Time-Series Forecasting for Out-of-Distribution Generalization Using Invariant Learning

Haoxin Liu [1] **Harshavardhan Kamarthi** [1] **Lingkai Kong** [1] **Zhiyuan Zhao** [1] **Chao Zhang** [1] **B. Aditya Prakash** [1]

Abstract

Time-series forecasting (TSF) finds broad applications in real-world scenarios. Due to the dynamic nature of time-series data, it is crucial to equip TSF models with out-of-distribution (OOD) generalization abilities, as historical training data and future test data can have different distributions. In this paper, we aim to alleviate the inherent OOD problem in TSF via invariant learning. We identify fundamental challenges of invariant learning for TSF. First, the target variables in TSF may not be sufficiently determined by the input due to unobserved core variables in TSF, breaking the conventional assumption of invariant learning. Second, time-series datasets lack adequate environment labels, while existing environmental inference methods are not suitable for TSF.

To address these challenges, we propose FOIL, a model-agnostic framework that enables time-series **F**orecasting for **O**ut-of-distribution generalization via **I**nvariant **L**earning. FOIL employs a novel surrogate loss to mitigate the impact of unobserved variables. Further, FOIL implements a joint optimization by alternately inferring environments effectively with a multi-head network while preserving the temporal adjacency structure, and learning invariant representations across inferred environments for OOD generalized TSF. We demonstrate that the proposed FOIL significantly improves the performance of various TSF models, achieving gains of up to 85%.

[1]School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, USA. Correspondence to: Haoxin Liu <hliu763@gatech.edu>, B. Aditya Prakash <badityap@cc.gatech.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Time-series (TS) data are ubiquitous across various domains, including public health (Kamarthi et al., 2021b; Rodriguez et al., 2021), finance (Sezer et al., 2020), and urban computing (Tabassum et al., 2021). Time-series forecasting (TSF), a foundational task in analyzing TS data, involving predicting future events or trends based on historical TS data, has received a longstanding research focus. TSF faces certain challenges due to the dynamic and complex nature of TS data: First, distributions of TS data change over time. Second, the inherent complexity of TSF is compounded by unforeseen exogenous factors, such as policy interventions and climate changes in the context of influenza forecasting.

Given the dynamic nature of TS data, where unforeseen distribution shifts can occur between historical training and future testing data, the TSF task asks for robust out-of-distribution (OOD) generalization abilities. Instead, existing TSF models employ empirical risk minimization to greedily incorporate all correlations within the data to minimize average training errors. However, as not all correlations persist in unknown test distributions, these models may lack OOD generalization abilities. Note that existing works on temporal distribution shifts (Du et al., 2021; Kim et al., 2021; Liu et al., 2022; Fan et al., 2023) merely focus on mitigating the marginal distribution shifts of the input. These methods are not generalizable enough for the OOD problem, which consists of various types of distribution shifts (Liu et al., 2021c), such as conditional distribution shifts, etc.

In this paper, we propose to alleviate the OOD generalization problem of TSF via invariant learning (IL). IL seeks to identify and utilize invariant features that maintain stable relationships with targets across different environments while discarding unstable correlations introduced by variant features. Although IL has witnessed wide theoretical and empirical success in various domains (Koyama & Yamaguchi, 2020; Ye et al., 2023; Weber et al., 2022), it remains unexplored yet non-trivial to apply IL for TSF because of the following challenges: First, TS data breaks IL's conventional assumption. In TS data, there are always variables that directly affect targets but remain unobserved, such as the outbreak of an epidemic, sudden temperature changes, policy adjustments, etc. IL fails to consider these unob-

served core variables, leading to poor OOD generalization in TSF. Second, TS data are usually collected without explicit environment labels. Although some general IL with environment inference methods have been proposed, their neglect of TS data characteristics results in suboptimal inferred time-series environments.

Thus, we propose a novel TSF approach for out-of-distribution generalization, namely FOIL (Forecasting for Out-of-distribution TS generalization via Invariant Learning). Our contributions are summarized as follows:

- We investigate the out-of-distribution generalization problem of time-series forecasting. To the best of our knowledge, we are the first to introduce invariant learning to TSF and identify two essential gaps, including the non-compliance of IL's conventional assumption and the lack of environment labels.
- We propose FOIL, a practical and model-agnostic invariant learning framework for TSF. FOIL leverages a simple surrogate loss to ensure the applicability of IL and designs an efficient environment inference module tailored for time-series data.
- We conduct extensive experiments on diverse datasets along with three advanced forecasting models ('backbones'). FOIL proves effectiveness by uniformly outperforming all baselines in better forecasting accuracy.

2. Preliminaries and Problem Definition

We formally introduce the TSF task and discuss why it is an OOD generalization problem. We then introduce the problem OOD-TSF, formulating TSF as an OOD problem.

We denote slanted upper-cased letters such as $\boldsymbol{X}$ as random variables and calligraphic font letters $\mathcal{X}$ as its sample space. Upright bold upper-cased letters such as $\mathbf{X}$, bold lower-cased letters such as $\boldsymbol{x}$ and regular lower-cased letters such as x denote deterministic matrices, vectors and scalars, respectively.

2.1. Time-Series Forecasting: An Out-of-Distribution Generalization View

TSF models take a time series as input and output future values of some or all of its features. Let the input time-series variable be denoted as $\boldsymbol{X} \in \mathbb{R}^{l \times d_{\text{in}}}$, where l is the length of the *lookback window* decided by domain experts and d_{in} is the feature dimension at each time step. The output variable of the forecasts generated of *horizon window* length h is denoted as $\boldsymbol{Y} \in \mathbb{R}^{h \times d_{\text{out}}}$, where d_{out} is the dimension of targets at each time step. For the sample at time step t, denoted as $(\mathbf{X}_t, \mathbf{Y}_t)$, $\mathbf{X}_t \in \boldsymbol{X} = [\mathbf{x}_{t-l+1}, \mathbf{x}_{t-l+2}, \ldots, \mathbf{x}_t]$ and $\mathbf{Y}_t \in \boldsymbol{Y} = [\mathbf{y}_{t+1}, \mathbf{y}_{t+2}, \ldots, \mathbf{y}_{t+h}]$. Thus, the TSF model parameterized by θ is denoted as $f_\theta : \mathcal{X} \to \mathcal{Y}$. In this paper, we focus on univariate forecasting with covariates, i.e., $d_{\text{out}} = 1$ and $d_{\text{in}} \geq 1$, but our method can be easily generalized to the multivariate forecasting setting by using multiple univariate forecasting (Gruver et al., 2023; Lim & Zohren, 2021).

Existing TSF models usually assume the training distribution is the same as the test distribution and use empirical risk minimization (ERM) for model training. However, training and test sets of TSF represent historical and future data, respectively. Given the dynamic nature of time series, the test distribution may diverge from the training distribution. In this paper, we consider TSF under the more realistic situation where $P^{\text{train}}(\boldsymbol{X}, \boldsymbol{Y}) \neq P^{\text{test}}(\boldsymbol{X}, \boldsymbol{Y})$, i.e., unknown $P^{\text{test}}(\boldsymbol{X}, \boldsymbol{Y})$, which can be defined as follows:

Problem 1. Out-of-Distribution Generalization for Time-Series Forecasting (**OOD-TSF**): Given a time-series training dataset $\mathcal{D}^{\text{train}} = \{(\mathbf{X}_t, \mathbf{Y}_t)\}_{t=1}^{T}$, the task is to learn an out-of-distribution generalized forecasting model $f_\theta^* : \mathcal{X} \to \mathcal{Y}$ parameterized by θ which achieves minimum error on testing set $\mathcal{D}^{\text{test}}$ with unknown distribution $P^{\text{test}}(\boldsymbol{X}, \boldsymbol{Y})$.

