Large Scale Financial Time Series Forecasting with Multi-faceted Model

Defu Cao*
defucao@usc.edu
University of Southern California
Los Angeles, California, USA

Yixiang Zheng*
yixiangzheng@usc.edu
University of Southern California
Los Angeles, California, USA

Parisa Hassanzadeh
parisa.hassanzadeh@jpmchase.com
J.P.Morgan AI Research
New York, USA

Simran Lamba
simran.lamba@jpmchase.com
J.P.Morgan AI Research
New York, USA

Xiaomo Liu
xiaomo.liu@jpmchase.com
J.P.Morgan AI Research
New York, USA

Yan Liu
yanliu.cs@usc.edu
University of Southern California
Los Angeles, California, USA

ABSTRACT

Data-driven approaches using deep neural networks have been successful in modeling complex financial time series and generating accurate predictions without requiring extensive domain knowledge. However, most of the existing models that assume independent and identically distributed (*i.i.d.*) data may not generalize well to novel situations or distributional shifts across or inside financial scenarios. To address this challenge, we introduce an invariant learning-based regularizer with relaxed bounds that expands the range of feasible solutions and mitigates over-convergence issues in Invariant Risk Minimization (IRM). In practice, the regularizer can be incorporated into both linear and nonlinear financial time series forecasting models. Experimental results on real-world large-scale financial datasets show that our proposed method enables more robust and adaptable financial forecasting models, enhancing the overall performance and generalizability of financial forecasting on both in-distribution and out-of-distribution (OOD) samples.

CCS CONCEPTS

• **Mathematics of computing** → **Distribution functions**; *Multivariate statistics*; • **General and reference** → Experimentation.

KEYWORDS

Financial time series; Forecasting algorithm; Distributional shifts

ACM Reference Format:
Defu Cao, Yixiang Zheng, Parisa Hassanzadeh, Simran Lamba, Xiaomo Liu, and Yan Liu. 2023. Large Scale Financial Time Series Forecasting with Multi-faceted Model. In *4th ACM International Conference on AI in Finance (ICAIF '23), November 27–29, 2023, Brooklyn, NY, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3604237.3626868

*Both authors contributed equally to this research.

ICAIF '23, November 27–29, 2023, Brooklyn, NY, USA

ACM ISBN 979-8-4007-0240-2/23/11.
https://doi.org/10.1145/3604237.3626868

1 INTRODUCTION

Global financial markets are fundamentally driven by investors' forward-looking views and expectations about future economic and business conditions [26]. As such, accurate financial forecasting is crucial for understanding market dynamics and enabling sound investment decisions [9]. However, producing reliable forecasts requires years of experience and deep domain expertise that is difficult to attain. Data-driven deep learning approach can be used to tackle financial forecasting tasks. Deep neural networks have recently achieved state-of-the-art results in modeling complex time series and generating accurate predictions [29]. They are capable of learning intricate patterns from large amounts of data without requiring extensive feature engineering or domain knowledge.

Financial forecasting, by itself, is inherently more difficult than typical time series forecasting due to a variety of factors [1] such as external variables, non-stationarity, high degree of noise, and volatility. On the one hand, the adaptive nature of market participants, who constantly update their strategies in response to new information, contributes to the non-stationary nature of financial time series [5]. On the other hand, different companies within the market also have different business strategies based on their unique circumstances, business models, and risk appetites. This discrepancy in business strategies across companies further contributes to the distributional shift of financial data and returns.

Despite the demonstrated success of long short-term memory (LSTM) [8] and Transformer-based [12] deep neural networks in time series forecasting tasks, a growing trend is emerging toward achieving comparable results utilizing simply linear layers [31] with its ability to capture the characteristics of trends and seasonality inside the observed time series and multilayer perceptron (MLP)-based model for handling covariates and non-linear dependencies [10]. Given the imperative for rapid response in financial interaction prediction, we envisage a rising preference for lightly deep neural network models that deliver robust predictive capabilities. Such models are anticipated to gain increasing traction within the field of financial time series prediction.

Furthermore, creating separate financial forecasting models for individual companies can be costly and require access to abundant, company-specific data. This is even harder for emerging companies that might not have much data available and may also produce zero-shot settings. Moreover, these models are often tailored to specific companies and their unique business tactics, making them difficult

to apply to other firms. However, most if not all of the previous existing deep learning-based financial forecasting models rely on the assumption of independent and identically distributed (*iid*.) data, which constrains the generalization capabilities when confronted with unencountered scenarios or when tested financial data exhibits distributional shifts. Recently, it has been observed that models, such as Invariant Risk Minimization (IRM) [4], leveraging invariant correlations across multiple training distributions can learn the underlying invariant causal relationships, thus maintaining stable performance in novel situations. Nevertheless, such models may be overly constrained, potentially hindering the achievement of desired results in various applications.

To tackle the aforementioned challenges, we propose a standardized algorithm to generate automated financial forecasts for multiple companies using an invariant learning-based regularization with relaxed constraints. This novel regularization approach effectively mitigates the overfitting issue observed in IRM under excessively restrictive assumptions by expanding the feasible solution space. In other words, models employing this regularization can discern consistent underlying logical relationships across diverse known financial data scenarios, consequently producing stable forecasting results for both familiar and new companies. By providing such a more flexible solution space, our proposed regularization enables more robust and adaptable linear/non-linear financial forecasting models.

