. The processed dataset contains 2,109 instances with 3 variables. We use the previous 400 steps to predict the acceleration measurements over the subsequent 50/100/200 steps.

FNSPID (Dong et al., 2024). The FNSPID dataset is a comprehensive Financial News and Stock Price Integration Dataset, combining stock price movements with corresponding financial news headlines. It covers thousands of stocks from major exchanges, featuring irregularly timed news events alongside stock price variables such as open, close, and volume. The alignment of sparse news signals with continuous price data results in multiple multi-modal IMTS. We select 10 IMTS and 6 variables for irregular time series forecasting. To facilitate stock trend and volatility forecasting, each instance is organized into windows of sequential market activities. We use the integrated data from the previous 730 days as observed data to predict the stock price fluctuations and market sentiment for the following 30/180/365 days.

GDELT (Leetaru & Schrodt, 2013). This dataset is sourced from the Global Database of Events, Language, and Tone, which monitors world broadcast, print, and web news in multiple languages. It captures physical activities and socio-political events globally, resulting in a massive collection of IMTS with varying temporal resolutions based on event occurrences. The dataset includes multiple variables such as the Goldstein scale, average tone, and event counts across thousands of distinct themes and locations. To adapt it for irregular time series forecasting, we select 6 IMTS instances and 4 variables. We use the event dynamics from the previous 365 days as the observed data to predict the event intensity and regional stability in the subsequent 30/90/180 days.

Seabirds (Browning et al., 2018). This dataset tracks the movement and behavioral patterns of seabirds using high-resolution GPS and tri-axial accelerometer sensors. It contains IMTS representing multiple distinct movement variables, such as altitude, speed, and body orientation, collected from multiple species during foraging trips. Given the irregular nature of wildlife tracking due to varying

signal quality and nesting cycles, the data is segmented into multiple IMTS instances, each spanning a specific duration of flight or diving activity. We select 108 IMTS and 3 variables for irregular time series forecasting. We use the sensor measurements from the first 1440 minutes as the observed data to predict the bird's trajectory and behavioral states in the subsequent 60/360/720 minutes.

Table 4: The statistics of evaluation datasets.

Dataset	Domain	# Variables	# Samples	Avg. # Obs.	Max Length	Missing Rate	Missing Pattern Complexity	Sampling Irregularity	Skewness
PhysioNet	Healthcare	41	12000	74	48.00	0.856	0.614	0.640	0.180
Human Activity	Activity	12	25	6593	304861.22	0.750	0.165	0.396	0.116
USHCN	Climate	5	1114	314	48.00	0.780	0.567	0.999	2.001
CESNET	Cybersecurity	5	11	1700	1999.87	0.000	0.000	0.637	1.497
Pamap2	Healthcare	4	2	40601	540880.10	0.229	0.122	0.005	-0.032
EPA-Air	Environment	4	8	5273	6575.00	0.708	0.171	0.057	1.505
ClusterTrace	Engineering	3	3	5005	1525.58	0.057	0.119	4.111	0.795
APTC	Engineering	3	2109	399	1232.00	0.001	0.001	0.000	-0.060
FNSPID	Finance	6	10	3495	5303.00	0.001	0.001	0.606	1.100
GDELT	Social	4	6	1684	1460.69	0.042	0.188	1.279	3.140
Seabirds	Nature	3	108	2442	10078.00	0.001	0.001	0.283	0.691

A.3 METRICS

The metrics are defined as follows.

$$
\begin{aligned}
MSE &= \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2, \qquad MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i|, \\
MAPE &= \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i-\hat{y}_i|}{|y_i|}\times 100\%, \qquad SMAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i-y_i|}{(|y_i|+|\hat{y}_i|)/2}, \\
RMSE &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}, \qquad WAPE = \frac{\sum_{i=1}^{n}|y_i-\hat{y}_i|}{\sum_{i=1}^{n}|y_i|}, \\
MSMAPE &= \frac{100\%}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i-y_i|}{\max(|y_i|+|\hat{y}_i|+\epsilon,\,0.5+\epsilon)/2}, \\
PCC &= \frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}}, \\
R^2 &= 1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}, \qquad CPN = |CP(\hat{\mathbf{y}})-CP(\mathbf{y})|, \\
DA &= \frac{1}{n-1}\sum_{i=1}^{n-1}\mathbb{I}[sign(y_{i+1}-y_i)=sign(\hat{y}_{i+1}-\hat{y}_i)], \\
DTW &= \min_{\pi}\sum_{(i,j)\in\pi}\|\hat{y}_i-y_j\|, \qquad IER = \frac{H(\hat{\mathbf{y}})}{H(\mathbf{y})}.
\end{aligned}
\tag{4}
$$

where y_i represents the actual value, $\hat{y}_i$ represents the predicted value, n is the number of observations, $sign(\cdot)$ is the sign function, and $\mathbb{I}[\cdot]$ is the indicator function which takes the value 1 if the condition inside the brackets is met, and 0 otherwise. π denotes a valid alignment path. $H(\cdot)$ denotes the information entropy of the time series. $CP(\mathbf{y}) = \sum_{i=3}^{n}\mathbb{I}\left(sign(\Delta y_i) \neq \text{sign}(\Delta y_{i-1})\right)$. We set the parameter ϵ to its proposed default of 0.1. For error-based metrics, lower values indicate better performance. Below we introduce the non-error-based metrics.

- Pearson Correlation (PCC) measures the linear relationship between the predicted series and the ground-truth series. Its value ranges from -1 to 1, where higher values indicate

that the model better captures the trend of the time series. A value of 0 indicates no linear correlation, while negative values indicate an inverse trend.

- Coefficient of Determination (R^2) measures how well the model's predictions account for the variance in the true series. Its value ranges from $-\infty$ to 1, where values closer to 1 indicate that the model better captures the variability of the true data. A negative value suggests that the model performs worse than a simple mean-based prediction.

- Directional Accuracy (DA) measures whether the model correctly predicts the direction of change in the time series. Its value ranges from 0 to 1, with higher values indicating a better ability to capture the trend direction.

- Dynamic Time Warping (DTW) measures the overall similarity between the predicted series and the ground-truth series under a non-linear alignment of the time axis. Lower values indicate that the predicted series is more similar to the ground-truth series in terms of overall shape, reflecting better forecasting performance.