2.2. Invariant Learning: Out-of-Distribution Generalization with Environments

Environment Labels. Invariant learning (IL), backed by the invariance principle (Arjovsky et al., 2019) from causality, is a popular solution for OOD generalization. IL assumes heterogeneity in observed data: dataset is collected from multiple environments, formulated as $\mathcal{D} = \cup_e \mathcal{D}^e = \cup_e \{(\mathbf{X}_i^e, \mathbf{Y}_i^e)\}_{i=1}^{|\mathcal{D}^e|}$; each environment e has a distinct distribution $P^e(\boldsymbol{X}, \boldsymbol{Y})$, termed heterogeneous environments. In time-series data, temporal environments can be seasons, temperatures, policies, etc. Let $\text{supp}(\boldsymbol{E})$ denote all environments, the objective function is formulated as:

$$\boldsymbol{\mathcal{R}}_{\text{IL}}(f_\theta) = \max_{e \in \text{supp}(\boldsymbol{E})} \mathbb{E}_{P(\boldsymbol{X}, \boldsymbol{Y}|e)} \left[\ell(f_\theta(\mathbf{X}), \mathbf{Y})) | e \right], \quad (1)$$

where OOD generalization is achieved by minimizing the empirical risk under the worst-performing environment.

Invariant Features. To optimize Eq. 1, IL proposes to identify and utilize invariant features that maintain stable relationships with target variables across different environments. For instance, in forecasting the number of flu cases, temperature changes belong to invariant features (Mourtzoukou & Falagas, 2007; Mäkinen et al., 2009), while hospital records are variant features since the proportion of influenza cases over all records may vary across different seasons.

Sufficiency and Invariance Assumption. Most IL methods are proposed based on the following conventional assumption (Gong et al., 2016; Rojas-Carulla et al., 2018; Kuang et al., 2020; Arjovsky et al., 2019; Liu et al., 2021a; Lin et al., 2022):

Assumption 2.1 (Conventional Assumption of Invariant Learning)**.** The input features $\boldsymbol{X}$ is a mixture of invariant features $\boldsymbol{X}_{\mathrm{I}}$ and variant features $\boldsymbol{X}_{\mathrm{V}}$. $\boldsymbol{X}_{\mathrm{I}}$ possesses the following properties:

a. **Sufficiency property:** $\boldsymbol{Y} = g(\boldsymbol{X}_{\mathrm{I}}) + \epsilon$, where $g(\cdot)$ can be any mapping function, and ϵ is random noise.

b. **Invariance property:** for all $e_i, e_j \in \mathrm{supp}(\boldsymbol{E})$, we have $P^{e_i}(\boldsymbol{Y}|\boldsymbol{X}_{\mathrm{I}}) = P^{e_j}(\boldsymbol{Y}|\boldsymbol{X}_{\mathrm{I}})$ holds.

Thus, $\boldsymbol{X}_{\mathrm{I}}$ is assumed to provide sufficient and invariant predictive power for $\boldsymbol{Y}$ and is theoretically proven to guarantee optimal OOD performance for Eq. 1 (Liu et al., 2021a).

To better understand the above, we employ the structural causal model (SCM) (Pearl et al., 2000) shown in Figure 1(a). We define invariant features $\boldsymbol{X}_{\mathrm{I}}$ as the subset of input features $\boldsymbol{X}$ that directly cause $\boldsymbol{Y}$, following (Arjovsky et al., 2019; Peters et al., 2016; Lin et al., 2022). Environment $\boldsymbol{E}$ can be interpreted as the confounder between $\boldsymbol{X}_{\mathrm{I}}$ and $\boldsymbol{X}_{\mathrm{V}}$. Specifically, the correlation between $\boldsymbol{X}_{\mathrm{V}}$ and $\boldsymbol{Y}$ is spurious, mediated through $\boldsymbol{X}_{\mathrm{V}} \leftarrow \boldsymbol{E} \rightarrow \boldsymbol{X}_{\mathrm{I}} \rightarrow \boldsymbol{Y}$. Conversely, the causal relationship $\boldsymbol{X}_{\mathrm{I}} \rightarrow \boldsymbol{Y}$ is invariant. Generally, IL aims to achieve OOD generalization using such $\boldsymbol{X}_{\mathrm{I}}$ to predict $\boldsymbol{Y}$.

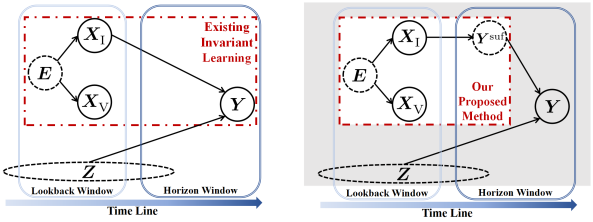
3. Challenges

Considering the theoretical and empirical successes of invariant learning (Arjovsky et al., 2019; Koyama & Yamaguchi, 2020; Krueger et al., 2021; Ye et al., 2023), a natural question arises: **Can we directly apply invariant learning (IL) to OOD-TSF?** Unfortunately, there are two main reasons rendering a direct application problematic. Firstly, the existence of unobserved variables in time-series (TS) data breaks the conventional Assumption 2.1 of IL. Secondly, TS datasets usually lack adequate environment labels.

TS data break IL's conventional assumption. Recall Assumption 2.1, where invariant features $\boldsymbol{X}_{\mathrm{I}}$ are assumed to provide sufficient and invariant predictive power for $\boldsymbol{Y}$ in IL. However, in TSF tasks, there are always variables that directly affect $\boldsymbol{Y}$ but are not included in the input features $\boldsymbol{X}$, such as the outbreak of a novel epidemic, sudden temperature changes, policy adjustments, etc. These *unobserved core variables*, denoted as $\boldsymbol{Z}$, exist due to their absence from the whole dataset or the lookback window.

In the SCM shown in Figure 1(a), we use $\boldsymbol{Z} \rightarrow \boldsymbol{Y}$ and the dash circle to describe the core effect of $\boldsymbol{Z}$ on $\boldsymbol{Y}$ and the unobserved issue of $\boldsymbol{Z}$ respectively. Clearly, there exists a gap between the SCM modeled by the existing IL methods and the SCM underlying TS data , due to the existence of $\boldsymbol{Z}$.

The existence of unobserved $\boldsymbol{Z}$ breaks both two parts of the IL's conventional assumption 2.1: First, $\boldsymbol{Z}$ breaks the sufficiency property part, obviously. Thus, existing IL methods actually absorb the influence of $\boldsymbol{Z}$ on $\boldsymbol{Y}$, leading to the overfitting issue, especially with deep models. Second, $\boldsymbol{Z}$ breaks the invariance property part when $\boldsymbol{Z}$ and $\boldsymbol{E}$ are not independent, for example, influenza outbreaks occur more frequently in winter. Formally, if there exists $e_i, e_j \in \mathrm{supp}(\boldsymbol{E})$ such that $P^{e_i}(\boldsymbol{Z}|\boldsymbol{X}_{\mathrm{I}}) \neq P^{e_j}(\boldsymbol{Z}|\boldsymbol{X}_{\mathrm{I}})$, then we have $P^{e_i}(\boldsymbol{Y}|\boldsymbol{X}_{\mathrm{I}}) = \sum_{\mathbf{Z}} P(\boldsymbol{Y}|\boldsymbol{X}_{\mathrm{I}}, \mathbf{Z}) P^{e_i}(\mathbf{Z}|\boldsymbol{X}_{\mathrm{I}}) \neq P^{e_j}(\boldsymbol{Y}|\boldsymbol{X}_{\mathrm{I}})$. Thus, existing IL methods lacks reliable OOD generalization ability for TSF.



(a) Existing IL methods. (b) Our proposed method.

Figure 1. The structural causal model (SCM) for (a) existing invariant learning methods and (b) our proposed method. The key difference is that our method targets the sufficiently predictable part of the target, i.e., $\boldsymbol{Y}^{\mathrm{suf}}$ rather than the raw $\boldsymbol{Y}$, thus making invariant learning feasible.

TS datasets usually lack environment labels. Firstly, most IL methods (Arjovsky et al., 2019; Ahuja et al., 2021; Krueger et al., 2021; Pezeshki et al., 2021; Sagawa et al., 2019) require explicit environment labels as input, which are often unavailable in TSF datasets. Due to the complexity of temporal environments, manual annotation is often difficult, expensive, and sometimes suboptimal. Secondly, existing IL with environment inference methods are fundamentally not applicable for TSF: (1) Existing IL methods show certain limitations when applying to TSF tasks: HRM (Liu et al., 2021a) and KernelHRM (Liu et al., 2021b) are based on low-dimensional raw features, while TS data are typically high-dimensional; EIIL (Creager et al., 2021) needs delicate initialization; ZIN (Lin et al., 2022) requires additional information satisfying specific conditions; and EDNIL (Huang et al., 2022) is designed for classification tasks. (2) Existing IL methods primarily cater to static data and thus overlook the characteristics of time-series data, leading to suboptimal inferred environments.