Specifically, leveraging access to an extensive repository of 3rd party corporate financial records of Standard & Poor's 500 (S&P 500) companies, we use seven different industry sectors' data from their financial statements to train the unified multi-faceted model with environmentally invariant information. This unified model is poised to streamline processes ranging from equity research to algorithmic trading by enabling rapid fundamental analysis at scale across diverse companies. In experiments, we show that our method designed under such inductive bias of environmental invariance can improve previous methods that only target a single environment. Specifically, when compared to the baseline models with mixup strategy [32] or IRM, our proposed model demonstrates a significant over-performing in S&P 500 datasets in the 'Revenue' forecasting and 'EBITDA' forecasting. Our main contributions can be summarized as follows:

- Departing from single-scenario forecasting, we seek to devise a financial time series solution capable of maintaining robust forecasting efficacy with multiple sectors. As per our understanding, such financial forecasting task has not received extensive scholarly attention previously.
- We propose a uniform multi-faceted algorithm with a regularizer toward learning the invariant representation to tackle distributional shifts within or across companies. Leveraging the innate correlations across diverse recognized scenarios, our solution can yield stable forecasting results adequate for large-scale financial time series analysis.
- We conduct extensive experiments and sensitivity analyses to demonstrate the effectiveness of our proposed algorithm on both familiar sectors and zero-shot settings. Notably, our model outperforms other baselines on the 'EBITDA' forecasting task of S&P 500 datasets by an average of 27.87%.

2 RELATED WORK

2.1 Time Series Forecasting

Time series forecasting tasks can be broadly classified into long-term forecasting [19, 20, 35] and short-term forecasting [7, 16] based on the forecast horizon. Notably, the Transformer architecture [28] has emerged as a widely accepted algorithm for long-term predictions, for example, Informer [34] and Crossformer [33], while short-term predictions predominantly employ Graph Neural Networks (GNNs) [6, 14].

As for financial forecasting, Papadimitriou et al. presents a framework that can produce reliable and robust forecasts of financial metrics in the financial statements such as income statement, cash flow and balance sheet as well as points out the distinction between the commonly used univariate model setting and the multivariate model setting. Clustering Guided model [22] first uses a clustering algorithm to group companies based on their financial characteristics and select the best time series model for each cluster. Then it uses a clustering based deep learning model to train and forecast the financial metrics for each cluster. In addition, TStI [5] transforms time series data into Gramian angular fields (GAF) images and then uses an ensemble of convolutional neural networks (CNNs) trained on different resolutions of GAF images to predict the future trend of the U.S. market based on the S&P 500 index. Recently, [2] proposes to forecast the price of the stock from National Stock Exwchange (NSE) using historical data as well as sentiment analysis of public opinion and news headlines.

For distributional shifts in time series, AdaRNN [13] addresses the temporal covariate shift by proposing an adaptive RNN model. After that, DIVERSIFY [18] extends AdaRNN into an end-to-end framework and mainly focuses on the time series classification task. While most of the related methods are limited to forecasting individual companies or narrow subsets of companies, our goal is to leverage data from diverse companies across heterogeneous environments to build a robust financial forecasting model that remains robust and adaptable across various scenarios. Please refer to [27] for more comprehensive related works.

2.2 Towards Invariant Representation

Invariant causal prediction (ICP) uses invariance under different environments to infer causality [23]. Invariant Risk Minimization (IRM) [4] proposes to find invariant information between environments to enable the out-of-distribution generalization for causal estimates. CoCo [30] maximizes an objective where the only solution is the causal solution by leveraging the causal invariance across environments. Besides, invariant Causal Representation Learning (iCaRL) [17] enables out-of-distribution (OOD) generalization via nonlinear classifiers. Recently, IRM-based learning has been shown to be fragile[25], and the criticism is mainly focused on the feasible space of the IRM-based might be overly restrictive for acquiring a solution with good generalization on unseen environments [3]. However, loosening the over-strict feasible space of IRM to acquire a solution with good generalization on unseen environments should be one solution to improve IRM-based methods' validity and reliability, which motivates us to propose this work under the highly dynamic and ever-evolving nature of financial markets.

3 PRELIMINARY

3.1 Problem Fomulation

Financial forecasting can be considered as predicting key items from a company's financial statements: the income statement, balance sheet, and cash flow statement obtained from financial filings. Considering the multitude of factors and data that potentially influence a company's financial performance, it is reasonable to view financial data as a collection of multivariate time series. Consequently, the forecasting task can be modeled as the multivariate time series forecasting and be written as follows:

$$\left\{y^e_{t+1}, \ldots, y^e_{t+H}\right\} = f\left(\left\{y^e_1, \ldots, y^e_t; \mathbf{x}^e_1, \ldots, \mathbf{x}^e_t\right\} \mid \Theta\right) \tag{1}$$

where $\left\{y^e_{t+1:t+T}\right\}_{e=1}^{\mathcal{E}}$ represents the forecast targets derived from the historical time series $\left\{y^e_{1:t}\right\}_{e=1}^{\mathcal{E}}$. Meanwhile, $\left\{\mathbf{x}^e_{1:t}\right\}_1^N$ stands for a set of multivariate time-varying vectors $\left(\left\{x^e_1, \ldots, x^e_D\right\} \in \mathbf{x}^e\right)$ of dimension D, which are associated with $\left\{y^e_{1:t}\right\}_{e=1}^{\mathcal{E}}$. The uniform multi-faceted function f is a model that calculates the model parameters Θ based on historical data from time series i. This model is then used to predict the values for future steps within a horizon H, extending from time t to $t+H$. The global model parameters, represented as Θ, can be learned conjointly with y and x, encompassing the entire $\mathcal{E}$ financial sectors (also note as environments).

3.2 Invariant Risk Minimization

Recent work develops methods for learning an invariant representation, such as IRM [4], connecting the regularizer and model invariance. It minimizes the empirical risk while constraining the representation learning model $f : \mathcal{X} \to \mathbb{R}^d$ to find an embedding space where our predictor $g_w : \mathbb{R}^d \to \mathcal{Y}$ has parameter w that is simultaneously optimal for all environments:

$$\begin{aligned} &\min_{w,f} \sum_{e\in\mathcal{E}} R^e(w, f), \\ &s.t.\ w \in \arg\min_{\bar{w}}\ R^e(w, f),\ \text{for all } e \in \mathcal{E} \end{aligned} \tag{2}$$

where $R^e(w, f) = \mathbb{E}_{x,y\sim p^e(x,y)}[l_y(g_w(f(x)), y)]$ refers to the risk of f, w in environment e. $l_y : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is the loss function computing the difference between the predicted outcome and actual outcome. In this paper, without loss of generality, we apply square error $(\hat{y} - y)^2$ as the loss function and follow the notation method of IRM in this section unless specifically mentioned.

