CAPE: Covariate-Adjusted Pre-Training for Epidemic Time Series Forecasting

Zewen Liu
Department of Computer Science
Emory University
Atlanta, Georgia
zewen.liu@emory.edu

Juntong Ni
Department of Computer Science
Emory University
Atlanta, Georgia
juntong.ni@emory.edu

Max S. Y. Lau
Rollins School of Public Health
Emory University
Atlanta, Georgia
msy.lau@emory.edu

Wei Jin
Department of Computer Science
Emory University
Atlanta, Georgia
wei.jin@emory.edu

Abstract

Accurate forecasting of epidemic infection trajectories is crucial for safeguarding public health. However, limited data availability during emerging outbreaks and the complex interaction between environmental factors and disease dynamics present significant challenges for effective forecasting. In response, we introduce CAPE, a novel epidemic pre-training framework designed to harness extensive disease datasets from diverse regions and integrate environmental factors directly into the modeling process for more informed decision-making on downstream diseases. Based on a covariate adjustment framework, CAPE utilizes pre-training combined with hierarchical environment contrasting to identify universal patterns across diseases while estimating latent environmental influences. We have compiled a diverse collection of epidemic time series datasets and validated the effectiveness of CAPE under various evaluation scenarios, including full-shot, few-shot, zero-shot, cross-location, and cross-disease settings, where it outperforms the leading baseline by an average of 9.9% in full-shot and 14.3% in zero-shot settings. The code will be released upon acceptance.

1 Introduction

Infectious disease outbreaks consistently challenge public health systems, affecting both individual well-being and economic stability [1]. Effective management of these outbreaks hinges on accurate epidemic forecasting, which involves predicting future incidences like infection cases and hospitalizations [2, 3, 4]. Over the years, various models have been developed to address this need. These include mechanistic models like SIR [5] and statistical models like ARIMA [6, 7], as well as advanced machine learning methods such as LSTM and GRU [8], which have proven instrumental in forecasting disease spread and supporting informed public health decision-making.

Despite the advancements, current models are typically trained for specific diseases within particular geographic regions, limiting their ability to integrate insights from diverse sources spanning multiple pathogens and spatiotemporal contexts. This narrow focus can impede a comprehensive understanding of disease dynamics and the design of effective outbreak responses, especially during novel or emergent outbreaks when observations are typically scarce. Given the extensive and diverse outbreak data collected over decades and across various geographies, pre-training on such a broad dataset

Preprint. Under review.

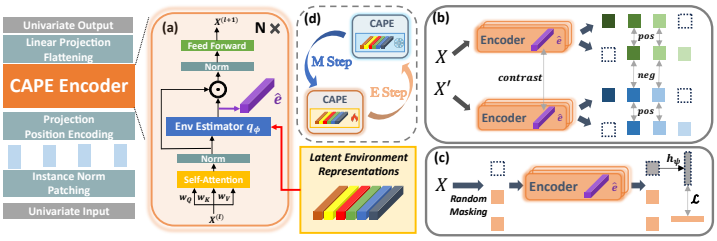

Figure 1: (a) CAPE encoder and environment estimator with latent representations; (b) Hierarchical environment contrasting for temporal and environment representations; (c) Random masking and reconstruction with environment estimation to capture universal patterns; (d) EM algorithm to iteratively optimize model parameters and environment representations.

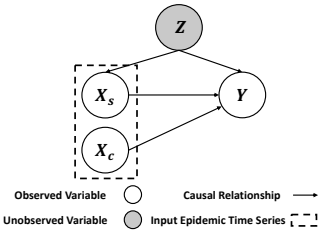

Figure 2: Structural causal model for epidemic foresting, where Z refers to the environment states, and X_s and X_c refer to the spurious and causal factors of the input.

could potentially enable the development of more generalizable models with greater applicability and adaptability across different pathogens and contexts. This raises an important question: *Can we leverage lessons from diverse historical disease time series to develop a generalized model that enhances epidemic forecasting accuracy?*