4. Our Methodology

We propose **FOIL** (Forcasting for-Out-of-distribution generalization via Invariant Learning), a model-agnostic environment-aware invariant learning framework, serving as a practical solution for the OOD-TSF problem.

4.1. Overview

High-level Idea. Our main idea is to use IL with environment inference targeting at the sufficiently predictable part of the target (we call it $\boldsymbol{Y}^{\text{suf}}$), see Figure 1(b). Specifically, inspired by the Wold's decomposition theorem (Anderson, 2011; Nerlove et al., 2014), we assume that $\boldsymbol{Y}$ can be decomposed into deterministic and uncertain parts relative to the input $\boldsymbol{X}$. Formally, $\boldsymbol{Y} = q(\boldsymbol{Y}^{\text{suf}}, \boldsymbol{Z})$, with $q(\cdot,\cdot)$ as any mapping function. Here, $\boldsymbol{Y}^{\text{suf}} \in \mathcal{Y}$, determined by the input $\boldsymbol{X}$, is deterministic, i.e., sufficiently predictable. Thus, targeting at $\boldsymbol{Y}^{\text{suf}}$, the Assumption 2.1 of sufficiency and invariance property holds, making invariant learning feasible. Additionally, considering the unpredictability of unobserved $\boldsymbol{Z}$, the optimal OOD prediction can be achieved if we are able to uncover $\boldsymbol{Y}^{\text{suf}}$ via invariant features $\boldsymbol{X}_{\text{I}}$. To this end, we propose FOIL, which serves as a practical solution for applying IL to the OOD-TSF problem.

Overall Framework. As shown in Figure 2, FOIL consists of three parts:

(1) *Label Decomposing Component* ($\boldsymbol{\mathcal{C}}_{\text{LD}}$), which decomposes sufficiently predictable $\boldsymbol{Y}^{\text{suf}}$ from observed $\boldsymbol{Y}$.

(2) *Time-Series Environment Inference Module* ($\boldsymbol{\mathcal{M}}_{\text{TEI}}$), which infers temporal environments based on learned representations from $\boldsymbol{\mathcal{M}}_{\text{TIL}}$.

(3) *Time-Series Invariant Learning Module* ($\boldsymbol{\mathcal{M}}_{\text{TIL}}$), which learns invariant representations across inferred environments from $\boldsymbol{\mathcal{M}}_{\text{TEI}}$.

In FOIL, $\boldsymbol{\mathcal{C}}_{\text{LD}}$ is the preliminary step for $\boldsymbol{\mathcal{M}}_{\text{TIL}}$ and $\boldsymbol{\mathcal{M}}_{\text{TEI}}$; $\boldsymbol{\mathcal{M}}_{\text{TIL}}$ and $\boldsymbol{\mathcal{M}}_{\text{TEI}}$ are then jointly optimized via alternating updates. During the testing phase, only $\boldsymbol{\mathcal{M}}_{\text{TIL}}$ is utilized for prediction.

As the first work of IL for TSF, FOIL is designed as a model-agnostic framework that seamlessly incorporates various off-the-shelf deep TSF models. Specifically, the backbone model can be *any deep TSF model*, denoted $\phi(\boldsymbol{X})$. We append a regressor $\rho(\cdot)$, typically a fully connected layer, on top of the learned output representations from the backbone model $\phi(\cdot)$. We denote the combined model succinctly as as $f_\theta(\boldsymbol{X}) = \rho(\phi(\boldsymbol{X}))$. $\boldsymbol{\mathcal{M}}_{\text{TIL}}$ and $\boldsymbol{\mathcal{M}}_{\text{TEI}}$ leverage the output representation $\phi(\boldsymbol{X})$, both for achieving model-agnostic and for accommodating high-dimensional inputs of TSF. We will next introduce each part.

4.2. The Label Decomposing Component

$\boldsymbol{\mathcal{C}}_{\text{LD}}$ is used to decompose the sufficiently predictable $\boldsymbol{Y}^{\text{suf}}$ from the observed $\boldsymbol{Y}$. However, accurately obtaining $\boldsymbol{Y}^{\text{suf}}$ is nearly unfeasible, owing to the lack of information about the underlying generation function and unobserved variables $\boldsymbol{Z}$. Instead of introducing additional data, such as external datasets as the agent for $\boldsymbol{Z}$, we aim to alleviate this problem

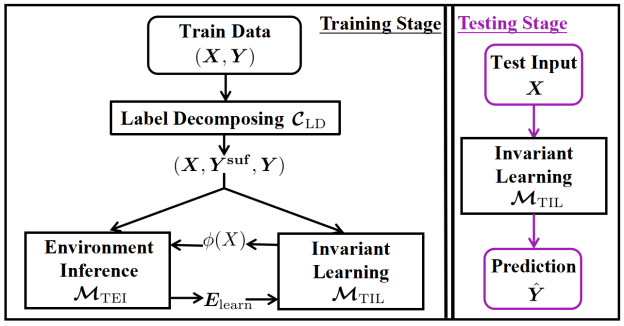


Figure 2. The overall framework of our proposed FOIL.

more practically via a surrogate loss to mitigate the effect of $\boldsymbol{Z}$. Firstly, we add the following assumption:

$$\boldsymbol{Y} = q(\boldsymbol{Y}^{\text{suf}}, \boldsymbol{Z}) = \alpha(\boldsymbol{Z})(\boldsymbol{Y}^{\text{suf}}) + \beta(\boldsymbol{Z})\mathbf{1}, \tag{2}$$

where $\alpha(\cdot) : \mathbb{R}^{d_Z} \to \mathbb{R}$ and $\beta(\cdot) : \mathbb{R}^{d_Z} \to \mathbb{R}$ could be any mapping function, and $\mathbf{1} \in \mathbb{R}^{h \times d_{\text{out}}}$ is an all-one matrix. This assumption follows the dynamic nature of observed $\boldsymbol{Y}$'s distribution (Cheng et al., 2015). Specifically, this assumption encompasses two aspects: (1) The relationships between $\boldsymbol{Z}$ and $\boldsymbol{Y}^{\text{suf}}$ are additive and multiplicative, which is a widely adopted assumption about unobserved variables (Hoyer et al., 2008; Maeda & Shimizu, 2021; Sancho et al., 1982; Wooldridge, 1997). (2) $\boldsymbol{Z}$ exerts a consistent influence in one horizon window, which can be readily extended by partitioning the horizon window into multiple segments.

Thus, the residual $\boldsymbol{Res}$ between ground truth $\boldsymbol{Y}$ and predicted $\hat{\boldsymbol{Y}}$, i.e., $\boldsymbol{Res} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$, absorb the effect of $\boldsymbol{Z}$ on $\boldsymbol{Y}$ via values of mean $\mu(\boldsymbol{Res})$ and standard deviation $\sigma(\boldsymbol{Res})$. Thus, we propose an Instance Residual Normalization (IRN) method to mitigate the effect of $\boldsymbol{Z}$. For the residual $\boldsymbol{Res}_t$ of instance t, IRN method can be formulated as:

$$\tilde{\mathbf{Res}}_t = \frac{\mathbf{Y}_t - \mu(\mathbf{Y}_t)}{\sigma(\mathbf{Y}_t)} - \frac{\hat{\mathbf{Y}}_t - \mu\left(\hat{\mathbf{Y}}_t\right)}{\sigma(\hat{\mathbf{Y}}_t)} = \tilde{\mathbf{Y}}_t - \tilde{\hat{\mathbf{Y}}}_t \tag{3}$$

IRN in Eq. 3 ensures the residuals to have a mean of 0 and a variance of $2 - 2\text{cov}(\hat{\mathbf{Y}}, \mathbf{Y})$, where cov denotes the covariance.