- Change Point Number Difference (CPN) measures the difference in the number of change points between the predicted series and the ground-truth series. Smaller values indicate that the predicted series is closer to the ground-truth in terms of structural changes and periodic patterns, reflecting better forecasting performance.

- Information Entropy Ratio (IER) measures the consistency between the predicted series and the ground-truth series in terms of complexity and uncertainty. Values closer to 1 indicate that the predicted series matches the ground-truth in terms of complexity and information content, reflecting better forecasting performance. Values significantly less than 1 suggest that the predicted series is overly smooth, while values significantly greater than 1 indicate that the prediction contains excessive noise.

A.4 Irregular Time Series Characteristics

Missing Rate. Missing Rate measures the proportion of missing values in the dataset based on the original observation mask, reflecting the overall degree of data incompleteness. Algorithm 1 details the calculation process.

Algorithm 1 Calculating Missing Rate

Input: Mask matrix $M \in \{0,1\}^{T \times D}$ (1 for observed, 0 for missing)
Output: Missing rate value $\rho \in [0,1]$
1: $T, D \leftarrow$ dimensions of M
2: $total_elements \leftarrow T \times D$
3: $missing_count \leftarrow \sum_{i=1}^{T} \sum_{j=1}^{D} (1 - M_{ij})$
4: **return** $\rho \leftarrow \frac{missing_count}{total_elements}$

Missing Pattern Complexity. Missing Pattern Complexity measures the structural complexity and uncertainty of missing patterns across different time steps based on the original observation mask. We use missing entropy to represent the missing pattern complexity. Algorithm 2 details the calculation process.

Algorithm 2 Calculating Missing Entropy

Input: Mask matrix $M \in \{0,1\}^{T \times D}$
Output: Normalized missing entropy $\eta \in [0,1]$
1: Define missing pattern vectors $\mathbf{r}_i = (1 - M_{i1}, \ldots, 1 - M_{iD}) \in \{0,1\}^D$ for $i = 1 \ldots T$
2: Identify K unique patterns $\{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_K\}$ from $\{\mathbf{r}_1, \ldots, \mathbf{r}_T\}$
3: **for** each unique pattern $\boldsymbol{\pi}_k$ **do**
4: $\quad p_k \leftarrow \frac{1}{T} \sum_{i=1}^{T} 1(\mathbf{r}_i = \boldsymbol{\pi}_k)$
5: **end for**
6: $H \leftarrow -\sum_{k=1}^{K} p_k \log(p_k)$
7: $H_{max} \leftarrow \min(\log T, D \log 2)$
8: **return** $\eta \leftarrow \frac{H}{H_{max} + \epsilon}$ {where $\epsilon = 10^{-12}$}

Sampling Irregularity. Sampling Irregularity measures the irregularity of sampling times using the coefficient of variation of adjacent sampling intervals. We use coefficient of variation of adjacent sampling intervals to represent the sampling irregularity. Algorithm 3 details the calculation process.

Algorithm 3 Calculating Sampling Irregularity

Input: Observation time stamps $t = \{t_1, t_2, \ldots, t_T\}$
Output: Coefficient of variation for intervals CV_{time}
1: Sort t such that $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(T)}$
2: **for** $i = 1$ **to** $T - 1$ **do**
3: $\quad \Delta_i \leftarrow t_{(i+1)} - t_{(i)}$
4: **end for**
5: $\mu_\Delta \leftarrow \text{mean}(\Delta_1, \ldots, \Delta_{T-1})$
6: $\sigma_\Delta \leftarrow \text{std}(\Delta_1, \ldots, \Delta_{T-1})$
7: **return** $CV_{time} \leftarrow \frac{\sigma_\Delta}{\mu_\Delta + \epsilon}$

Skewness. Skewness measures the overall asymmetry of the data distribution by averaging the skewness values computed separately for each variable. Algorithm 4 details the calculation process.

Algorithm 4 Calculating Average Skewness

Input: Time series $X \in \mathbb{R}^{T \times D}$, Mask matrix $M \in \{0,1\}^{T \times D}$
Output: Average skewness value γ_{avg}
1: $\mathcal{J} \leftarrow \emptyset$ {Set of valid variables}
2: **for** each variable $j = 1$ **to** D **do**
3: $\quad$ Extract observed values $x^{(j)} = \{X_{ij} \mid M_{ij} = 1\}$
4: $\quad$ **if** $|x^{(j)}| >$ threshold **then**
5: $\quad\quad \mu_j, \sigma_j \leftarrow \text{mean}(x^{(j)}), \text{std}(x^{(j)})$
6: $\quad\quad \gamma_j \leftarrow \frac{\frac{1}{n_j} \sum_{k=1}^{n_j} (x_k^{(j)} - \mu_j)^3}{\sigma_j^3}$
7: $\quad\quad \mathcal{J} \leftarrow \mathcal{J} \cup \{j\}$
8: $\quad$ **end if**
9: **end for**
10: **return** $\gamma_{avg} \leftarrow \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \gamma_j$

A.5 Other Experiment results

A.5.1 Model Efficiency Analysis

We compare six categories of irregular forecasting methods on USHCN in terms of MSE, inference time, and memory usage, as shown in Figure 8. Among temporal segmentation and aggregation methods, APN achieves the lowest MSE and fastest inference with moderate memory consumption,

demonstrating that effective temporal aggregation does not necessarily require substantial computational resources. However, efficiency varies considerably within this category. TiWeaver consumes less memory than TFMixer but incurs higher inference latency, while Hi-Patch requires substantially more memory, potentially due to the additional overhead of hierarchical patch-graph modeling. A similar disparity appears among data-format transformation methods, where ASTGI incurs higher latency than GraFITi and HyperIMTS, possibly reflecting the computational cost of adaptive spatio-temporal graph construction. Differential equation-based methods generally incur longer inference times, particularly GNeuralFlow, which suggests that modeling continuous dynamics can introduce substantial computational overhead. In contrast, imputation-based mTAN achieves low inference latency but requires relatively high memory, while sampling context-aware GRU-D maintains low memory consumption through relatively lightweight recurrent modeling. Multi-scale Warpformer also incurs additional latency, potentially associated with temporal warping and multi-scale representation learning. These results highlight that greater modeling complexity does not necessarily yield better accuracy, and low memory consumption does not guarantee fast inference. Therefore, forecasting accuracy, inference latency, and memory usage should be jointly considered when evaluating irregular forecasting methods.