To address the above question, we draw inspiration from the success of large pre-trained transformer-based models [9] and develop a *pre-trained epidemic forecasting model* using extensive disease time series data to distill generalizable knowledge across pathogens and contexts. The pre-trained model can be subsequently fine-tuned for specific diseases or geographical regions. While it is possible to adapt general time series foundation models [10, 11] to epidemic forecasting, their pre-trained corpus mostly consists of non-epidemic data, which may not accurately capture epidemic dynamics and infection trajectories, potentially degrading forecasting accuracy. Although an early effort has been made in epidemic pre-training [12], it overlooks critical external factors such as temperature, elevation, and public health policies and interventions – factors are known to influence the dynamics of disease spread in space and time [13] – potentially yielding suboptimal performance. For instance, dengue infection spread may exhibit distinct dynamics in different geographical regions due to variations in temperature and humidity [14]. Without accounting for these external factors, models risk failing to capture their complex interplay with pathogens and producing inaccurate forecasts. Throughout this paper, we refer to these external factors as *environments*.

Nevertheless, the need to robustly and effectively account for the environment further intensifies the challenge of developing an epidemic pre-training framework that is generalizable across varying pathogens and contexts. A major obstacle is the shift in the temporal distribution of infection trajectories between training and test datasets, often driven by the changes in the environment. Insufficient consideration of such distribution shifts can obscure the relationship between historical infection data and future predictions (for a detailed discussion, see Appendix A.8), compromising a model's ability to make accurate forecasts. As such, it is crucial to disentangle the influence of changing environments from other more intrinsic factors (e.g., a pathogen's infection rate) affecting disease transmission dynamics. Yet, exact and explicit mechanisms by which the environment influences the disease dynamics of a particular pathogen are often not fully understood, which necessitates a sophisticated modeling approach to identify and separate these latent environmental influences.

Our Solution. To integrate insights from extensive historical diseases and effectively model environmental factors, we propose **C**ovariate-**A**djusted **P**retraining for **E**pidemic forecasting (**CAPE**) to capture the universal patterns of disease dynamics, as shown in Figure 1. Our approach addresses the challenges of optimizing the model with limited observations of a single disease infection trajectory and the complex influence of the environment by *combining a pre-training framework with explicit environment modeling*. Drawing on principles from causal analysis and covariate adjustment [15], CAPE aims to estimate the latent environments and control for their influences for epidemic forecasting. Specifically, during the pre-training phase, CAPE utilizes environment-aware self-supervised learning, including random masking (Figure 1(c)) and hierarchical environment contrasting (Figure 1(b)), to enhance its understanding of the disease dynamics and environmental influence. Furthermore, an environment estimator is introduced, which estimates dynamic environments based on latent

environment representations learned during pre-training using *Expectation-Maximization* algorithm. Our contributions can be summarized as follows:

- We propose a novel epidemic pre-training framework, namely CAPE, that learns representations of environments and performs covariate adjustment on the input epidemic time series data, which aims to disentangle the inherited disease dynamics from the environment.
- We assemble a diverse collection of epidemic time series datasets from various diseases and regions, serving as a crucial testbed for evaluating pre-trained epidemic forecasting models. This allows for extensive testing across multiple scenarios, including few-shot, zero-shot, cross-location, and cross-disease evaluations.
- We demonstrate the effectiveness of pre-training on epidemic datasets, showcasing superior performance across various downstream datasets and settings. Notably, CAPE surpasses the best baseline by an average of 9.9% in the full-shot setting and 18.1% in the zero-shot setting across all tested downstream datasets.
- We provide an in-depth analysis of how pre-training and environment estimation affect downstream performance and mitigate the impact of distribution shifts.