Finally, we derive the following simple and effective surrogate loss to mitigate the effect of $\boldsymbol{Z}$, instead of directly decoupling $\boldsymbol{Y}^{\text{suf}}$ in $\boldsymbol{\mathcal{C}}_{\text{LD}}$:

$$\ell_{\text{suf}}(\hat{\boldsymbol{Y}}, \boldsymbol{Y}) = \text{MSE}(\tilde{\boldsymbol{Res}}, \mathbf{0}) = \ell(\tilde{\hat{\boldsymbol{Y}}}, \tilde{\boldsymbol{Y}}), \tag{4}$$

where $\text{MSE}(\tilde{\boldsymbol{Res}}, \mathbf{0}) = \frac{1}{h}\sum_{j=1}^{h}(\tilde{\mathbf{Res}}_{t+j})^2$. Note that our IRN fundamentally differs from the existing instance normalization (IN) methods. Existing methods adopt IN to $\boldsymbol{X}$, and reverse IN to $\hat{\boldsymbol{Y}}$ based on $\mu(\boldsymbol{X})$ and $\sigma(\boldsymbol{X})$, aiming to

address non-stationary problem of $\boldsymbol{X}$ (Kim et al., 2021; Liu et al., 2022). While, our IRN method directly aligns the mean and variance between $\hat{\boldsymbol{Y}} = f(\boldsymbol{X})$ and $\boldsymbol{Y}$, thus removing error caused by $\boldsymbol{Z}$ under the introduced assumption. Since $\boldsymbol{Z}$ is not contained in $\boldsymbol{X}$, existing methods usually fail to achieve our goal.

4.3. The Time-Series Environment Inference Module

$\boldsymbol{\mathcal{M}}_{\text{TEI}}$ aims to infer environments $\boldsymbol{E}_{\text{infer}}$, thereby providing environment labels for the time-series invariant learning module $\boldsymbol{\mathcal{M}}_{\text{TIL}}$. We consider inferring effective and reasonable temporal environments with two goals:

(1) Sensitive to the encoded invariant features. In FOIL, $\boldsymbol{\mathcal{M}}_{\text{TEI}}$ and $\boldsymbol{\mathcal{M}}_{\text{TIL}}$ are adversarial: $\boldsymbol{\mathcal{M}}_{\text{TEI}}$ infers environments based on the variant features not discarded by $\boldsymbol{\mathcal{M}}_{\text{TIL}}$; $\boldsymbol{\mathcal{M}}_{\text{TIL}}$ discards variant features based on inferred environments from $\boldsymbol{\mathcal{M}}_{\text{TEI}}$. Ultimately, when $\boldsymbol{\mathcal{M}}_{\text{TIL}}$ only utilizes invariant features, $\boldsymbol{\mathcal{M}}_{\text{TEI}}$ is unable to infer effective environments. Thus, we propose to infer informative environments that are sensitive to the variant features encoded in the currently learned representations, formulated as:

$$\min_{\boldsymbol{E}_{\text{infer}}} H\left(\boldsymbol{Y}^{\text{suf}} | \phi^*(\boldsymbol{X}), \boldsymbol{E}_{\text{infer}}\right) - H\left(\boldsymbol{Y}^{\text{suf}} | \phi^*(\boldsymbol{X})\right), \quad (5)$$

where H is Shannon conditional entropy, $\phi^*(\boldsymbol{X})$ are learned representations from $\boldsymbol{\mathcal{M}}_{\text{TIL}}$ and frozen in $\boldsymbol{\mathcal{M}}_{\text{TEI}}$.

(2) Preserving the temporal adjacency structures. To ensure that the inferred environments are reasonable in the context of TSF, we consider preserving the inherent characteristic of time-series data, i.e., the temporal adjacency structure. Specifically, instances that are temporally adjacent should possess similar temporal environments. This can also be viewed as a type of regularization to prevent inferred environments from overfitting to random noises.

Intuitively, the approach to infer environments is to optimize Eq. 5, with a plugin for preserving the temporal adjacency structure. To this end, we present an EM-based clustering solution in the representation space, implemented through a multi-head neural network. Each head is an environment-specific regressor, playing the role of each cluster's center. Specifically, the regressor $\rho^{(e)}$ is specific for environment e. And the representation $\phi^*(\boldsymbol{X})$ is shared and frozen in $\boldsymbol{\mathcal{M}}_{\text{TEI}}$. We describe the M step and E step next.

M Step: Optimizing Environment-Specific Regressors
In the M step, we optimize $\{\rho^{(e)}\}$ to better fit the data from current environment partition $\boldsymbol{E}_{\text{infer}}$ of E step as:

$$\begin{aligned}\min_{\{\rho^{(e)}\}} \boldsymbol{\mathcal{L}}_{\text{TEI}} &= \mathbb{E}_{e \in \boldsymbol{E}_{\text{infer}}} \boldsymbol{\mathcal{R}}^{(e)}_{\text{suf}}(\rho^{(e)}, \phi^*) \\ &= \sum_{e \in \boldsymbol{E}_{\text{infer}}} \frac{1}{|\mathcal{D}_e|} \sum_{(\mathbf{X},\mathbf{Y}) \in \mathcal{D}_e} \ell_{\text{suf}}\left(\rho^{(e)}\left(\phi^*(\mathbf{X})\right), \mathbf{Y}\right)\end{aligned} \quad (6)$$

E Step: Estimating Environment Labels
Next, in the E step, we reallocate the environment partitions. For instance $(\mathbf{X}_t, \mathbf{Y}_t)$, we reassign its environment label $\boldsymbol{E}_{\text{infer}}(t)$ via the following two steps:

- Step 1: Reallocating based on the distances with the center of each cluster (environment). We use the loss with respect to regressor $\rho^{(e)}$ to describe the distance with the center of cluster e. Thus, we reassign $\boldsymbol{E}_{\text{infer}}(t)$ according to the shortest distance, as follows:

$$\boldsymbol{E}_{\text{infer}}(t) \leftarrow \arg\min_{e \in \boldsymbol{E}_{\text{infer}}} \left\{ \ell_{\text{suf}}\left(\rho^{(e)}\left(\phi^*(\mathbf{X}_t)\right), \mathbf{Y}_t\right)\right\} \quad (7)$$

- Step 2: Reallocating to preserve temporal adjacency structure. We propose an environment label propagation solution to achieve this goal, as follows:

$$\boldsymbol{E}_{\text{infer}}(t) \leftarrow \text{mode}\left\{\boldsymbol{E}_{\text{infer}}(t+j)\right\}_{j=-r}^{r}, \quad (8)$$

  where mode implements majority voting by considered temporal neighbors selected via the radius $r \in \mathbb{Z}^+$.

In summary, we iteratively execute M step and E step to obtain the inferred environments $\boldsymbol{E}^*_{\text{infer}}$. Due to the fixed second term of Eq. 5, our solution represents a practical instantiation of Eq. 5.

4.4. The Time-Series Invariant Learning Module

$\boldsymbol{\mathcal{M}}_{\text{TIL}}$ is used to learn invariant representations $\phi^*(\boldsymbol{X})$ across inferred environments $\boldsymbol{E}^*_{\text{infer}}$ from $\boldsymbol{\mathcal{M}}_{\text{TEI}}$. Specifically, $\boldsymbol{\mathcal{M}}_{\text{TIL}}$ aims to learn the $\phi^*(\boldsymbol{X})$ which encode and solely encode all the information of invariant features $\boldsymbol{X}_{\text{I}}$ thus achieving both invariant and sufficient predictive capability targeting at $\boldsymbol{Y}^{\text{suf}}$. Such $\phi^*(X)$ has been theoretically proven (Liu et al., 2021a) to be obtained by optimizing the following objective function:

$$\phi^* = \arg\max_{\phi} I(\mathbf{Y}^{\text{suf}}; \phi(\mathbf{X}) - I(\mathbf{Y}^{\text{suf}}; \boldsymbol{E}^*_{\text{learn}} | \phi(\mathbf{X})), \quad (9)$$

where $I(\cdot;\cdot)$ measures Shannon mutual information. The first and second terms correspond to ensure sufficiency and invariance property of $\phi(\boldsymbol{X})$, respectively.