4 METHODOLOGY

4.1 Overview

As shown in Algorithm 1, we first collect data from heterogeneous sectors and split them into the training set and test set. Then, we initialize the forecasting model M with parameters $\alpha \sim \mathcal{N}(0, 1)$. Note that the selection of the backbone for our predictive model M is flexible and not confined to linear or nonlinear structures. For illustrative purposes, we will be using a Multilayer Perceptron (MLP) financial model [21] as an example in the following explanation. Then, we train the MLP model with forecasting objective function as well as the optimization-based regularizer.

Enabling a representation learning model to adapt invariant information in different environments is a challenging task since most of the machine learning models assume that the testing samples are drawn from the same distribution as training data. Recent work develops methods for learning an invariant representation, such as mixup [32], IRM [4] and CoCo [30]. However, IRM-based approaches are unstable with the strict approximation of invariant features [25]. To this end, we propose to learn invariant representations in a high-dimensional nonlinear feature space through neural networks that approximate the underlying relationships in a looser manner.

Algorithm 1 Uniform Multi-faceted Forecasting Model for Large Scale Financial Time Series

Input: Observational time series X^e, where $e \in \{1, ..., n\}$ is the $\mathcal{E}$ (Environments) index, label Y^e.
Outcome: Predicted time series, $\hat{Y}^e$.
Initialize model M with parameters $\alpha \leftarrow \mathcal{N}(0, 1)$.
Split X^e into training X^e_{train} and test X^e_{test} data
while less than training epoch **do**
  for each environment e **do**
    Extract data X^e from environment e
    Split X_e into $X_{e,train}$ and $X_{e,test}$
    Compile M
    Fit M on X^e_{train} using mini-batches
    Calculate y^e via model M
    $\mathrm{L}^{\mathrm{e}}_{\mathrm{MSE}} = l_y(\hat{y}^e, y^e)$//Calculate outcome loss
    $\mathrm{L}^{\mathrm{e}}_{\mathrm{OB}} = \|(\nabla R^e(w, f))^{\circ c} \circ w\|_2^2$ //Apply optimization-based regularizer
    Update α according to:
    $\alpha \leftarrow \alpha - \lambda\nabla_\alpha(\mathrm{L}^{\mathrm{e}}_{\mathrm{MSE}} + \gamma\mathrm{L}^{\mathrm{e}}_{\mathrm{OB}})$
Output: Learnt parameters α for forecasting model M; Estimated outcome $\hat{Y}^e$.

4.2 Optimization-based Invariant Regularizer

The feasible space for the optimization presented in Section 3.2 may impose overly stringent conditions for identifying a solution that effectively generalizes to unseen environments. Our paper seeks to investigate whether superior solutions could potentially exist outside this feasible space under certain specific scenarios.

Let us denote the variance of the outcome y, given the embedding $f(x)$, in the test environment e, as $\sigma^2_e(y|f(x))$. Suppose that f^* represents the optimal solution of the Invariant Risk Minimization (IRM). The distribution of x in the test environment is represented as $p_e(x)$. We posit that there might exist a superior solution outside the feasible space of IRM, if there exists a data representation function f' such that, in the test environment e, $g_w(f(x))$ exhibits L-Lipschitz continuity with respect to w on the support set of x. And we have:

$$\mathbb{E}_{x\sim p_e(x)}[\sigma^2_e(y|f^*(x)) - \sigma^2_e(y|f'(x))] > L||w' - w''||_2^2 \tag{3}$$

where w' is the optimal predictor parameter trained with f' as encoder on the training environments and w'' is the optimal predictor parameter trained with f' as encoder on the testing environment.

The previously discussed equation indicates that if the optimal solution for the predictor across diverse environments remains bounded, a superior representation of the learning model could exist outside the IRM's feasible space. This insight suggests an opportunity to judiciously *relax the parameter constraints*, consequently improving performance. This enhancement can be achieved by compelling the model to formulate an embedding space, in which a predictor g_w can be identified that stays near to an optimal solution within each environment.

Nonetheless, the optimization issue outlined above presents computational challenges due to the necessity of calculating the w_e^* for every environment. By strengthening condition 1 on g_w a notch, restricting the $R^e(w, f)$ to possess a Lipschitz continuous gradient, we can make this more manageable. As such, the magnitude of the gradient of w, denoted as $\nabla_w R^e(w, f)$, can serve as an indicator of the distance separating w and the nearest w_e^*. Then, we can optimize the following tractable formulation:

$$\min_{\alpha} \sum_{e \in \mathcal{E}} \Big[\underbrace{R^e(w, f)}_{\text{Empirical risk}} + \lambda \underbrace{\max(||\nabla_w R^e(w, f)||_2^2, \epsilon)}_{\text{Invariant regularization}} \Big], \tag{4}$$

where ϵ is a hyperparameter to control how loose the regularization is. In practice, to avoid numerical issues and simplify the optimization landscape, we approximate the $\max(||\nabla_w R^e(w, f)||_2^2, \epsilon)$ term with a smoothed surrogate. Specifically, we replace it with the squared Hadamard power[1] function $||\nabla_w R^e(w, f))^{\circ c}||_2^2$, where $(\nabla_w R^e(w, f))^{\circ c}$ represents taking the element-wise power of the gradient vector $\nabla_w R^e(w, f)$.

The element-wise Hadamard power replacement for the $\max(\cdot, \epsilon)$ function approximates its key properties. When $\|(\nabla_w R^e(w, f))^{\circ c}\|$ is small, its value approaches zero rapidly, mimicking the behavior of the maximum function for small inputs. Additionally, for regions where w is distant from optimality along some dimensions, $(\nabla_w R^e(w, f))^{\circ c}$ provides stronger regularization compared to a simple threshold. This acts to restrain large gradient updates that could hinder convergence. Critically, the surrogate remains fully differentiable, maintaining the benefits of gradient-based optimization. Together, these properties allow the Hadamard power formulation to uphold the aims of limiting uninformative updates, while introducing the favorable characteristics of smoothness into the objective. The result is an optimization procedure that can traverse complex landscapes efficiently to obtain high-quality solutions.