2 Related Work and Problem Definition

Epidemic Forecasting Models. Traditionally, epidemic forecasting employs models like ARIMA [6], SEIR [16], and VAR [17]. ARIMA predicts infections by analyzing past data and errors, SEIR models population transitions using differential equations, and VAR captures linear inter-dependencies by modeling each variable based on past values. Recently, deep learning models—categorized into RNN-based, MLP-based, and transformer-based—have surpassed these methods. RNN-based models like LSTM [18] and GRU [19] use gating mechanisms to manage information flow. MLP-based models use linear layers [20] or multi-layer perceptrons [21, 22] for efficient data-to-prediction mapping. Transformer-based models [23, 24, 25] apply self-attention to encode time series and generate predictions via a decoder. However, these models are limited in that they typically utilize data from only one type of disease without considering valuable insights and patterns from diverse disease datasets.

Pre-trained Time Series Models. To enhance performance and enable few-shot or zero-shot capabilities, transformer-based models often employ pre-training on large datasets, which typically use masked data reconstruction [26, 27] or promote alignment across different contexts [28, 29, 30]. For example, PatchTST [31] segments time series into patches, masks some, and reconstructs the masked segments. Larger foundational models like MOMENT [32] aim to excel in multiple tasks (e.g., forecasting, imputation, classification) but require substantial data and computational resources. In the epidemic context, Kamarthi et al. [12] pre-trained on various diseases, improving downstream performance and highlighting pre-training's potential in epidemic forecasting. Nevertheless, all these models overlook the influence of the environment, and zero-shot ability in epidemic forecasting, along with the factors affecting the pre-training process, remain unanswered. In this study, we introduce environment modeling and conduct a thorough analysis of these questions.

Problem Definition. In this study, we adopt a univariate setting: Given a historical time series input: $\mathbf{x} \in \mathbb{R}^{T\times 1}$, where T is the size of lookback window, the goal of epidemic forecasting is to map $\mathbf{x}$ into target trajectories (e.g. infection rates): $\mathbf{y} \in \mathbb{R}^h$, where h denotes the size of the forecast horizon. We define X and Y as the random variables of input $\mathbf{x}$ and target $\mathbf{y}$ respectively. During pre-training, a representation function $g_\theta : \mathbb{R}^{T\times 1} \to \mathbb{R}^{T\times d}$, where d denotes the dimension of the latent space and θ being the parameter of the model, extracts universal properties from a large collection of epidemic time series datasets $\mathcal{D}_{\text{pre}} = \{D'_1, D'_2, \ldots, D'_S\}$. Then, a set of self-supervised tasks $\mathcal{T}_{\text{pre}} = \{\mathcal{T}_i\}_{i=1}^R$ is defined, where each task $\mathcal{T}_i$ transforms a sample $\mathbf{x} \sim \mathcal{D}_{\text{pre}}$ into a pair of new input and label: $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$, and optimizes a loss $\mathcal{L}_{\mathcal{T}_i} = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{pre}}}[\ell_{\mathcal{T}_i}(h_\psi(g_\theta(\tilde{\mathbf{x}})), \tilde{\mathbf{y}})]$, with $\ell_{\mathcal{T}_i}$ being the task-specific metric and h_ψ the task-specific head.

3 Proposed Method

3.1 Model Design

3.1.1 Causal Analysis for Epidemic Forecasting

As environments influence both historical infection patterns and future disease spread, we draw inspiration from causal inference [33, 34] and introduce a Structural Causal Model where we treat the environment Z as a confounder that influences both the independent variable (e.g., historical data X) and the dependent variable (e.g., future infections Y). Furthermore, we adopt a causal decomposition approach [35] that separates X into two components (Figure 2): (1) a *spurious* factor X_s that is environment-dependent, and (2) a *causal* factor X_c that remains environment-independent. Both factors influence the target Y, with X_s reflecting the impact of environment Z. Since epidemic dynamics are driven by a finite set of critical factors, such as public health policies, we model Z with the following assumption:

Assumption 3.1. The environment variable Z follows a categorical distribution $p(Z)$ and takes on one of K discrete environmental states, denoted as z_k. Each state z_k is associated with a unique latent representation $\mathbf{e}_k \in \mathbb{R}^{h_e}$, capturing the unique features specific to that environment.