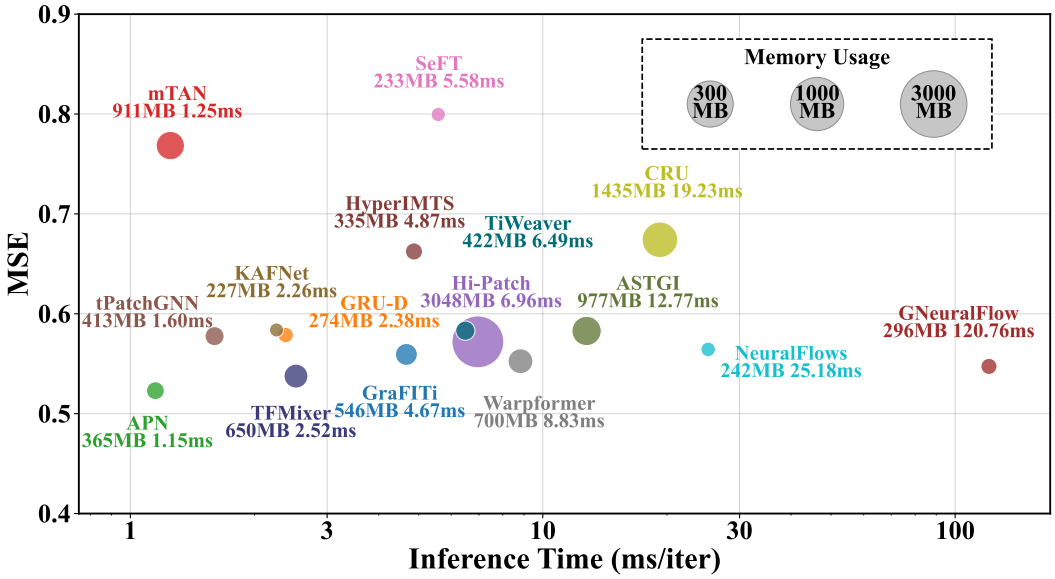


Figure 8: Inference efficiency comparison under BITS.

A.5.2 Full results of performance comparison on irregular time series forecasting

We provide the full results in Table 5, Table 6, Table 7 and Table 8.

A.6 Leaderboard

To provide an overall comparison, we further rank models in Figure 10. MSE rank ranks models by MSE on each task and averages their ranks across all tasks. Aggregated MSE normalizes the MSE of each model against GRU-D on each task and computes the geometric mean of the resulting relative errors. DTW rank ranks models by DTW on each task and averages their ranks across all tasks. Aggregated DTW normalizes the DTW of each model against GRU-D on each task and computes the geometric mean of the resulting relative errors. Lower values indicate better overall performance. Warpformer participates in per-task rankings where results are available but is excluded from the overall leaderboard due to missing results on Pamap2.

Table 5: Performance comparison of irregular time series forecasting models in terms of MSE ($\times 10^{-2}$). Best results are in **bold** and second-best are underlined.