Therefore, the optimization-based invariant regularizer is: $\mathrm{L}_{\mathrm{OB}}^{\mathrm{e}} = \|(\nabla R^e(w, f))^{\circ c}\|_2^2$, where the power c applied element-wise to the gradient is a tunable hyperparameter, which is set greater than one. This strengthens regularization for large gradient magnitudes compared to $c = 1$ (vanilla IRM's setting), guiding optimization. Inspired by prior work on causal discovery from observational data [30], we additionally multiply the gradient by the predictor weights w. This attenuates the influence of non-causal features with $w \approx 0$, focusing the objective on truly predictive dimensions. Together, these modifications yield our final regularize:

$$\mathrm{L}_{\mathrm{OB}}^{\mathrm{e}} = \left\|(w \circ \nabla R^e(w, f))^{\circ c}\right\|_2^2, \tag{5}$$

where $\circ$ denotes the Hadamard product. Importantly, our formulation presents a flexible and principled framework for learning underlying invariant structures through gradient-based optimization objectives.

[1]Hadamard power is defined as, for $y = x^{\circ c}$, we have $y_i = x_i^c$

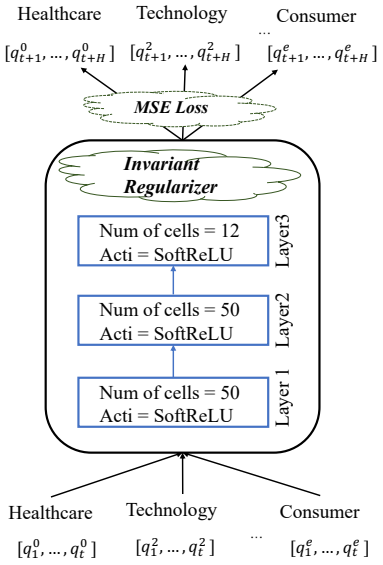

Figure 1: The proposed uniform architecture for multi-step forecasting with optimization-based Invariant regularizer and data-driven backbone, which is three-layer MLP as an example .

4.3 Objective Function

As shown in Figure 1, our proposed regularizer can be seamlessly integrated with any existing machine learning model to enhance its stability across diverse environments. We Utilize the Mean Squared Error (MSE) as the loss function, i.e., $\mathrm{L}_{\mathrm{MSE}}^{\mathrm{e}} = \frac{1}{n^e} \sum_{i=1}^{n^e} (y_i - \tilde{y}_i)^2$, for prediction loss, where the $\tilde{y}_i$ is the predicted value from backbone forecasting model (MLP). In addition, we incorporate our novel regularizer to derive a model that exhibits consistent performance across different sectors. Thus, the loss for the sector e from the heterogeneous environments can be written as:

$$\mathrm{L}^{\mathrm{e}} = \mathrm{L}_{\mathrm{MSE}}^{\mathrm{e}} + \gamma \mathrm{L}_{\mathrm{OB}}^{\mathrm{e}}, \tag{6}$$

where γ is the hyperparameter for controlling the importance of the proposed regularizer.

5 EXPERIMENTS

5.1 Datasets

The assessment and prediction of a company's future profitability and input-output ratio are essential for the development and investment of the company. Financial analysis and forecasting are data-driven and mostly depend on the combination of different types of data which include company filings, industry reports, and so on. In this project, we mainly use the financial statements of a company: balanced sheet, income statements, and cash flow statements.

Standard & Poor's 500 Index (S&P 500) is a market-capitalization-weighted index of the 500 largest U.S. publicly traded companies [11].

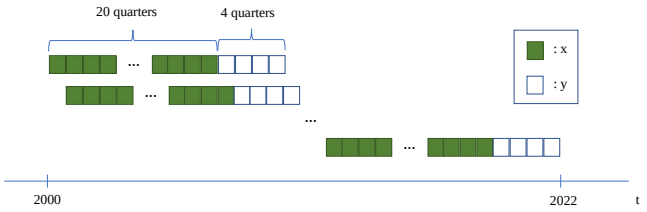

Figure 2: An illustration of the data allocation for future 4 quarter forecasting.

We evaluate our model in S&P 500 with companies' quarterly data for 10 sectors: Basic Materials (21 companies), Communication Services (26 companies), Consumer Cyclical (58 companies), Consumer Defensive (36 companies), Energy (22 companies), Financial Services (69 companies), Healthcare (65 companies), Industrials (73 companies), Technology (71 companies), Utilities (30 companies) from the years 2000 to 2022. Sliding Window (Section 5.2) is used to build the time series samples, where we can get 19,513 samples from 7 sectors as seen data for training and 11,188 samples from 3 sectors as unseen data (referring to zero-shot setting) for generation ability evaluation. As a financial project, our target is to predict the future performance of companies with multiple quarters. Thus we forecast future one-year metrics using the past 5 years' data.

5.2 Experimental Setting

We conduct our experiments using four NVIDIA GeForce GTX 2080 GPUs with 16G memory. For each sector in the training seen stage, we split the data into training, test, and validation sets with an 80/10/10 ratio. The hyperparameter γ was determined through a systematic grid search, with the search space encompassing the values 1, 0.1, 0.01, and 0.001. To ensure an unbiased comparison, the reported results represent the arithmetic mean of three independent runs. We set batch size to 32 for training.

Out-of-distribution (OOD) Setting. Based on all the sectors, we separate 10 sectors into 7 seen data (Healthcare, Technology, Basic Materials, Energy, Financial Services, Utilities, Communication Services) for the training stage as well as the evaluation stage and 3 sectors (Consumer Defensive, Consumer Cyclical, Industrials) as the unseen environments to build the OOD setting for evaluation stage, also referring to the zero-shot setting in our results.