In constructing a predictive model for input $\mathbf{x}$, we define $\hat{Y}$ as the predicted time series $\hat{\mathbf{y}}$ and model the predictive distribution $p_\Theta(\hat{Y}|X)$ using $f_\Theta(\mathbf{x}) = h_\psi(g_\theta(\mathbf{x}))$, where $\Theta = \{\theta, \psi\}$. Training typically involves maximizing the log-likelihood of $p_\Theta(\hat{Y}|X)$, which in practice translates to minimizing the errors over the pre-training dataset $\mathcal{D}_{\text{pre}}$:

$$\Theta^* = \arg\min_\Theta -\frac{1}{|\mathcal{D}_{\text{pre}}|} \sum\nolimits_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_{\text{pre}}} \|\mathbf{y} - f_\Theta(\mathbf{x}))\|^2. \tag{1}$$

As the environment Z impacts the distribution of the observed data through $p(X, Y|Z) = p(X|Z)p(Y|X, Z)$, we formulate the following objective:

$$\Theta^* = \arg\min_\Theta \mathbb{E}_{p(Z)}[\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p(Y,X|Z)}[\|\mathbf{y} - f_\Theta(\mathbf{x}))\|^2]]. \tag{2}$$

The above equation suggests that the optimal Θ^* depends on the environment distribution $p(Z)$. If we simply maximize the likelihood $p_\Theta(\hat{Y}|X)$, the confounding effect of Z on X and Y will mislead the model to capture the shortcut predictive relation between the input and the target trajectories, which necessitates explicit modeling of the environment during pre-training. Given that input infection trajectories inherently reflect the influence of the environment, it is crucial to develop mechanisms that disentangle the correlations between infection trajectories and environmental factors.

In this study, we switch to optimize $p_\Theta(\hat{Y}|do(X))$, where the *do*-operation intervenes the variable X and removes the effects from other variables (i.e., Z in our case), thus effectively isolating the disease dynamics from environmental influences. In practice, this operation is usually conducted via covariate adjustment, particularly *backdoor adjustment* [36], which controls for the confounder and uncovers the true causal effects of interest. The theoretical foundation for this is explained through: $p(Y|do(X)) = \int p(Y|X, Z = z)p(Z = z)dz$ (see Appendix A.1). Under Assumption 3.1, this simplifies over different environmental states:

$$p(Y|do(X)) = \sum\nolimits_Z p(Y|X, Z = z)p(Z = z). \tag{3}$$

However, obtaining detailed environmental information, or $\mathbf{e}_k$, can be challenging due to variability in data availability and quality. To address this, we resort to a data-driven approach that treats $\mathbf{e}_k$ as learnable parameters and thus allows us to dynamically infer the environmental distribution directly from the observed data. Specifically, we implement an environment estimator $q_\phi(Z|X)$ that infers the probability of environment states based on historical inputs together with the latent representations of each state. Then, we derive a variational lower bound (see Appendix A.1):

$$\begin{aligned} &\log p_\Theta(\hat{Y}|do(X)) \geq \\ &\mathbb{E}_{q_\phi(Z|X)}\left[\log p_\Theta(\hat{Y}|X, Z)\right] - \mathrm{KL}\left(q_\phi(Z|X) \parallel p(Z)\right), \end{aligned} \tag{4}$$

where the first term maximizes the model's predictive power and the second term regularizes the environment estimator to output a distribution close to the prior distribution $p(Z)$.

3.1.2 Model Instantiation

To instantiate and train a model that performs the covariate adjustment, we need to model the environment estimator $q_\phi(Z|X)$ and the predictor $p_\Theta(\hat{Y}|X,Z)$.