Dataset	Pred	TFM	TiW	AST	APN	KAF	HiP	HIMTS	GNF	tPGNN	GFT	Warp	CRU	NF	mTAN	SeFT	GRU-D
PhysioNet	1	0.619	0.606	0.553	**0.439**	0.654	0.559	0.493	1.220	0.455	0.594	1.290	1.298	1.136	1.029	1.120	1.082
	6	0.486	0.522	0.470	**0.443**	0.497	0.528	0.481	0.929	0.443	0.444	1.018	1.114	0.943	0.798	1.160	0.909
	12	0.569	0.573	0.530	0.530	0.550	0.566	0.791	0.972	**0.507**	0.663	1.066	1.088	0.909	0.842	1.251	0.879
	Avg	0.558	0.567	0.518	0.471	0.567	0.551	0.588	1.040	**0.468**	0.567	1.125	1.167	0.996	0.890	1.177	0.957
Human Activity	200	**0.254**	0.278	0.277	0.259	0.265	0.292	0.261	1.852	0.266	0.259	0.396	0.819	0.977	0.521	0.841	0.399
	1200	**0.354**	0.390	0.385	0.358	0.373	0.401	0.369	1.877	0.403	0.373	1.704	0.871	1.017	0.540	1.691	0.714
	3000	0.481	0.538	0.509	**0.481**	0.511	0.492	0.542	1.794	0.481	0.502	2.079	1.001	1.019	0.701	1.675	0.569
	Avg	**0.363**	0.402	0.390	0.366	0.383	0.395	0.391	1.841	0.383	0.378	1.393	0.897	1.004	0.587	1.402	0.561
USHCN	1	53.76	58.31	58.28	**52.30**	58.38	57.19	66.24	54.73	57.76	55.93	55.25	67.41	56.43	76.84	79.95	57.87
	6	**48.50**	51.16	49.03	60.86	65.30	63.62	68.48	51.62	74.51	50.98	50.58	61.28	57.50	68.92	74.17	74.00
	12	**49.18**	58.77	49.78	51.42	69.73	54.16	72.17	71.03	75.36	71.83	53.74	61.77	57.86	58.70	75.12	75.01
	Avg	**50.48**	56.08	52.36	54.86	64.47	58.32	68.96	59.13	69.21	59.58	53.19	63.49	57.26	68.15	76.41	68.96
CESNET	24	2.293	**2.153**	2.164	2.298	2.336	2.340	2.297	2.323	2.313	2.234	2.711	2.308	2.304	2.329	3.028	2.617
	168	2.332	**2.156**	2.160	2.307	2.362	2.443	2.696	2.343	2.339	2.307	2.339	2.331	2.348	2.348	5.077	2.326
	336	2.357	2.160	**2.156**	2.353	2.355	2.523	2.978	2.351	2.351	2.303	4.098	2.376	2.340	2.324	5.082	2.328
	Avg	2.327	**2.156**	2.160	2.319	2.351	2.435	2.657	2.339	2.334	2.281	3.049	2.338	2.331	2.334	4.396	2.424
Pamap2	200	0.294	0.274	0.246	0.225	0.305	0.365	0.308	1.619	0.335	**0.216**	-	0.847	1.540	0.350	1.860	1.770
	1200	0.635	0.667	0.651	0.754	0.830	0.722	**0.609**	1.796	0.757	0.697	-	1.574	1.630	1.140	1.881	1.803
	3000	**0.869**	0.943	1.042	1.914	1.879	1.495	0.946	1.996	1.465	1.150	-	1.470	1.480	1.127	1.880	1.752
	Avg	**0.599**	0.628	0.646	0.964	1.005	0.861	0.621	1.804	0.852	0.688	-	1.297	1.550	0.872	1.874	1.775
EPA-Air	24	**0.742**	0.807	0.790	2.093	1.613	1.104	0.851	2.530	1.749	0.791	2.001	1.623	1.747	2.076	2.949	2.401
	72	0.881	0.830	**0.766**	1.454	1.567	1.146	0.841	2.241	1.587	0.946	2.461	1.563	1.979	2.027	2.718	1.546
	168	**0.750**	0.803	0.778	1.647	1.438	0.926	1.336	1.937	1.400	0.778	2.341	1.330	1.474	1.883	2.414	2.202
	Avg	0.791	0.813	**0.778**	1.731	1.539	1.059	1.009	2.236	1.579	0.838	2.268	1.505	1.733	1.995	2.694	2.050
ClusterTrace	24	**3.919**	3.997	4.060	4.250	4.497	4.372	4.470	4.721	4.619	4.478	4.600	4.853	4.471	5.058	4.950	4.630
	48	4.356	4.344	**4.249**	4.465	4.485	4.285	4.316	4.471	4.463	4.459	4.710	4.797	4.471	4.923	4.980	4.464
	72	4.321	4.656	**4.222**	4.237	4.495	4.247	4.432	4.536	4.256	4.557	4.571	5.161	4.609	4.316	4.800	4.418
	Avg	4.199	4.332	**4.177**	4.317	4.492	4.301	4.406	4.576	4.446	4.498	4.627	4.937	4.517	4.766	4.910	4.504
APTC	50	0.391	0.391	**0.390**	0.390	0.391	0.392	0.391	0.395	0.561	0.391	0.571	0.391	0.394	0.392	0.430	0.391
	100	0.377	0.377	0.383	0.379	**0.377**	0.377	0.377	0.378	0.624	0.377	0.634	0.377	0.379	0.377	0.436	0.377
	200	0.376	**0.376**	0.376	0.377	0.376	0.376	0.376	0.376	0.622	0.376	0.401	0.376	0.379	0.376	0.412	0.376
	Avg	**0.381**	0.381	0.383	0.382	0.381	0.382	0.381	0.383	0.602	0.381	0.535	0.381	0.384	0.382	0.426	0.381
FNSPID	30	0.157	**0.155**	1.490	0.606	0.373	0.313	0.839	0.831	0.848	0.527	1.940	1.311	0.805	0.289	1.826	0.359
	180	**0.202**	0.232	0.717	0.318	0.350	0.661	0.555	0.910	0.645	0.663	1.771	1.011	0.764	0.888	1.690	1.289
	365	0.638	0.514	1.001	1.431	**0.396**	1.011	0.528	1.073	1.187	1.464	1.691	1.073	0.612	0.802	1.651	1.877
	Avg	0.332	**0.300**	1.069	0.785	0.373	0.662	0.641	0.938	0.893	0.885	1.801	1.132	0.727	0.660	1.722	1.175
GDELT	30	1.913	1.899	**1.876**	1.902	1.893	1.945	1.882	1.900	1.912	1.878	2.100	1.897	1.902	2.047	2.880	1.891
	90	1.918	**1.874**	1.884	1.902	1.886	1.902	1.914	1.912	1.920	2.105	2.690	1.964	1.902	1.921	2.650	1.889
	180	1.746	**1.713**	1.719	1.719	1.732	2.018	1.751	1.782	1.746	1.876	2.130	1.730	1.724	1.786	2.631	1.731
	Avg	1.859	1.829	**1.826**	1.841	1.837	1.955	1.849	1.865	1.859	1.953	2.307	1.864	1.843	1.918	2.720	1.837
Seabirds	60	0.016	0.018	**0.008**	0.036	0.083	0.103	0.063	0.111	0.116	0.110	0.081	0.063	0.197	0.131	0.121	0.141
	360	0.034	0.020	**0.020**	0.067	0.094	0.061	0.048	0.131	0.098	0.047	0.120	0.130	0.173	0.150	0.140	0.181
	720	0.046	0.040	0.039	0.139	0.114	0.130	**0.035**	0.161	0.138	0.067	0.150	0.150	0.140	0.068	0.161	0.170
	Avg	0.032	0.026	**0.022**	0.081	0.097	0.098	0.049	0.134	0.117	0.075	0.117	0.114	0.170	0.116	0.141	0.164

Note. "-" indicates out-of-memory (OOM) even with the batch size set to 1. Avg is computed from the three displayed prediction-length values.
Abbreviations. TFM: TFMixer; TiW: TiWeaver; AST: ASTGI; KAF: KAFNet; HiP: Hi-Patch; HIMTS: HyperIMTS; GNF: GNeuralFlow; tPGNN: tPatchGNN; GFT: GraFITi; Warp: Warpformer; NF: NeuralFlows.

A.7 Limitations and Future Work

Although BITS covers diverse datasets and forecasting models, it cannot exhaustively represent all real-world irregular scenarios and emerging architectures. Future work will expand the dataset and baseline collections to accommodate new developments in irregular time series forecasting.

Table 6: Performance comparison of irregular time series forecasting models in terms of MAE ($\times 10^{-2}$). Best results are in **bold** and second-best are underlined.