Sliding Window Setting. For each company, we do the window to partition the dataset into subsections in order to increase the dimension shape of the time series dataset. We set window slide t=24 which means 24 quarters as one sample (the first 20 quarters/5 years data as input and the last 4 quarters/1 year data for evaluation). The window framework is demonstrated in Figure 2.

Features and Targets. Following [21]'s multivariate time series setting, we choose cost of goods sold (COGS), selling, general and administrative expenses (SG&A), RD expenses (RD_EXP) which are more significant and basic financial metrics among all financial data as fixed input features from the income statement. For the sake of simplicity, we make predictions for revenue (REV) and Earnings Before Interest, Taxes, Depreciation, and Amortization (EBITDA) which can also be calculated as EBITDA = REV - COGS - SG&A - RD_EXP, respectively.

Missing value. The S&P 500 dataset contains some examples of missing values. For instance, many companies may lose information before 2010. For this kind of missing data, we experiment with different interpolation methods and find that linear interpolation is the best for our time series data. Linear interpolation is an approach used to estimate a value within a range based on two known endpoint values. Another major missing data is research development expenses that many companies not have invested in which will leave null in statements. Thus we replace all the null in research development expenses with the value 0.

For each method, we predict quarterly Revenue/EBITDA by the past 20 quarters' data and use prediction value to predict the next quarter's Revenue/EBITDA for 4 times to get yearly prediction result and the evaluation result(SMAPE) is based on the yearly value (sum of 4 quarters).

5.3 Evaluation Metrics

In reality, the magnitude of financial metrics values for different companies may be largely different. We choose the symmetric mean absolute percentage error (SMAPE) which is an accurate measurement based on the percentage as our evaluation metrics:

$$\text{SMAPE} = \frac{200\%}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{F_t + A_t}, \tag{7}$$

where F_t is the true value and A_t is the prediction value in our system and n is the total time steps we need to forecast.

SMAPE is sensitive to outliers, in particular, when true data and prediction are opposite in sign, the error may be up to 200% which will seriously skew the final result. Following [21], we remove the data points of 80% and 90% and verify they have a significant financial change due to mergers & acquisitions (M&A) etc.

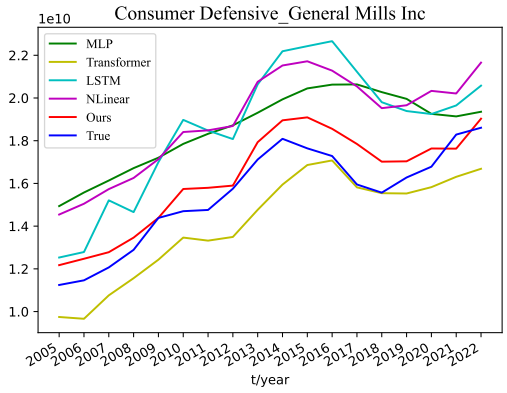

Figure 3: Revenue prediction results (SMAPE) for General Mills Inc in Consumer Defensive sector (out-of-sample).

5.4 Results

Baselines. For the purposes of establishing a robust comparison, we construct the Linear model [31] with its variants, including

Table 1: SMPAE of Revenue setting with baselines and our proposed method. Hereafter for the tables, we remove outliers with SMAPE over 0.8/0.9. The best results are marked in bold and second optimal in underlined respectively with 0.8 & 0.9.

REV	Sectors	LSTM	MLP	Transformer	AdaRNN	Linear	NLinear	DLinear	Ours
SEEN	Healthcare	54.6/59.4	44.1/48.8	16.5/18.6	31.5/35.4	19.9/21.1	34.4/37.5	30.2/32	**15.4/15.6**
	Technology	32.4/32.6	47.5/50.4	16.1/16.1	32/32.8	21.6/21.6	35.7/36.1	30.6/30.6	**13.4/13.4**
	Basic Material	15.1/15.1	27.2/27.2	19.1/19.1	14.4/14.4	10.8/10.8	13.8/13.8	12.2/12.2	**10.2/10.2**
	Energy	32.8/33.3	52.2/54.9	25.3/25.3	42/42	24.4/24.4	38.1/38.1	31.5/31.5	**23.3/23.3**
	Financial Services	27.7/28.5	40.5/42	10.6/10.6	19.4/19.4	12/12	24/24.7	19.8/20.3	**10.1/10.1**
	Utilities	14/14	17.1/17.1	8.1/8.1	13.1/13.1	8.7/8.7	10.7/10.7	**7.8/7.8**	8.8/8.8
	Communication Services	24/24	39.4/39.4	15/15	26.9/26.9	12.8/12.8	23.2/23.2	19.3/19.3	**11.7/11.7**
Zero-shot	Consumer Defensive	26.4/26.7	38.1/38.5	12.1/12.1	32.6/32.9	15.3/15.3	27.6/27.9	22.4/22.6	**9.6/9.6**
	Consumer Cyclical	26.8/27	29.8/30.1	13.6/13.6	22.3/22.5	13.9/14	23.7/23.8	18.9/19	**11.9/12**
	Industrials	18.3/18.3	30.2/30.2	12.9/12.9	14.7/15.2	**8.8/8.8**	17/17	13.5/13.6	**8.8**/8.9

Table 2: SMPAE of EBITDA setting with baselines and our proposed method.