Latent Environment Estimator $q_\phi(Z|X)$. We model $p(Z|X)$ using a latent environment estimator $q_\phi(Z|X)$. Since environmental influences vary over time, we apply patching [31] to manage granularity in environment estimation. This prevents overly specific or generalized estimations that could obscure key temporal fluctuations. We divide the input $\mathbf{x}$ into C non-overlapping patches, $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_C]$, where $\mathbf{x}_c \in \mathbb{R}^{T/C}$. Then, a self-attention layer f_{enc} captures temporal dependencies between patches, producing contextualized representations $\mathbf{h}_c^{(l)} = f_{\text{enc}}(\mathbf{x}_c^{(l)})$ for each patch at layer l. Subsequently, since the environment influences only the spurious component of the input, we introduce a transformation $\mathbf{W}_s^{(l)}$ to capture the spurious component of $\mathbf{h}_c^{(l)}$. Finally, we model $q_\phi(Z|X)$ as a cross-attention layer that captures the relation between each patch and the latent environment representations:

$$\pi_{k,c}^{(l)} = \text{Softmax}\left((\mathbf{W}_k^{(l)}\mathbf{e}_k)^\top \cdot (\mathbf{W}_s^{(l)}\mathbf{h}_c^{(l)})\right), \tag{5}$$

where $\pi_{k,c}^{(l)}$ is the output probability of the environment z_k for the c-th patch, and $\mathbf{W}_k^{(l)}$ is a transformation layer for $\mathbf{e}_k$. Such operation not only takes into account the contextualized representation of the current time period, but also considers the latent environment representations, which made it possible to infer the densities of other environment distributions with different latent representations.

Epidemic Predictor $p_\Theta(\hat{Y}|X,Z)$. Unlike previous studies, which do not explicitly model environment states, we incorporate these states directly into the input using their latent representations $\mathbf{e}_k$. Specifically, we model the predictor $p_\Theta(\hat{Y}|X,Z)$ by employing a weighted sum over the combined representations of each environment and the input using Hadamard product, i.e., $f_{\text{enc}}(\mathbf{x}_c^{(l)}) \odot \mathbf{e}_k$. Finally, we apply a feed-forward layer to compute the output representations, serving as the input for the next layer. Integrating these components, the CAPE encoder can be expressed as:

$$\mathbf{x}_c^{(l+1)} = \sigma\left(\mathbf{W}_f^{(l)} \sum_{k=1}^{K} \pi_{k,c}^{(l)} \left[f_{enc}(\mathbf{x}_c^{(l)}) \odot \mathbf{e}_k\right]\right), \tag{6}$$

where σ represents the activation function and $\mathbf{W}_f^{(l)}$ denotes the learnable parameters of the feedforward layer. Assuming L layers are stacked, we acquire the final representation $\mathbf{x}^{(L)} = [\mathbf{x}_1^{(L)}, \mathbf{x}_2^{(L)}, \ldots \mathbf{x}_C^{(L)}] = g_\theta(\mathbf{x}) \in \mathbb{R}^{C \cdot d}$ and apply a task-specific head to predict the target variable $\mathbf{y} = h_\psi(\mathbf{x}^{(L)})$, where h_ψ is a linear transformation.

3.2 Pre-training Objectives for Epidemic Forecasting

CAPE captures diverse epidemic time series dynamics through self-supervised learning tasks that identify universal patterns in the pre-training dataset. While previous studies neglected the confounding effects of environmental factors on input-label pairs in $\mathcal{T}_{\text{pre}}$, CAPE seamlessly integrates environment estimation into the self-supervised framework.

Random Masking with Environment Estimation. To capture features from large unlabeled epidemic time series data, we employ a masked time series modeling task [12, 32] (Figure 1(c)) that masks 30% of input patches. As depicted in Figure 2, the generation of X depends on the environment Z, indicating that accurate patch reconstruction requires capturing both temporal and environmental dependencies. Unlike prior studies that overlook the environment's role, we utilize an environment estimator $q_\phi(Z|X)$ to infer Z, aiding both reconstruction and estimator training. During pre-training, input $\mathbf{x}$ is transformed into masked input and label pairs $(\tilde{\mathbf{x}}, \mathbf{x})$, with the original time series serving as label y. The reconstruction $\hat{\mathbf{x}} = h_\psi(g_\theta(\tilde{\mathbf{x}}))$ is optimized using Mean Squared Error (MSE): $\mathcal{L}_{\text{recon}}(\mathbf{x}, \hat{\mathbf{x}}) = \text{MSE}(\hat{\mathbf{x}}, \mathbf{x})$.