Considering the unavailability of $\boldsymbol{Y}^{\text{suf}}$, we present the following practical loss function as the instantiation of Eq. 9 via our surrogate loss in Eq. 4:

$$\begin{aligned}\min_{\rho,\phi} \mathcal{L}_{\text{TIL}} =& \mathbb{E}_{e \in \boldsymbol{E}^*_{\text{infer}}} \boldsymbol{\mathcal{R}}^{(e)}_{\text{suf}}(\rho, \phi) + \lambda_1 \boldsymbol{\mathcal{R}}_{\text{ERM}}(\rho, \phi) \\ &+ \lambda_2 \text{Var}_{e \in \boldsymbol{E}^*{}_{\text{infer}}}\left[\boldsymbol{\mathcal{R}}^{(e)}_{\text{suf}}(\rho, \phi)\right],\end{aligned} \quad (10)$$

where λ_1, λ_2 are hyper-parameters, $\boldsymbol{\mathcal{R}}_{\text{ERM}}(\rho, \phi) = \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\ell(\rho(\phi(\mathbf{X})), \mathbf{Y})\right]$ is the ERM loss on raw $\boldsymbol{Y}$,

$\mathcal{R}^{e}_{\text{suf}}(\rho, \phi)$ defined in Eq. 10 is the loss of inferred environment e on $\boldsymbol{Y}^{\text{suf}}$, and $\text{Var}_{e \in \boldsymbol{E}^{*}_{\text{infer}}}\left[\mathcal{R}^{(e)}_{\text{suf}}(\rho, \phi)\right]$ implies the variance of loss across inferred environments. The first and second terms are jointly used to ensure the sufficient predictive power of $\phi(\boldsymbol{X})$ for $\boldsymbol{Y}^{\text{suf}}$, where λ_1 controls the trade-off between introducing information of $\mu(\boldsymbol{Y}^{\text{suf}}), \sigma(\boldsymbol{Y}^{\text{suf}})$ and the influence of $\boldsymbol{Z}$. The third term further balanced by λ_2 ensures the invariance property and is robust to marginal distribution shifts of input, theoretically guaranteed by (Krueger et al., 2021) and further balanced by λ_2.

The overall algorithm is summarized in Appendix A. Compared to the backbone, FOIL slightly increases the parameter count due to additional multiple regressors.

5. Experiments

5.1. Setup

Datasets. We conduct experiments on four popular real-world datasets commonly used as benchmarks: the daily reported exchange rates dataset (**Exchange**) (Lai et al., 2018), the weekly reported ratios of patients seen with influenza-like illness dataset (**ILI**) (Kamarthi et al., 2021a), and two hourly reported electricity transformer temperature datasets (**ETTh1** and **ETTh2**) (Zhou et al., 2021). We adhere to the general setups and target variables selections, following previous literatures (Wu et al., 2021; 2022; Nie et al., 2022).

Backbones. As previously mentioned, our proposed FOIL is a model-agnostic framework. We select three different types of TSF models as backbones. **Informer** (Zhou et al., 2021) proposes an efficient transformer for long-term TSF. **Crossformer** (Zhang & Yan, 2022) better utilizes cross-dimension dependency, making it more sensitive to spuriouse correlations. **PatchTST** (Nie et al., 2022) employs channel-independent and patching strategies to achieve state-of-the-art performance.

Baselines. We comprehensively compare the following twelve distribution shifts baselines: (1) Two advanced methods for handling temporal distribution shifts in TSF: **NST** (Liu et al., 2022) and **RevIN** (Kim et al., 2021). (2) Six well-acknowledged general OOD methods following (Gagnon-Audet et al., 2022), adopted due to the lack of OOD methods for TSF: (a) Methods requiring environment labels: **GroupDRO** (Sagawa et al., 2019), **IRM** (Arjovsky et al., 2019), **IB-ERM** (Ahuja et al., 2021), **VREx** (Krueger et al., 2021) and **SD** (Pezeshki et al., 2021). (b) Methods not requiring environment labels: **EIIL** (Creager et al., 2021). (3) Two hybrid methods: **IRM+RevIN** and **EIIL+RevIN**.

Implementation. Regarding the horizon window length, we considered a range from short to long-term TSF tasks. For ETTh1, ETTh2, and Exchange, the lengths are [24, 48, 96, 168, 336, 720] with a fixed lookback window size of 96 and a consistent label window size of 48 for the decoder. For the weekly reported ILI, the lengths are [4, 8, 12, 16, 20, 24], representing the next one month to six months, with a fixed lookback window size of 36 and a consistent label window size of 18 for the decoder.. Note that, we lack the availability of suitable environment labels. We address this by dividing the training set into k, tuned from 2 to 10, equal-length time segments to serve as predefined environment labels. When the backbone is equipped with our FOIL, the model architecture of the backbone remains unchanged.

Evaluation. We employ the widely-adopted evaluation metrics: mean squared error (**MSE**) and mean absolute error (**MAE**). We report average performance over three independent runs for each model.

Reproducibility. All training data, testing data and code are available at: `https://github.com/AdityaLab/FOIL`. More experimental details are revealed in Appendix B.

5.2. Results

As shown in Table 1, we present results for both original versions and corresponding FOIL equipped versions of backbones, yielding the following observations:

(1) Overall, FOIL consistently and significantly improves the performance of various TSF backbones across all datasets and forecasting lengths with improvements reaching up to 85% on MSE, thereby demonstrating FOIL's effectiveness. For the state-of-the-art PatchTST, FOIL consistently enhances performance, achieving up to 30% improvement. For the lower-performing Informer, FOIL shows more significant improvements, frequently by an order of magnitude, yielding competitive results.

(2) FOIL excels in short-term forecasting compared to long-term forecasting, as the higher uncertainty of the latter hinders learning invariant features. Moreover, FOIL's most significant improvement in the ILI dataset is attributed to the serious OOD shifts in its test data, particularly during the unseen COVID-19 period.

5.3. Comparison with Distribution Shifts Methods

In this section, we conduct a comparative analysis of the performance disparities between FOIL and existing distribution shifts methods. We employ the Informer as the forecasting backbone. The forecasting length is set as 16 for ILI and 96 for others. Similar observations are found in other settings. We measure the relative improvement compared to the best-performing baseline on each metric and dataset.

As shown in Table 2, our observations include:

(1) FOIL achieves the best performance across all datasets. The average improvements on MSE and MAE are more