EBITDA	Sectors	LSTM	MLP	Transformer	AdaRNN	Linear	NLinear	DLinear	Ours
Seen	Healthcare	60.4/74.9	44.1/48.8	23/23.4	32.2/33	37.4/39.3	29.8/31	43.4/45.7	**20.5/21.4**
	Technology	44.3/47.2	47.5/50.4	32.7/34	40.8/41	45.8/49.7	38.4/40.3	52.7/57	**28.6/29.1**
	Basic Material	42.5/45.6	**27.2/27.2**	35.8/38.5	35.2/37.4	27.9/29.6	27.4/29.2	29.7/31.4	28.1/29.8
	Energy	45.9/47.3	52.2/54.9	36.2/36.8	**30/30.6**	34.8/36.4	37.1/39.6	43/42.4	32/34.5
	Financial Services	41.6/43.5	40.5/42	29.6/30.5	30/31.3	33.3/34	28.6/29.2	38.1/39.1	**22.6/23.5**
	Utilities	30/30.4	**17.1/17.1**	23.8/24.5	29.4/29.7	22.7/22.7	21.3/21.1	22/22.4	21.1/21.5
	Communication Services	27.7/29.6	39.4/39.4	39.2/43	39.1/41	27.9/28.4	31.8/31.8	30.6/31.2	**27.0/27.6**
Zero-shot	Consumer Defensive	30.1/30.3	38.1/38.5	26.3/26.6	25.7/25.9	18.5/18.6	19/19	19.2/19.2	**17.8/17.8**
	Consumer Cyclical	74.3/76.9	29.8/30.1	68.4/72.3	39/40.3	39/39.5	25.8/26.4	46.3/47.2	**21.9/22.5**
	Industrials	42.7/43.3	30.2/30.2	41.4/43.5	26.4/26.7	22.7/23	20.2/20.7	27.1/27.3	**18.3/18.7**

Table 3: Sensitivity analysis: SMPAE of Revenue with proposed our regularizer's different power.

REV	Sectors	ERM	IRM	Ours(Penalty **2)	Ours(Penalty **3)	Ours(Penalty **4)
SEEN	Healthcare	44.1/48.8	58.6/62.3	52/56.4	56.3/60.1	**23.4/25**
	Technology	47.5/50.4	58/59.7	45.7/48.5	56.2/57.9	**29.9/29.9**
	Basic Material	27.2/27.2	30.5/30.5	27.6/27.6	30/30	**20.4/20.4**
	Energy	52.2/54.9	58.7/64.7	53.8/57.4	56.2/62.4	**35.9/35.9**
	Financial Services	40.5/42	46.7/49.2	42.4/44.1	45.3/47.5	**28/29**
	Utilities	17.1/17.1	17.8/17.8	17.9/17.9	17.2/17.2	**13.7/13.7**
	Communication Services	39.4/39.4	46/47.6	42.7/43.6	44/45.2	**17.6/17.6**
Zero-shot	Consumer Defensive	38.1/38.5	50.5/51.3	42.3/42.9	46.1/46.8	**17.8/17.8**
	Consumer Cyclical	29.8/30.1	38.5/39.4	34/34.8	36/36.6	**18/18.3**
	Industrials	30.2/30.2	37.7/37.9	33/33.2	35.9/36.1	**13.3/13.4**

NLinear with a simple normalization for the input sequence and DLinear, which is a combination of a Decomposition scheme used in FEDformer [35]with linear layers as well as non-linear models including LSTM [15], multi-layer perceptron (MLP) [21], Transformer [34] as our ERM model in a mixup way [24, 32] for both Revenue (REV) and EBITDA. In addition, we compare our method with AdaRNN [13] for distributional shifts. All the baselines have been widely recognized and adopted in the field for their ability to handle time series data, thereby serving as appropriate reference points for our study. According to Figure 1, we add the proposed invariant regularization with γ=0.01 for EBITDA and γ=1 for Revenue into the NLinear model to build our proposed system.

Examining every sector's results are shown in Table 1 for the Rev setting and Table 2 for the EBITDA setting. As shown in those tables, our proposed system, as a unified multi-faceted model, establishes a new benchmark for both SEEN and zero-shot scenarios, reinforcing its capability to tackle diverse sector-based scenarios. Importantly, our findings rest on the Linear-based model in which the parameters of this system (12, 301) represent only 1/856 of those of the transformer (10, 531, 077). In addition, under the EBITDA setting, prevalent models such as LSTM, Transformer, and DLinear exhibit significant difficulties when dealing with OOD data from the Consumer Cyclical sector, leading to substantial prediction errors of 74.3%, 68.4%, and 46.3%, respectively. In stark contrast, our

Table 4: Sensitivity analysis: SMPAE of EBITDA with proposed our regularizer's different power.

EBITDA	Sectors	ERM	IRM	Ours(Penalty **2)	Ours(Penalty **3)	Ours(Penalty **4)
SEEN	Healthcare	39.4/39.2	40.6/42.3	**37**/**36.8**	37.6/37.8	37.4/37.2
	Technology	36.4/37.1	44/45.3	44.8/48.2	**35.1**/**36.2**	39.4/41.2
	Basic Material	31.9/33	33.8/34.9	**30**/**31.7**	32/34.2	33.2/34.8
	Energy	39.7/41.5	44.5/46.9	**38.7**/**40.1**	39.1/40.2	43.9/44.5
	Financial Services	33.3/33.7	42.7/44.3	36.3/37.8	**29.3**/**29.3**	33.3/33.5
	Utilities	29.7/29.1	26.8/27.2	26.4/**26.4**	**26.2**/26.5	28.3/28.7
	Communication Services	**32.3**/**32.8**	37.8/39.3	36.1/36.6	37.4/38.4	38.6/40.1
Zero-shot	Consumer Defensive	29.7/29.8	27.5/27.7	27.1/27.2	31.1/31.2	**26.5**/**26.7**
	Consumer Cyclical	**39.2**/**41.5**	57/58.9	48.4/51.5	46.1/48.6	45/48
	Industrials	42.8/44.1	45.7/47.3	**37.8**/**38.8**	39.7/40.6	43.9/45.9

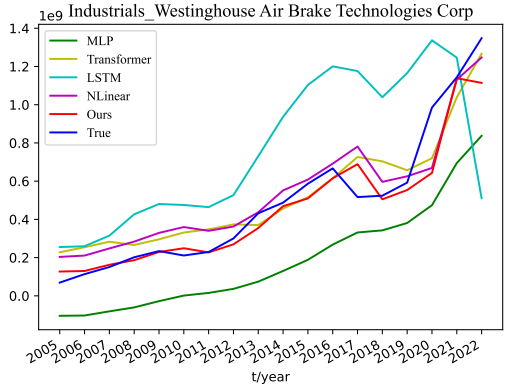

Figure 4: EBITDA prediction results (SMAPE) for Westinghouse Air Brake Technologies Corp in Industrials sector (out-of-sample).

model leverages invariant representation to effectively manage the Consumer Cyclical sector data, thereby limiting the prediction error to a mere 21.9%. On a broader scale, our proposed model outperforms the next best model by a substantial margin, reducing the prediction error by 16.10% on the Rev setting and 27.87% on the EBITDA setting.