Hierarchical Environment Contrasting. Two consecutive time series samples, $\mathbf{x}$ and $\mathbf{x}'$, can include overlapping regions when divided into multiple patches. These overlapping patches, although identical, can exhibit contextual variations influenced by their different adjacent patches. As indicated by Eq. (5), such variations can alter the latent patch-wise representations, leading to inconsistencies in the environmental estimates for the same patch across the samples. To ensure that each patch's

environment remains *context-invariant*, we propose a hierarchical environment contrasting scheme inspired by [30]. We define an *aggregated latent environment representation* $\hat{\mathbf{e}}_c^{(l)} = \sum_{k=1}^{K} \mathbf{e}_k \pi_{k,c}^{(l)}$ to represent the weighted environment states for the c-th patch. For contrastive loss computation, we use the combined representation $\hat{\mathbf{E}}_{j,c}^{(l)} = \sigma(\mathbf{W}_f^{(l)}(\hat{\mathbf{e}}_c^{(l)} \odot \mathbf{h}_c^{(l)}))$ for c-th patch of sample j. Additionally, $\hat{\mathbf{E}'}_{j,c}^{(l)}$ denotes the representation in the context of $\mathbf{x}'$. Finally, we compute a patch-wise contrastive loss:

$$\begin{aligned}\mathcal{L}_{\text{CL}}(j,c) &= -\hat{\mathbf{E}}_{(j,c)} \cdot \hat{\mathbf{E}'}_{(j,c)} \\ &+ \log\left(\sum_{b\in B} \exp\left(\hat{\mathbf{E}}_{(j,c)} \cdot \hat{\mathbf{E}'}_{(b,c)}\right) + \mathbb{I}_{j\neq b} \exp\left(\hat{\mathbf{E}}_{(j,c)} \cdot \hat{\mathbf{E}}_{(b,c)}\right)\right) \\ &+ \log\left(\sum_{t\in\Omega} \exp\left(\hat{\mathbf{E}}_{(j,c)} \cdot \hat{\mathbf{E}'}_{(j,t)}\right) + \mathbb{I}_{c\neq t} \exp\left(\hat{\mathbf{E}}_{(j,c)} \cdot \hat{\mathbf{E}}_{(j,t)}\right)\right).\end{aligned}$$

where B is the batch size, Ω denotes the overlapping patches, and $\mathbb{I}$ is the indicator function. The above equation contains three key terms: (1) The first term encourages the representations of the same patch from two different contexts to be similar, which preserves the context-invariant nature of environments. (2) The second term (*Instance-wise Contrasting*) treats $\hat{\mathbf{e}}_c^{(l)}$ from different samples in the batch as negative pairs, which promotes dissimilar representations, and enhances diversity among instances. (3) The third term (*Temporal Contrasting*) treats the representations of different patches from overlapping regions (Ω) as negative pairs, which encourages differences across temporal contexts.

Pre-Training Loss. Given a batch of B samples $\mathbf{X} \in \mathbb{R}^{B\times T}$, we combine the reconstruction loss and the contrastive loss, yielding the final loss function for pre-training:

$$\mathcal{L}_{\text{final}} = \sum\nolimits_{\mathbf{x}\in\mathbf{X}} \mathcal{L}_{\text{recon}}(\mathbf{x}, \hat{\mathbf{x}}) + \alpha\, \mathcal{L}_{\text{CL}}(\hat{\mathbf{E}}^{(L)}, \hat{\mathbf{E}}'^{(L)}),\ \mathbf{X} \sim \mathcal{D}_{\text{pre}}$$

where L is the number of layers, and α is the hyperparameter used to balance the contrastive loss and the reconstruction loss. Further analysis can be found in Appendix A.10.

3.3 Optimization of the CAPE Framework

To effectively maximize the variational lower bound in Eq. (4), we employ the *Expectation-Maximization* (EM) algorithm to iteratively update the latent environments and epidemic predictor. The pseudo algorithm for the optimization procedure is provided in Appendix A.3.