Dataset	Pred	TFM	TiW	AST	APN	KAF	HiP	HIMTS	GNF	tPGNN	GFT	Warp	CRU	NF	mTAN	SeFT	GRU-D
PhysioNet	1	3.765	3.555	3.882	3.554	4.980	4.004	**3.454**	6.061	3.706	3.972	6.646	6.366	5.898	5.576	5.891	5.524
	6	**3.415**	3.636	3.692	3.659	3.763	3.963	3.914	5.548	3.540	3.604	5.250	5.891	5.609	5.093	5.391	5.525
	12	**3.525**	3.729	3.805	3.751	4.110	3.859	5.451	5.869	3.710	4.538	5.453	6.016	5.298	5.052	5.281	5.144
	Avg	**3.568**	3.640	3.793	3.655	4.284	3.942	4.273	5.826	3.652	4.038	5.783	6.091	5.602	5.240	5.521	5.398
Human Activity	200	**2.814**	3.060	3.167	2.974	3.054	3.373	3.028	10.641	3.029	2.963	4.365	6.450	7.523	5.191	6.190	4.341
	1200	**3.568**	3.875	4.001	3.688	3.871	4.149	3.831	10.459	4.156	3.949	10.427	6.724	7.806	5.310	10.294	6.300
	3000	**4.431**	4.794	4.843	4.534	4.826	4.859	5.153	10.411	4.718	4.794	11.317	7.343	7.725	6.190	10.249	5.504
	Avg	**3.604**	3.910	4.004	3.732	3.917	4.127	4.004	10.504	3.968	3.902	8.703	6.839	7.685	5.564	8.911	5.382
USHCN	1	**25.15**	33.81	37.09	30.97	35.16	36.97	50.59	35.49	36.39	33.34	40.90	42.81	35.06	45.15	50.37	36.52
	6	**24.40**	31.15	32.70	41.02	43.41	43.73	43.97	34.17	50.86	33.25	35.67	41.05	37.23	44.83	50.18	49.05
	12	**27.79**	35.54	33.58	35.70	47.81	40.37	46.68	46.62	51.09	52.29	37.28	40.94	39.37	36.67	50.54	48.77
	Avg	**25.78**	33.50	34.46	35.90	42.13	40.36	47.08	38.76	46.11	39.63	37.95	41.60	37.22	42.22	50.36	44.78
CESNET	24	9.723	**9.019**	9.384	10.211	10.363	10.174	10.050	10.262	10.371	9.981	11.545	10.316	10.243	10.382	11.363	11.318
	168	10.035	**9.073**	9.209	10.166	10.369	10.365	12.280	10.392	10.272	10.424	10.162	10.378	10.400	10.383	15.399	10.305
	336	10.259	**9.080**	9.428	10.101	10.414	10.436	13.340	10.358	10.352	10.389	13.600	10.687	10.288	10.367	15.442	10.333
	Avg	10.006	**9.057**	9.340	10.159	10.382	10.325	11.890	10.337	10.332	10.265	11.769	10.460	10.310	10.377	14.068	10.652
Pamap2	200	3.279	2.839	2.898	2.762	3.201	4.386	3.448	9.100	4.063	**2.649**	-	6.820	8.504	4.107	7.571	9.071
	1200	5.481	**4.869**	5.702	6.056	6.138	6.255	5.253	9.644	6.142	5.954	-	8.854	8.891	7.739	7.480	9.955
	3000	6.387	**5.885**	7.556	10.277	10.803	8.859	6.578	10.076	9.481	7.570	-	8.238	8.311	7.852	7.540	9.217
	Avg	5.049	**4.531**	5.385	6.365	6.714	6.500	5.093	9.607	6.562	5.391	-	7.971	8.569	6.566	7.530	9.414
EPA-Air	24	6.125	**6.068**	6.566	11.071	8.813	7.674	6.709	11.791	9.422	6.862	9.968	8.767	9.400	9.762	12.806	10.854
	72	6.632	**6.220**	6.435	8.555	8.668	7.496	6.663	10.811	8.687	7.645	11.736	8.983	10.042	10.469	12.641	8.694
	168	6.419	6.416	**6.365**	9.449	8.385	6.923	9.433	10.343	8.422	6.627	10.793	8.673	8.638	10.018	11.071	10.589
	Avg	6.392	**6.235**	6.455	9.692	8.622	7.364	7.602	10.982	8.844	7.045	10.832	8.808	9.360	10.083	12.173	10.046
ClusterTrace	24	**14.730**	15.149	15.754	15.305	15.502	15.620	15.749	16.318	15.733	15.585	16.000	17.173	15.553	17.625	15.460	15.724
	48	15.305	15.900	**15.226**	15.414	15.479	15.338	15.857	15.674	15.415	15.469	16.871	16.010	15.555	16.125	15.481	15.416
	72	**15.137**	16.400	16.051	16.036	15.422	16.036	15.627	15.840	15.853	15.701	17.101	16.336	15.626	15.275	15.370	15.428
	Avg	**15.057**	15.816	15.677	15.585	15.468	15.665	15.744	15.944	15.667	15.585	16.657	16.506	15.578	16.342	15.437	15.523
APTC	50	4.679	**4.677**	4.683	4.693	4.678	4.684	4.678	4.709	5.737	4.678	5.791	4.679	4.704	4.684	4.974	4.678
	100	4.580	4.580	4.629	4.594	4.580	**4.580**	4.581	4.587	6.084	4.580	6.150	4.581	4.601	4.585	5.084	4.580
	200	4.552	4.551	4.554	4.562	4.551	**4.550**	4.552	4.550	6.060	4.552	4.743	4.552	4.576	4.556	4.850	4.550
	Avg	4.604	**4.603**	4.622	4.616	4.603	4.605	4.604	4.615	5.960	4.603	5.561	4.604	4.627	4.608	4.969	4.603
FNSPID	30	2.770	**2.702**	7.793	6.201	4.626	4.708	7.102	6.591	6.753	5.226	8.630	8.861	6.867	4.125	8.780	4.710
	180	3.214	**2.831**	7.236	4.076	4.427	5.493	5.474	6.690	5.603	5.991	8.140	7.981	6.523	7.127	8.481	9.476
	365	5.466	**4.152**	8.451	9.210	4.902	7.301	5.301	7.592	9.243	9.175	8.590	7.886	5.780	7.315	9.830	10.436
	Avg	3.817	**3.228**	7.827	6.496	4.652	5.834	5.959	6.958	7.200	6.797	8.453	8.243	6.390	6.189	9.030	8.207
GDELT	30	8.427	8.643	**8.397**	8.512	8.591	8.865	8.482	8.583	8.404	8.480	9.320	8.443	8.482	8.967	9.891	8.492
	90	8.484	8.554	8.413	8.603	8.614	8.731	8.515	8.622	**8.314**	9.902	9.741	8.829	8.576	8.585	9.450	8.538
	180	8.384	8.453	8.317	8.420	8.583	10.197	8.413	8.742	**8.272**	9.317	9.280	8.444	8.524	8.656	9.311	8.386
	Avg	8.432	8.550	8.376	8.512	8.596	9.264	8.470	8.649	**8.330**	9.233	9.447	8.572	8.527	8.736	9.551	8.472
Seabirds	60	1.000	0.916	**0.623**	1.466	2.667	2.951	1.898	2.011	2.554	2.839	2.981	2.005	3.569	3.800	2.631	3.111
	360	1.339	**0.953**	1.168	1.940	2.827	1.890	1.634	2.601	2.253	1.797	2.191	2.521	3.025	3.290	2.851	3.961
	720	1.580	1.398	**1.332**	3.275	2.731	2.011	1.384	2.551	2.616	2.220	2.240	2.380	2.974	2.001	2.350	3.030
	Avg	1.306	1.089	**1.041**	2.227	2.742	2.284	1.639	2.388	2.474	2.285	2.471	2.302	3.189	3.030	2.611	3.367