Furthermore, we deliver a case study of the revenue predictions for General Mills Inc from the Consumer Defensive sector, which is an unseen sector. We use the past 20 quarters (5 years) data to predict the next 4 quarters (1 year) revenue iteratively to get prediction results. Figure 3 shows that our model works much better than other methods. Besides, the EBITDA yearly prediction results of Westinghouse Air Brake Technologies Corp in Figure 4 show that the forecast of our model is more accurate than all other methods in this case.

In summary, through experiments on real-world financial datasets, we show that our approach generates models with stronger generalization ability. Specifically, the model learns latent invariant relationships that hold across different market conditions by jointly training on data from multiple economic scenarios. However, unlike traditional invariant models, it does so without overly constraining the solution space, enabling it to capture more complex relationships. As a result, the learned model produces stable and accurate forecasts on both in-sample data from known scenarios as well as out-of-sample data from new scenarios.

6 ANALYSIS

6.1 Towards Invariant Representation Learning

To empirically demonstrate the superior flexibility and effectiveness of our proposed regularizer over Mixup [32] and IRM [4] we apply the three regularizers to the NLinear and Transformer backbones, respectively. As depicted in Figure 5, models with mixup tend to be relatively less self-assured with over-smoothed labels, which leads to high entropy for both in-distribution and out-of-distribution samples. Besides, IRM's overly strict feasible space can potentially lead to unstable results, specifically, an increase rather than a decrease in prediction error in unseen scenarios. Contrastingly, our proposed regularizer consistently delivers performance improvements across all unseen scenarios. This experimental evidence underscores the robustness and adaptive capabilities of our proposed regularizer, highlighting its potential for broader application.

6.2 Sensitivity Analysis on L_{OB}

We investigate the effect of the power value c in Equation 5 of on the proposed model (MLP as the backbone) to verify our claim previously. We vary the power value c in the range $\{2.0, 3.0, 4.0\}$ and the IRM. The Rev results are shown in Table 3, where the performance of the IRM is worse than the ERM and the performance of penalty with power 4 is the best. Specifically, over-constraining a model with environmental invariance assumptions can degrade overall performance when comparing ERM with IRM and our regularizer with power 2 and 3. On the other hand, judiciously relaxing such constraints within a model can lead to improved accuracy over traditional ERM approaches. Furthermore, for previously OOD scenarios, imposing properly constrained environmental invariance encourages the model to learn representations that are robust to variations in the input distribution. This enables stable predictions across a range of test-time environments not observed during training. For EBITDA, according to the result shown in Table 4, the performance of IRM is worse than the baseline which is the same as revenue. The results reveal that the best performance of all sectors is mainly concentrated on the penalty with power 2 and 3.

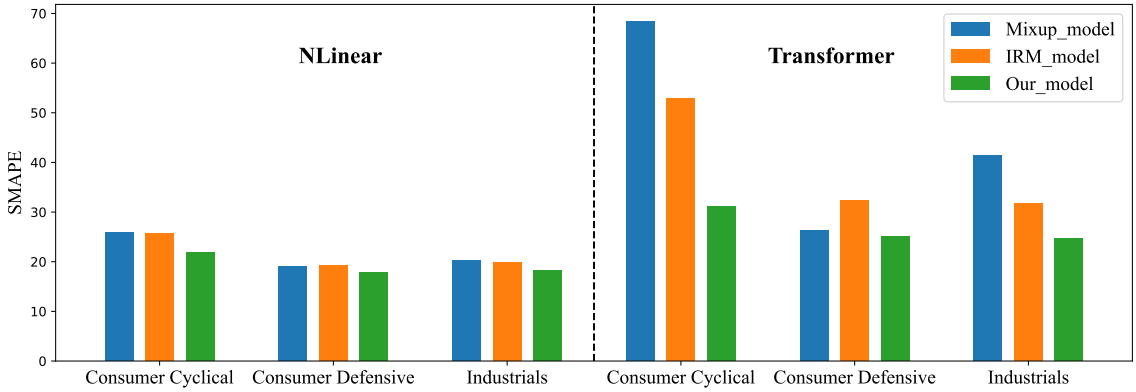

Figure 5: The comparison for Mixup, IRM, our regularizer on NLinear and Tranformer under EBITDA setting.

7 CONCLUSION

We propose an invariant learning-based regularizer with relaxed bounds, for large-scale financial forecasting tasks. This regularizer alleviates over-convergence by relaxing the constraints on feasible solutions, allowing the model to learn a wider range of invariant features. This regularizer can be incorporated into both linear and non-linear architecture for financial time series forecasting. In contrast, models without the proposed regularizer fail to generalize as well, with their performance suffering significantly when applied to new scenarios. Similarly, standard MLP models overfit to in-sample data and do not generalize to out-of-sample data even from known scenarios. Our model thus achieves the desirable property of learning generalizable, invariant relationships from the data without the typical loss in performance on in-sample data faced by most invariant learning methods. The clear ongoing future work is to inject the regularizer with more powerful forecasting models, for example, Transformers-based models combing with large language models, to achieve more accuracy and more stable results on financial datasets.

ACKNOWLEDGMENTS

This work was funded in part by J.P. Morgan AI Research. This paper was prepared for informational purposes in part by the Artificial Intelligence Research group of JPMorgan Chase & Co. and its affiliates ("JP Morgan"), and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

REFERENCES

[1] Yaser S Abu-Mostafa and Amir F Atiya. 1996. Introduction to financial forecasting. *Applied intelligence* 6 (1996), 205–213.