Method		Informer(AAAI'21)		with FOIL		Crossformer(ICLR'23)		with FOIL		PatchTST(ICLR'23)		with FOIL	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Exchange	24	0.812	0.736	**0.036**	**0.146**	0.083	0.233	**0.029**	**0.129**	0.092	0.229	**0.031**	**0.136**
	48	0.715	0.682	**0.063**	**0.191**	0.164	0.328	**0.054**	**0.175**	0.090	0.243	**0.052**	**0.171**
	96	0.782	0.710	**0.142**	**0.274**	0.214	0.381	**0.111**	**0.240**	0.142	0.291	**0.107**	**0.235**
	192	0.708	0.701	**0.236**	**0.369**	0.709	0.716	**0.213**	**0.349**	0.364	0.468	**0.226**	**0.351**
	336	1.587	1.063	**0.546**	**0.591**	2.158	1.231	**0.471**	**0.500**	0.512	0.540	**0.465**	**0.486**
	720	3.922	1.793	**0.712**	**0.679**	2.093	1.215	**1.193**	**0.833**	0.957	0.738	**0.925**	**0.722**
	IMP.			**80.58%**	**61.24%**			**61.90%**	**45.06%**			**30.60%**	**21.11%**
ILI	4	3.212	1.530	**0.736**	**0.593**	2.147	1.232	**0.332**	**0.400**	1.043	0.587	**0.616**	**0.507**
	8	3.668	1.642	**0.881**	**0.667**	2.678	1.403	**0.569**	**0.512**	0.638	0.557	**0.586**	**0.546**
	12	3.974	1.722	**1.069**	**0.768**	2.914	1.476	**0.706**	**0.575**	0.959	0.795	**0.560**	**0.519**
	16	4.187	1.773	**1.047**	**0.779**	3.496	1.628	**0.701**	**0.568**	0.726	0.563	**0.696**	**0.555**
	20	4.296	1.806	**1.011**	**0.797**	3.589	1.653	**0.702**	**0.596**	0.807	0.705	**0.571**	**0.541**
	24	4.445	1.844	**1.014**	**0.806**	3.513	1.633	**0.686**	**0.604**	1.072	0.850	**0.663**	**0.625**
	IMP.			**75.80%**	**57.37%**			**79.99%**	**64.03%**			**27.04%**	**16.91%**
ETTh1	24	0.219	0.392	**0.038**	**0.146**	0.194	0.400	**0.028**	**0.126**	0.031	0.136	**0.027**	**0.126**
	48	0.474	0.638	**0.065**	**0.190**	0.270	0.465	**0.042**	**0.156**	0.044	0.160	**0.041**	**0.154**
	96	0.965	0.892	**0.088**	**0.224**	0.146	0.312	**0.056**	**0.181**	0.061	0.190	**0.056**	**0.182**
	192	1.029	0.967	**0.148**	**0.299**	0.241	0.420	**0.075**	**0.209**	0.082	0.223	**0.078**	**0.215**
	336	0.677	0.769	**0.136**	**0.296**	0.246	0.425	**0.088**	**0.233**	0.100	0.246	**0.092**	**0.237**
	720	1.086	0.973	**0.132**	**0.288**	0.392	0.554	**0.104**	**0.254**	0.154	0.310	**0.120**	**0.272**
	IMP.			**85.38%**	**68.14%**			**73.04%**	**54.43%**			**10.48%**	**5.80%**
ETTh2	24	0.668	0.705	**0.121**	**0.275**	0.136	0.299	**0.071**	**0.198**	0.080	0.215	**0.071**	**0.197**
	48	0.999	0.866	**0.258**	**0.407**	0.122	0.274	**0.106**	**0.248**	0.106	0.248	**0.103**	**0.241**
	96	3.070	1.628	**0.222**	**0.369**	0.256	0.408	**0.137**	**0.286**	0.156	0.309	**0.140**	**0.289**
	192	3.548	1.768	**0.699**	**0.682**	1.257	1.034	**0.198**	**0.352**	0.217	0.374	**0.201**	**0.356**
	336	2.663	1.526	**0.801**	**0.756**	1.305	1.027	**0.234**	**0.389**	0.233	0.390	**0.216**	**0.372**
	720	2.335	1.422	**0.730**	**0.725**	1.579	1.158	**0.253**	**0.402**	0.317	0.448	**0.238**	**0.391**
	IMP.			**78.03%**	**58.82%**			**59.61%**	**44.42%**			**10.65%**	**6.64%**

Table 1. Performance comparison between original and FOIL equipped versions of backbones. The top-performing version is marked in **bold**. IMP. is the average percentage improvement across lengths of horizon window compared to the original version. FOIL consistently and significantly enhances the performance of various TSF backbones on all datasets and metrics across horizon window lengths.

than 10% and 5.5% respectively, showing the benefits of FOIL over existing distribution shift methods. Notably, though hybrid models additionally alleviate the temporal distribution shift problem and exhibit better performance than general OOD baselines, FOIL still outperforms hybrid models by over 11%. Therefore, our proposed surrogate loss in Eq. 4 is irreplaceable by current instance normalization methods as discussed in Section 4.2 and exhibits important benefits for alleviating unobserved core covariates issues in the TSF task.

(2) General OOD methods exhibit poor performances. This verifies that directly applying existing invariant learning methods for the TSF task is inappropriate, as discussed in Section 3.

(3) Among the existing general OOD methods, EIIL exhibits better performance than other baselines, due to their capability to infer proper environments from the data. Besides, the performances of EIIL also suggest the advantages of inferring environments at representation spaces as opposed to raw feature spaces for TSF's high-dimensional input. These observed advantages align with the considerations made in FOIL.

5.4. Ablation Study

To demonstrate the effectiveness of each module or loss in FOIL, we conduct an ablation study that introduces three ablated versions of FOIL: (1) FOIL \Suf: remove the surrogate loss in Eq. 4 for decomposing Sufficiently predictable $\boldsymbol{Y}^{\text{suf}}$ (2) FOIL \TEI: remove the whole Time-series Environment Inference module detailed in Section 4.3,i.e. set the number of environment as 1.(3) FOIL \LP: removed the Label Propagation approach in $\boldsymbol{\mathcal{M}}_{\text{TEI}}$ in Eq. 8. All other experiment setups follow Section 5.3. The ablation study results are shown in Figure 3(a).

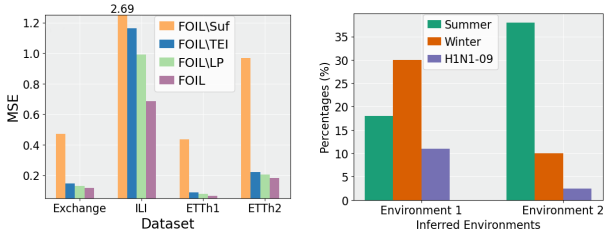
Though FOIL outperforms all ablated versions in forecasting accuracy, all designed modules and loss in FOIL show individual effectiveness through the ablation study. Specifically, the performance FOIL \Suf drops significantly more than other ablated versions, which indicates the necessity of mitigating unobserved covariate issues when applying invariant learning for TSF. Moreover, FOIL \TEI consistently outperforms FOIL \LP across all datasets, which validates the effectiveness of preserving the temporal adjacency structure for Time Series Forecasting (TSF).

Dataset			Exchange		ILI		ETTh1		ETTh2	
Type	Env. Known?	Method	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Base	No	ERM	0.782	0.710	3.974	1.722	0.965	0.892	3.070	1.628
General OOD (Invariant Learning)	Yes	GroupDRO	0.781	0.715	3.721	1.888	0.880	0.863	3.192	1.647
		IRM	0.716	0.688	3.608	1.732	0.495	0.646	2.910	1.581
		VREx	0.781	0.715	3.671	1.875	0.874	0.859	3.238	1.662
		SD	0.782	0.716	3.674	1.677	0.891	0.870	3.246	1.664
		IB-ERM	0.787	0.719	3.673	1.677	0.883	0.865	3.209	1.654
	No	EIIL	0.540	0.630	3.251	1.648	0.673	0.783	1.252	1.013
Temporal Shifts	NA	RevIN	0.169	0.296	1.350	0.867	0.108	0.248	0.236	0.387
		NST	0.151	0.281	1.351	0.871	0.118	0.260	0.258	0.406
Hybrid	Yes	IRM+RevIN	0.160	0.291	1.328	0.863	0.105	0.244	0.234	0.381
	No	EIIL+RevIN	0.170	0.309	1.205	0.820	0.097	0.241	0.343	0.483
Ours	No	FOIL	**0.136**	**0.274**	**1.047**	**0.768**	**0.088**	**0.224**	**0.210**	**0.358**
Improvement(%)			+9.93	+2.50	+12.94	+5.00	+9.27	+7.05	+10.26	+6.04

Table 2. Comparison with existing distribution shifts methods across four datasets using Informer backbone. The best results are in **bold**. NA means not considering environments. Our FOIL outperforms all existing distribution shift methods on all datasets and both metrics.

5.5. Case Study: Analysis of Inferred Environments

To justify the reasonableness of the environments inferred by FOIL, we conduct a case study on the ILI dataset by demonstrating the contribution disparities among three major components (Summer, i.e., June to August annually; Winter, i.e., December to February annually; and the H1N1-09 period, i.e., April 2009 to August 2010) when the total number of inferred environments is set to 2. The visualization of contributions from each component is shown in Figure 3(b). The visualization results align with public health perspectives in two ways: First, the major components of Environment 1 and 2 are distinguished by Winter and Summer, as influenza is a seasonal disease and typically spreads during the winter and ends before the summer. Second, the H1N1-09 period has more contributions in Environment 1 than 2, which aligns with the fact that the H1N1-09 period and winter flu seasons exhibit similarities. These observations support the ability of FOIL to infer meaningful environments in real-world TSF applications.