Note. "-" indicates out-of-memory (OOM) even with the batch size set to 1. Avg is computed from the three displayed prediction-length values; best and second-best markings are determined before rounding the averages. *Abbreviations.* TFM: TFMixer; TiW: TiWeaver; AST: ASTGI; KAF: KAFNet; HiP: Hi-Patch; HIMTS: HyperIMTS; GNF: GNeuralFlow; tPGNN: tPatchGNN; GFT: GraFITi; Warp: Warpformer; NF: NeuralFlows.

Table 7: Performance comparison of time series forecasting models and TSFMs in terms of MSE ($\times 10^{-2}$). Best results are in **bold** and second-best are underlined.

Dataset	Pred	Crossformer	DLinear	PatchTST	TimesNet	iTransformer	DuET	SRSNet	MIRA	Sundial	Timer
PhysioNet	1	1.391	6.965	6.692	7.573	6.777	**1.115**	1.142	9.060	7.757	12.829
	6	**0.850**	5.231	2.538	2.934	2.431	1.388	1.428	8.592	22.247	12.462
	12	**0.778**	5.261	2.302	3.105	2.334	1.657	1.669	8.567	21.873	12.512
	Avg	**1.006**	5.819	3.844	4.537	3.847	1.387	1.413	8.740	17.292	12.601
Human Activity	200	0.728	1.647	0.716	23.004	1.627	**0.265**	0.268	6.885	22.453	23.022
	1200	1.064	1.723	0.410	23.318	1.525	**0.370**	0.380	10.081	34.614	23.762
	3000	1.987	1.746	1.218	23.693	1.544	**0.534**	0.542	11.309	31.874	23.866
	Avg	1.260	1.705	0.781	23.338	1.565	**0.390**	0.397	9.425	29.647	23.550
USHCN	1	71.74	79.96	79.02	77.22	78.76	**52.58**	58.14	75.51	84.13	81.35
	6	73.76	74.46	74.48	74.41	74.26	**47.12**	53.21	72.58	76.90	74.98
	12	74.69	75.24	75.16	75.03	75.08	**47.32**	53.80	73.45	77.15	75.88
	Avg	73.40	76.55	76.22	75.55	76.03	**49.01**	55.05	73.85	79.39	77.40
CESNET	24	2.678	5.196	3.661	7.349	5.641	2.313	3.636	**2.221**	4.912	4.983
	168	2.563	5.276	3.417	7.361	4.765	2.252	3.640	**2.249**	5.373	5.311
	336	2.458	5.524	3.423	7.422	5.761	2.262	3.890	**2.224**	5.112	5.105
	Avg	2.566	5.332	3.500	7.377	5.389	2.276	3.722	**2.231**	5.132	5.133
Pamap2	200	0.481	1.150	2.081	0.784	2.159	**0.277**	0.280	3.402	10.418	11.260
	1200	0.761	1.548	0.724	0.902	4.427	0.576	**0.572**	6.460	11.244	12.098
	3000	1.073	1.997	4.487	1.081	1.994	0.823	**0.806**	7.272	11.527	11.503
	Avg	0.772	1.565	2.431	0.922	2.860	0.559	**0.553**	5.711	11.063	11.620
EPA-Air	24	1.616	1.788	1.206	1.334	0.917	0.682	**0.628**	6.784	8.385	7.968
	72	1.694	3.105	0.966	1.308	0.809	0.698	**0.667**	6.818	8.563	8.351
	168	2.555	3.532	1.074	1.238	1.764	0.775	**0.720**	7.143	8.514	8.161
	Avg	1.955	2.808	1.082	1.293	1.163	0.718	**0.672**	6.915	8.487	8.160
ClusterTrace	24	5.757	8.604	6.849	10.282	8.340	3.996	4.004	**3.964**	10.821	9.629
	48	6.078	4.845	7.416	7.269	6.504	**4.289**	4.340	4.384	10.333	10.958
	72	7.231	10.977	7.035	10.376	8.247	4.623	**4.463**	4.664	9.780	10.528
	Avg	6.355	8.142	7.100	9.309	7.697	4.303	**4.269**	4.337	10.311	10.372
APTC	50	**0.376**	0.376	0.377	0.376	0.376	0.398	0.400	0.405	0.379	0.413
	100	0.388	**0.375**	0.375	0.375	0.375	0.383	0.385	0.415	0.379	0.396
	200	**0.375**	0.375	0.378	0.375	0.375	0.381	0.385	0.420	0.378	0.387
	Avg	0.380	**0.375**	0.377	0.375	0.375	0.387	0.390	0.413	0.379	0.399
FNSPID	30	3.184	4.907	0.935	6.247	2.622	**0.319**	0.612	0.422	6.929	6.716
	180	3.348	4.666	0.989	6.962	3.064	**0.308**	15.548	0.466	8.500	8.058
	365	3.108	4.865	0.718	7.408	2.747	0.990	0.718	**0.430**	7.362	7.097
	Avg	3.213	4.813	0.881	6.872	2.811	0.539	5.626	**0.439**	7.597	7.290
GDELT	30	**1.946**	17.667	4.269	9.949	8.045	2.034	1.956	4.019	16.130	15.514
	90	2.008	14.633	4.787	9.737	7.869	2.014	**2.008**	3.800	17.639	16.466
	180	**1.728**	14.461	5.225	10.076	7.315	1.859	1.800	3.833	15.078	14.525
	Avg	**1.894**	15.587	4.760	9.921	7.743	1.969	1.921	3.884	16.282	15.502
Seabirds	60	2.503	0.179	0.025	0.034	0.020	0.012	**0.005**	0.021	0.040	0.100
	360	1.974	0.049	0.158	0.041	0.028	**0.024**	0.028	0.041	0.042	0.112
	720	3.790	**0.023**	0.026	0.050	0.066	0.045	0.024	0.052	0.052	0.137
	Avg	2.756	0.084	0.070	0.042	0.038	0.027	**0.019**	0.038	0.045	0.116

Table 8: Performance comparison of time series forecasting models and TSFMs in terms of MAE ($\times 10^{-2}$). Best results are in **bold** and second-best are underlined.