[2] Aditya Agarwal, Romit Ganjoo, Harsh Panchal, and Suchitra Khojewe. 2023. Deep Learning–Based Financial Forecasting of NSE Using Sentiment Analysis. In *Recurrent Neural Networks*. CRC Press, 263–288.

[3] Kartik Ahuja, Jun Wang, Amit Dhurandhar, Karthikeyan Shanmugam, and Kush R. Varshney. 2021. Empirical or Invariant Risk Minimization? A Sample Complexity Perspective. In *International Conference on Learning Representations*. https://openreview.net/forum?id=jrA5GAccy_

[4] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).

[5] Silvio Barra, Salvatore Mario Carta, Andrea Corriga, Alessandro Sebastian Podda, and Diego Reforgiato Recupero. 2020. Deep learning and time series-to-image encoding for financial forecasting. *IEEE/CAA Journal of Automatica Sinica* 7, 3 (2020), 683–692.

[6] Defu Cao, Jiachen Li, Hengbo Ma, and Masayoshi Tomizuka. 2021. Spectral temporal graph neural network for trajectory prediction. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1839–1845.

[7] Defu Cao, Yujing Wang, Juanyong Duan, Ce Zhang, Xia Zhu, Congrui Huang, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, et al. 2020. Spectral temporal graph neural network for multivariate time-series forecasting. *Advances in neural information processing systems* 33 (2020), 17766–17778.

[8] Jian Cao, Zhi Li, and Jian Li. 2019. Financial time series forecasting model based on CEEMDAN and LSTM. *Physica A: Statistical mechanics and its applications* 519 (2019), 127–139.

[9] Dawei Cheng, Fangzhou Yang, Sheng Xiang, and Jin Liu. 2022. Financial time series forecasting with multi-modality graph neural network. *Pattern Recognition* 121 (2022), 108218.

[10] Abhimanyu Das, Weihao Kong, Andrew Leach, Rajat Sen, and Rose Yu. 2023. Long-term Forecasting with TiDE: Time-series Dense Encoder. *arXiv preprint arXiv:2304.08424* (2023).

[11] Diane K Denis, John J McConnell, Alexei V Ovtchinnikov, and Yun Yu. 2003. S&P 500 index additions and earnings expectations. *the Journal of Finance* 58, 5 (2003), 1821–1840.

[12] Qianggang Ding, Sifan Wu, Hao Sun, Jiadong Guo, and Jian Guo. 2020. Hierarchical Multi-Scale Gaussian Transformer for Stock Movement Prediction.. In *IJCAI*. 4640–4646.

[13] Yuntao Du, Jindong Wang, Wenjie Feng, Sinno Pan, Tao Qin, Renjun Xu, and Chongjun Wang. 2021. Adarnn: Adaptive learning and forecasting of time series. In *Proceedings of the 30th ACM international conference on information & knowledge management*. 402–411.

[14] Weiwei Jiang and Jiayun Luo. 2022. Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications* (2022), 117921.

[15] Xuan-Hien Le, Hung Viet Ho, Giha Lee, and Sungho Jung. 2019. Application of long short-term memory (LSTM) neural network for flood forecasting. *Water* 11, 7 (2019), 1387.

[16] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2018. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In *International Conference on Learning Representations*.

[17] Chaochao Lu, Yuhuai Wu, Jośe Miguel Hernández-Lobato, and Bernhard Schölkopf. 2021. Nonlinear invariant risk minimization: A causal approach. *arXiv preprint arXiv:2102.12353* (2021).

[18] Wang Lu, Jindong Wang, Xinwei Sun, Yiqiang Chen, and Xing Xie. 2022. Generalized representations learning for time series classification. *arXiv preprint arXiv:2209.07027* (2022).

[19] Hao Niu, Guillaume Habault, Roberto Legaspi, Chuizheng Meng, Defu Cao, Shinya Wada, Chihiro Ono, and Yan Liu. 2023. Time-delayed Multivariate Time Series Predictions. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*. SIAM, 325–333.

[20] Hao Niu, Chuizheng Meng, Defu Cao, Guillaume Habault, Roberto Legaspi, Shinya Wada, Chihiro Ono, and Yan Liu. 2022. Mu2rest: Multi-resolution recursive spatio-temporal transformer for long-term prediction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 68–80.

[21] Antony Papadimitriou, Urjitkumar Patel, Lisa Kim, Grace Bang, Azadeh Nematzadeh, and Xiaomo Liu. 2020. A multi-faceted approach to large scale financial forecasting. In *Proceedings of the First ACM International Conference on AI in Finance*. 1–8.

[22] Urjitkumar Patel, Lisa Kim, and Antony Papadimitriou. 2021. Financial Forecasting with Clustering Guided Modeling. In *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 3267–3273.

[23] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78, 5 (2016), 947–1012.

[24] Francesco Pinto, Harry Yang, Ser Nam Lim, Philip Torr, and Puneet Dokania. 2022. Using mixup as a regularizer can surprisingly improve accuracy & out-of-distribution robustness. *Advances in Neural Information Processing Systems* 35 (2022), 14608–14622.

[25] Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. 2021. The Risks of Invariant Risk Minimization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=BbNIbVPJ-42

[26] Georgios Sermpinis, Andreas Karathanasopoulos, Rafael Rosillo, and David de la Fuente. 2021. Neural networks in financial trading. *Annals of Operations Research* 297 (2021), 293–308.

[27] Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. 2020. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied soft computing* 90 (2020), 106181.

[28] Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. 2022. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125* (2022).

[29] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *International Conference on Learning Representations*.

[30] Mingzhang Yin, Yixin Wang, and David M Blei. 2021. Optimization-based Causal Estimation from Heterogenous Environments. *arXiv preprint arXiv:2109.11990* (2021).

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

[32] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r1Ddp1-Rb

[33] Yunhao Zhang and Junchi Yan. 2022. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*.

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

[35] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. 2022. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*. PMLR, 27268–27286.