(a) Ablation study of our method and three ablated versions showing the effectiveness of the model design.

(b) Analysis of two inferred environments on ILI showing significant differences in component weights.

Figure 3. Results of analytical experiments.

6. Additional Related Works

6.1. Time Series Forecasting

Classical TSF models (Tsay, 2000; Ariyo et al., 2014; Box et al., 2015) often face limitations in capturing complex patterns due to their inherent model constraints. Recent advancements in deep learning methods, such as Recurrent Neural Networks (RNN) and Transformer (Rumelhart et al., 1986; Hochreiter & Schmidhuber, 1997; Vaswani et al., 2017), have led to sophisticated deep TSF models including Informer, Reformer, Autoformer, Fedformer, and PatchTST (Zhou et al., 2021; Kitaev et al., 2020; Wu et al., 2021; Zhou et al., 2022; Nie et al., 2022), significantly improving forecasting accuracy. However, these advanced models primarily rely on ERM with simple IID assumptions. Consequently, they exhibit shortcomings in OOD generalization when faced with potential distribution shifts in TS data.

6.2. Distribution Shifts in Time-Series Forecasting.

In addition to the aforementioned TSF methods in handling marginal distribution shifts (Passalis et al., 2019; Kim et al., 2021; Liu et al., 2022; Fan et al., 2023; Du et al., 2021), there are some efforts that have tackled OOD challenges in TSF. However, all have certain limitations. For example, DIVERSITY (Lu et al., 2022; 2023) is specifically designed for time series classification and detection tasks. OneNet (Zhang et al., 2023) is tailored for online forecasting scenarios by online ensembling. Pets (Zhao et al., 2023) focuses on distribution shifts induced by the specific phenomenon of performativity. This highlights the need for a

general OOD method applicable across diverse TSF scenarios and models.

Despite the existing benchmark WOODS (Gagnon-Audet et al., 2022) that evaluates IL methods combined with TSF models with a focus on TS classification tasks, our proposed approach addresses diverse datasets under realistic TSF scenarios, offering different and comprehensive problem formulation, methodology, and evaluations.

7. Conclusion and Discussion

In this paper, we formally study the fundamental out-of-distribution challenges in time-series forecasting tasks (OOD-TSF). We identify specific gaps when applying existing invariant learning methods to OOD-TSF, including theoretical violations of sufficiency and invariance assumptions and the empirical absence of environment labels in time-series datasets. To address these challenges, we introduce a model-agnostic framework named FOIL, which employs an innovative surrogate loss to alleviate the impact of unobserved variables. FOIL features a joint optimization strategy, which learns invariant representations and preserves temporal adjacency structure. Empirical evaluations demonstrate the effectiveness of FOIL by consistently improving the performances of different TSF models and outperforming other OOD solutions.

Beyond the scope of FOIL, it is important to recognize that invariant learning is not the only solution to enhance OOD generalization in TSF tasks. Alternative approaches or interpretations can require advanced causal analysis, feature selections, or learning dynamic temporal patterns. The using of additional information to enhance the sufficiency of predictions also deserves to be explored. We also emphasize the need for conscientious evaluations on underrepresented subgroups when implementing our approach in real-world scenarios for promoting fairness among subgroups. We expect that future research will delve into these open questions, contributing both theoretically and practically to advance the understanding of OOD-TSF challenges and achieve more reliable TSF models.

Impact Statement

Our work introduces a new methodology to improve the out-of-distribution generalization of time-series forecasting models and is applicable across wide range of domains and real-world applications including sensitive applications in public health, economics, etc. Therefore, care should be taken in alleviating biases and disparities in dataset as well as making sure the predictions of model do not pose ethical risks or lead to inequitable outcomes across various stakeholders relevant to specific applications our methods are used.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This paper was supported in part by the NSF (Expeditions CCF-1918770, CAREER IIS-2028586, Medium IIS-1955883, Medium IIS-2106961, PIPP CCF-2200269, IIS-2008334, CAREER IIS-2144338), CDC MInD program, Meta faculty gifts, and funds/computing resources from Georgia Tech.

References

Ahuja, K., Caballero, E., Zhang, D., Gagnon-Audet, J.-C., Bengio, Y., Mitliagkas, I., and Rish, I. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

Anderson, T. W. *The statistical analysis of time series*. John Wiley & Sons, 2011.

Ariyo, A. A., Adewumi, A. O., and Ayo, C. K. Stock price prediction using the arima model. In *2014 UKSim-AMSS 16th international conference on computer modelling and simulation*, pp. 106–112. IEEE, 2014.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

Cheng, C., Sa-Ngasoongsong, A., Beyca, O., Le, T., Yang, H., Kong, Z., and Bukkapatnam, S. T. Time series forecasting for nonlinear and non-stationary processes: a review and comparative study. *Iie Transactions*, 47(10): 1053–1071, 2015.

Creager, E., Jacobsen, J.-H., and Zemel, R. Environment inference for invariant learning. In *International Conference on Machine Learning*, pp. 2189–2200. PMLR, 2021.

Du, Y., Wang, J., Feng, W., Pan, S., Qin, T., Xu, R., and Wang, C. Adarnn: Adaptive learning and forecasting of time series. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pp. 402–411, 2021.

Fan, W., Wang, P., Wang, D., Wang, D., Zhou, Y., and Fu, Y. Dish-ts: a general paradigm for alleviating distribution shift in time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7522–7529, 2023.

Gagnon-Audet, J.-C., Ahuja, K., Darvishi-Bayazi, M.-J., Mousavi, P., Dumas, G., and Rish, I. Woods: Benchmarks for out-of-distribution generalization in time series. *arXiv preprint arXiv:2203.09978*, 2022.

Gong, M., Zhang, K., Liu, T., Tao, D., Glymour, C., and Schölkopf, B. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pp. 2839–2848. PMLR, 2016.

Gruver, N., Finzi, M., Qiu, S., and Wilson, A. G. Large language models are zero-shot time series forecasters. *arXiv preprint arXiv:2310.07820*, 2023.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

Huang, B.-W., Liao, K.-T., Kao, C.-S., and Lin, S.-D. Environment diversification with multi-head neural network for invariant learning. *Advances in Neural Information Processing Systems*, 35:915–927, 2022.

Kamarthi, H., Kong, L., Rodriguez, A., Zhang, C., and Prakash, B. A. When in doubt: Neural non-parametric uncertainty quantification for epidemic forecasting. *Advances in Neural Information Processing Systems*, 34: 19796–19807, 2021a.

Kamarthi, H., Kong, L., Rodriguez, A., Zhang, C., and Prakash, B. A. When in doubt: Neural non-parametric uncertainty quantification for epidemic forecasting. *Advances in Neural Information Processing Systems*, 34: 19796–19807, 2021b.

Kamarthi, H., Kong, L., Rodríguez, A., Zhang, C., and Prakash, B. A. Camul: Calibrated and accurate multi-view time-series forecasting. In *Proceedings of the ACM Web Conference 2022*, pp. 3174–3185, 2022.

Kim, T., Kim, J., Tae, Y., Park, C., Choi, J.-H., and Choo, J. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2021.

Kitaev, N., Kaiser, Ł., and Levskaya, A. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

Koyama, M. and Yamaguchi, S. When is invariance useful in an out-of-distribution generalization problem? *arXiv preprint arXiv:2008.01883*, 2020.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

Kuang, K., Xiong, R., Cui, P., Athey, S., and Li, B. Stable prediction with model misspecification and agnostic distribution shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 4485–4492, 2020.

Lai, G., Chang, W.-C., Yang, Y., and Liu, H. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 95–104, 2018.

Lim, B. and Zohren, S. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, 379(2194):20200209, 2021.

Lin, Y., Zhu, S., Tan, L., and Cui, P. Zin: When and how to learn invariance without environment partition? *Advances in Neural Information Processing Systems*, 35: 24529–24542, 2022.

Liu, J., Hu, Z., Cui, P., Li, B., and Shen, Z. Heterogeneous risk minimization. In *International Conference on Machine Learning*, pp. 6804–6814. PMLR, 2021a.