Dataset	Pred	Crossformer	DLinear	PatchTST	TimesNet	iTransformer	DuET	SRSNet	MIRA	Sundial	Timer
PhysioNet	1	9.073	20.636	20.626	15.178	19.574	**4.020**	4.291	18.336	20.270	24.464
	6	4.845	17.818	8.244	8.817	8.688	**4.577**	4.830	18.711	38.118	24.256
	12	**4.668**	18.031	9.015	8.948	8.867	5.066	5.303	18.979	37.640	24.278
	Avg	6.195	18.828	12.628	10.981	12.376	**4.554**	4.808	18.675	32.009	24.333
Human Activity	200	5.596	10.224	5.559	45.665	10.008	**2.992**	2.992	21.896	44.685	45.422
	1200	7.137	10.423	4.063	46.122	9.663	**3.778**	3.779	28.957	55.996	46.579
	3000	11.066	10.652	8.091	46.714	9.722	**4.710**	4.725	31.173	53.641	46.614
	Avg	7.933	10.433	5.904	46.167	9.798	**3.827**	3.832	27.342	51.441	46.205
USHCN	1	44.78	49.91	48.49	48.17	48.31	**25.33**	33.77	46.48	57.67	50.54
	6	46.43	51.01	49.63	49.16	49.55	**23.94**	32.28	47.40	47.04	49.60
	12	47.67	50.27	49.69	49.20	49.74	**24.11**	32.67	48.90	47.34	50.11
	Avg	46.29	50.40	49.27	48.84	49.20	**24.46**	32.91	47.59	50.68	50.08
CESNET	24	11.320	12.912	11.301	16.692	13.428	9.214	11.582	**9.124**	13.309	13.338
	168	11.210	13.177	10.954	16.755	12.480	**9.035**	11.521	9.328	13.856	13.883
	336	10.996	13.340	10.991	16.810	14.029	**9.071**	11.710	9.228	13.576	13.626
	Avg	11.175	13.143	11.082	16.752	13.312	**9.107**	11.604	9.227	13.580	13.616
Pamap2	200	5.233	8.522	10.114	5.493	10.183	**2.864**	2.893	9.696	19.898	21.927
	1200	6.222	10.515	5.646	5.907	15.657	**4.357**	4.372	14.797	20.700	22.561
	3000	7.826	11.981	14.849	6.442	9.674	**5.297**	5.365	16.244	21.166	22.526
	Avg	6.427	10.339	10.203	5.947	11.838	**4.173**	4.210	13.579	20.588	22.338
EPA-Air	24	9.240	10.289	7.846	7.749	7.114	5.615	**5.149**	14.928	20.295	19.913
	72	9.713	13.416	7.561	7.759	6.638	5.538	**5.481**	15.358	20.534	20.430
	168	12.831	14.941	7.681	7.761	9.914	6.056	**5.813**	16.146	20.471	20.139
	Avg	10.595	12.882	7.696	7.756	7.889	5.736	**5.481**	15.477	20.433	20.161
ClusterTrace	24	18.448	22.485	19.853	25.318	20.261	14.963	15.084	**14.901**	25.951	24.494
	48	20.219	15.917	20.860	20.247	19.002	**15.669**	15.867	15.815	25.559	26.012
	72	22.461	26.171	20.163	25.590	22.238	16.166	**15.880**	16.220	24.151	25.021
	Avg	20.376	21.524	20.292	23.718	20.500	**15.599**	15.610	15.645	25.220	25.176
APTC	50	4.620	4.550	4.553	**4.543**	4.543	4.709	4.728	4.674	4.566	4.775
	100	4.619	4.542	4.553	**4.534**	4.534	4.616	4.636	4.727	4.554	4.669
	200	4.598	4.537	4.543	**4.529**	4.529	4.586	4.612	4.756	4.548	4.605
	Avg	4.612	4.543	4.550	**4.535**	4.535	4.637	4.659	4.719	4.556	4.683
FNSPID	30	15.275	15.807	7.002	19.377	11.752	**3.617**	5.076	3.808	17.798	18.396
	180	15.978	15.788	7.206	20.270	12.976	**3.466**	24.486	4.570	20.692	20.765
	365	15.372	15.961	6.375	21.040	12.633	6.034	5.228	**4.485**	18.126	18.584
	Avg	15.542	15.852	6.861	20.229	12.454	4.372	11.597	**4.288**	18.872	19.248
GDELT	30	8.786	31.079	12.575	20.963	16.995	8.823	**8.706**	13.270	28.485	27.447
	90	9.354	26.966	12.850	21.124	17.410	**8.759**	8.783	12.868	30.159	28.360
	180	**8.440**	26.897	13.021	21.746	17.266	8.774	8.620	12.967	26.721	26.294
	Avg	8.860	28.314	12.815	21.278	17.224	8.785	**8.703**	13.035	28.455	27.367
Seabirds	60	13.227	3.798	1.075	1.251	0.988	0.787	**0.465**	0.854	1.398	2.140
	360	12.206	1.774	2.953	1.411	1.144	**1.061**	1.185	1.343	1.411	2.390
	720	16.814	**1.039**	1.101	1.588	1.822	1.475	1.135	1.631	1.666	2.648
	Avg	14.082	2.204	1.710	1.417	1.318	1.108	**0.928**	1.276	1.492	2.393

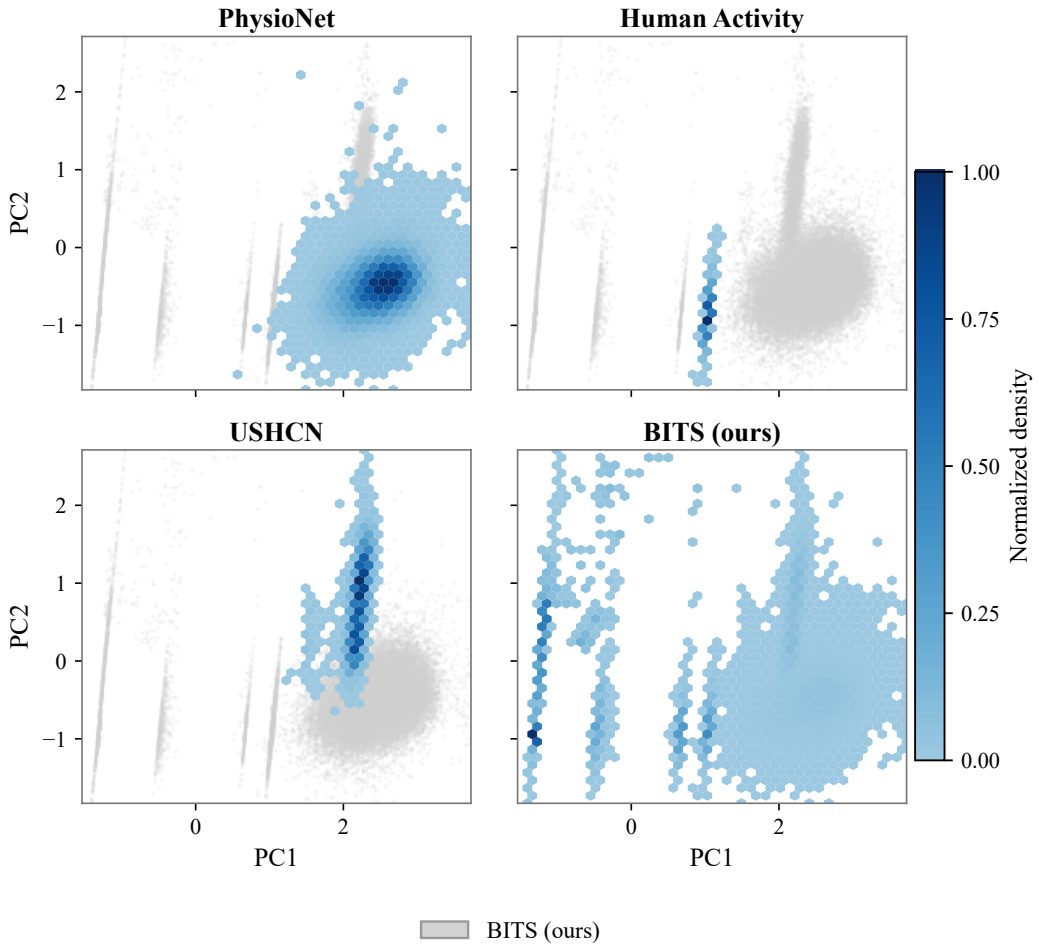


Figure 9: Coverage of irregularity characteristics across commonly used datasets and BITS. Hexbin plots visualize the distributions of four irregularity characteristics projected into a shared two-dimensional PCA space. Gray indicates the overall distribution of all 11 datasets in BITS, while blue represents the normalized density of each dataset or the entire BITS collection.

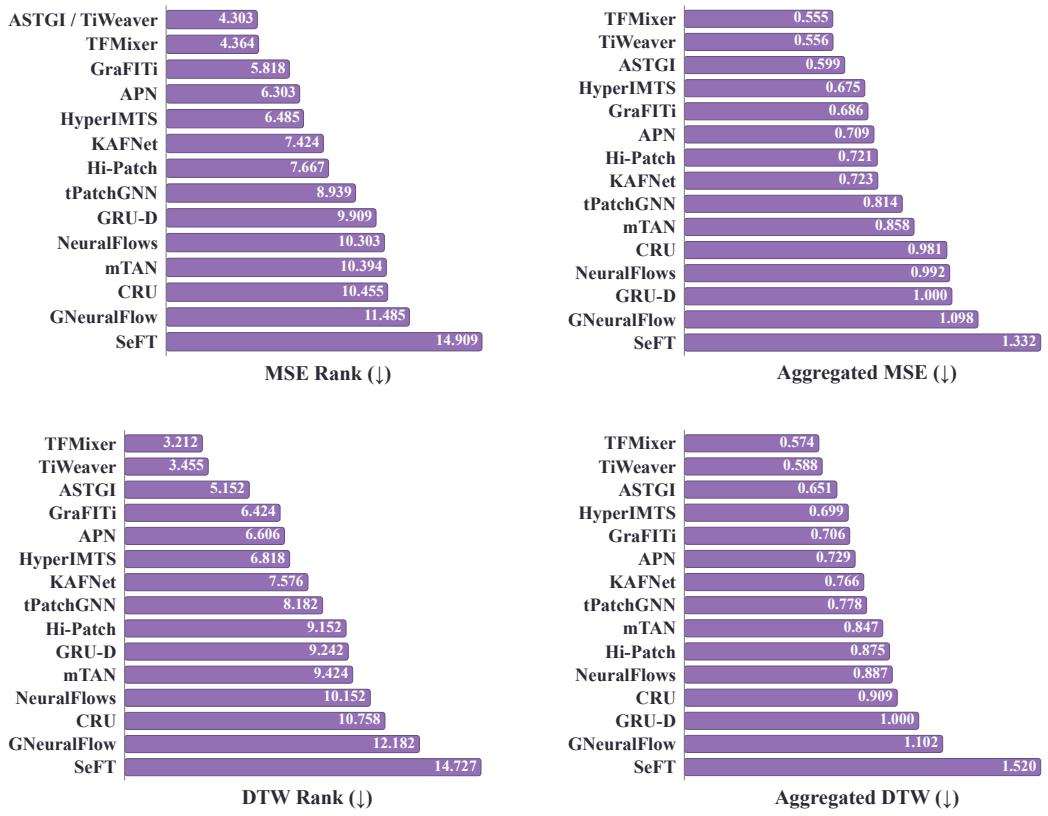


Figure 10: Overall performance ranking on BITS in terms of MSE rank, aggregated MSE, DTW rank and aggregated DTW. Lower is better.