Liu, J., Hu, Z., Cui, P., Li, B., and Shen, Z. Integrated latent heterogeneity and invariance learning in kernel space. *Advances in Neural Information Processing Systems*, 34: 21720–21731, 2021b.

Liu, J., Shen, Z., He, Y., Zhang, X., Xu, R., Yu, H., and Cui, P. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021c.

Liu, Y., Wu, H., Wang, J., and Long, M. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in Neural Information Processing Systems*, 35:9881–9893, 2022.

Lu, W., Wang, J., Sun, X., Chen, Y., and Xie, X. Out-of-distribution representation learning for time series classification. In *The Eleventh International Conference on Learning Representations*, 2022.

Lu, W., Wang, J., Sun, X., Chen, Y., Ji, X., Yang, Q., and Xie, X. Diversify: A general framework for time series out-of-distribution detection and generalization. *arXiv preprint arXiv:2308.02282*, 2023.

Maeda, T. N. and Shimizu, S. Causal additive models with unobserved variables. In *Uncertainty in Artificial Intelligence*, pp. 97–106. PMLR, 2021.

Mäkinen, T. M., Juvonen, R., Jokelainen, J., Harju, T. H., Peitso, A., Bloigu, A., Silvennoinen-Kassinen, S., Leinonen, M., and Hassi, J. Cold temperature and low humidity are associated with increased occurrence of respiratory tract infections. *Respiratory medicine*, 103(3): 456–462, 2009.

Mourtzoukou, E. and Falagas, M. E. Exposure to cold and respiratory tract infections. *The International Journal of Tuberculosis and Lung Disease*, 11(9):938–943, 2007.

Nerlove, M., Grether, D. M., and Carvalho, J. L. *Analysis of economic time series: a synthesis*. Academic Press, 2014.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2022.

Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M., and Iosifidis, A. Deep adaptive input normalization for time series forecasting. *IEEE transactions on neural networks and learning systems*, 31(9):3760–3765, 2019.

Pearl, J. et al. Causality: Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2): 3, 2000.

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

Pezeshki, M., Kaba, O., Bengio, Y., Courville, A. C., Precup, D., and Lajoie, G. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021.

Rodriguez, A., Tabassum, A., Cui, J., Xie, J., Ho, J., Agarwal, P., Adhikari, B., and Prakash, B. A. Deepcovid: An operational deep learning-driven framework for explainable real-time covid-19 forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 15393–15400, 2021.

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Sancho, J. M., San Miguel, M., Katz, S., and Gunton, J. Analytical and numerical studies of multiplicative noise. *Physical Review A*, 26(3):1589, 1982.

Sezer, O. B., Gudelek, M. U., and Ozbayoglu, A. M. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied soft computing*, 90:106181, 2020.

Tabassum, A., Chinthavali, S., Tansakul, V., and Prakash, B. A. Actionable insights in multivariate time-series for urban analytics. 2021.

Tsay, R. S. Time series and forecasting: Brief history and future research. *Journal of the American Statistical Association*, 95(450):638–643, 2000.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Weber, M. G., Li, L., Wang, B., Zhao, Z., Li, B., and Zhang, C. Certifying out-of-domain generalization for blackbox functions. In *International Conference on Machine Learning*, pp. 23527–23548. PMLR, 2022.

Wooldridge, J. M. Multiplicative panel data models without the strict exogeneity assumption. *Econometric Theory*, 13(5):667–678, 1997.

Wu, H., Xu, J., Wang, J., and Long, M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 34:22419–22430, 2021.

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., and Long, M. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

Ye, N., Zhu, L., Wang, J., Zeng, Z., Shao, J., Peng, C., Pan, B., Li, K., and Zhu, J. Certifiable out-of-distribution generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10927–10935, 2023.

Zhang, Y. and Yan, J. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*, 2022.

Zhang, Y.-F., Wen, Q., Wang, X., Chen, W., Sun, L., Zhang, Z., Wang, L., Jin, R., and Tan, T. Onenet: Enhancing time series forecasting models under concept drift by online ensembling. *arXiv preprint arXiv:2309.12659*, 2023.

Zhao, Z., Rodriguez, A., and Prakash, B. A. Performative time-series forecasting. *arXiv preprint arXiv:2310.06077*, 2023.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., and Jin, R. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, pp. 27268–27286. PMLR, 2022.

A. Algorithm

Algorithm 1 The training procedure of our FOIL.

Require: Time-series dataset $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{N}$
Ensure: An optimized predictor $\rho(\phi(\cdot)) : \mathcal{X} \to \mathcal{Y}$
1: Initialize $\rho(\cdot)$, $\{\rho^{(e)}(\cdot)\}$, $\phi(\cdot)$
2: Random assign environment label for each $(\mathbf{X}_i, \mathbf{Y}_i)$.
3: **while** not converged **do**
4: **Stage 1: Time-series Invariant Learning:** Update $\phi(\cdot)$, $\rho(\cdot)$ according to Equation 10.
5: **Stage 2: Time-series Environment Inference:**
6: **M Step:** Fit models according to Equation 6, update $\{\rho^{(e)}\}$.
7: **E Step:** Reallocate environment labels according to Equation 7 and Equation 8.
8: **end while**
9: **return** $\rho(\cdot)$ and $\phi(\cdot)$.

B. Additional Experimental Details

B.1. Datasets

We conduct experiments on four real-world datasets, commonly used as benchmark datasets:

- **Exchange** dataset records the daily exchange rates of eight currencies.
- **ETTh1** and **ETTh2** datasets record the hourly electricity transformer temperature, comprising two years of data collected from two separate counties in China. They include seven variables. We omitted ETTm1 and ETTm2 as they share the same data source as ETTh1 and ETTh2, but with different sampling frequencies.
- **ILI** dataset collects data on influenza-like illness patients weekly, with eight variables. We mainly follow (Wu et al., 2022) to preprocess data, split datasets into train/validation/test sets and select the target variables. All datasets are preprocessed using the zero-mean normalization method.

B.2. Backbones

As aforementioned, our proposed FOIL is a model-agnostic framework. We select three different types of TSF models as backbones. **Informer** (Zhou et al., 2021) proposes an efficient transformer for long-term TSF. **Crossformer** (Zhang & Yan, 2022) better utilizes cross-dimension dependency, making it more sensitive to spuriouse correlations. **PatchTST** (Nie et al., 2022) employs channel-independent and patching strategies to achieve state-of-the-art performance.

B.3. Baselines: General OOD Methods

- **Methods with Environment Labels: IRM** (Arjovsky et al., 2019) introduces a penalty to learn invariant predictors across different environments. On the basis of the invariance principle of IRM, **IB-ERM** (Ahuja et al., 2021) incorporates the information bottleneck constraint. **VREx** (Krueger et al., 2021) propose a penalty on the variance of training risks between environments as a simple agent of risk extrapolation. **SD** (Pezeshki et al., 2021) proposes a regularization method aimed at decoupling feature learning dynamics to achieve better OOD generalization.**GroupDRO** (Sagawa et al., 2019), a regularizer for worst-case group generalization, often considered to have general OOD generalization capabilities.
- **Methods without Environment Labels: EIIL** (Creager et al., 2021) infers the most informative environments for downstream learning invariant predictors by maximizing the penalty in IRM.

We omit AdaRNN (Du et al., 2021) for not being model-agnostic; DIVERSITY (Lu et al., 2022; 2023), as it's specific to time series classification and detection tasks; and multi-view TSF methods (Kamarthi et al., 2022), which treat each covariate as one view and inflate the parameter count, leading to unfair comparison.

B.4. Implementation

For the backbones, we utilize implementations and hyperparameter settings from the Time Series Library[1]. For general OOD methods, we employ the implementations and tune hyperparameter suggested by DomainBed[2]. For TSF methods, we use the implementations and hyperparameter settings from their corresponding papers. We have added an MLP to the end of PatchTST to utilize covariates effectively. For our proposed framework FOIL, we also incorporate RevIN like PatchTST to address the issue of non-stationarity. We perform affine transformation on each dimension of the raw covariate through learnable weight variables to better find invariant features and improve out-of-distribution generalization capabilities.

[1] https://github.com/thuml/Time-Series-Library
[2] https://github.com/facebookresearch/DomainBed