E-Step: Estimating Latent Environments. In the E-step, we aim to identify the environment states Z and the corresponding distribution $p(Z)$ that result in the target distribution $p(Y)$. This involves maximizing the expected likelihood of $p(Y|Z)$ given $p(Z)$. We freeze the epidemic predictor $p_\Theta(\hat{Y}|X,Z)$ and the environment estimator $q_\phi(Z|X)$, treating them as oracles, which means $p_\Theta(\hat{Y}|X,Z) = p(Y|X,Z)$ and $q_\phi(Z|X) = q(Z|X)$. While actively updating the environment representations $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, ...\mathbf{e}_k]$, the optimization of the environment states Z is learned through maximizing $\mathbb{E}_{p(Z)}[p(Y|Z)] = \mathbb{E}_{p(X)}[\mathbb{E}_{q_\phi(Z|X)} p_\Theta(Y|X,Z)]$, which is equivalent to minimizing the expected reconstruction loss:

$$\mathbf{E}^{t+1} \leftarrow \arg\min_{\mathbf{E}} \left[\mathbb{E}_{\mathbf{x}\sim p(X)}[\mathcal{L}_{\text{recon}}(\mathbf{x}, \hat{\mathbf{x}})]\right]. \tag{7}$$

We use subscript t to denote the pre-update distribution and derive the updated distribution $p^{t+1}(Z)$ as $q_{\phi_t}^{t+1}(Z)$, along with the updated environment representations $\mathbf{E}^{t+1}$.

M-Step: Optimizing Epidemic Predictor. In the M-step, we aim to optimize the epidemic predictor by maximizing its predictive power and regularizing the environment distribution. During this step, the environment representations $\mathbf{E}^{t+1}$ are held fixed. We have the following theorem:

Theorem 3.2. *Assuming $q_{\phi_t}^{t+1}(Z) = p^{t+1}(Z)$ and an L2 norm is applied on ϕ, the variational lower bound in Eq. (4) can be approximated as follows:*

$$\mathbb{E}_{p(X)}\left[\mathbb{E}_{q_{\phi_t}^{t+1}(Z|X)}\left[\log p_{\Theta_{t+1}}^{t+1}(\hat{Y}|X,Z)\right]\right] - C, \tag{8}$$

which is equivalent to minimizing the expected reconstruction loss $\mathbb{E}_{\mathbf{x}\sim p(X)}[\mathcal{L}_{recon}(\mathbf{x}, \hat{\mathbf{x}})]$.

Table 1: Univariate forecasting results with horizons ranging from 1 to 16 future steps. The lookback window length is set to 36 and all models are evaluated using MSE. Note that performance rankings are distinguished by *color coding: Best, Second Best, Third Best*. $\Delta(\%)$ stands for the relative improvement of CAPE over the baselines in terms of average MSE over all horizons.

Dataset	Horizon	*Statistical Model*	*RNN-Based*		*MLP-Based*	*Transformer-Based*						CAPE
						Non-Pre-trained			*Pre-trained*			
		ARIMA	LSTM	GRU	Dlinear	Informer	Autoformer	Fedformer	PEM	MOMENT	PatchTST	
ILI USA	1	0.138	0.338	0.259	0.220	0.175	0.457	0.368	0.179	0.269	0.195	0.155
	2	0.203	0.377	0.301	0.247	0.370	0.710	0.380	0.226	0.321	0.264	0.200
	4	0.354	0.458	0.386	0.376	0.517	0.670	0.433	0.304	0.397	0.385	0.270
	8	0.701	0.579	0.529	0.506	0.597	0.842	0.570	0.538	0.510	0.535	0.404
	16	1.121	0.691	0.626	0.617	0.812	0.835	0.701	0.570	0.610	0.485	0.516
	Avg	0.503	0.489	0.420	0.393	0.494	0.703	0.490	0.363	0.421	0.373	**0.309**
	Δ (%)	38.57%	36.81%	26.43%	21.37%	37.45%	56.05%	36.94%	14.88%	26.60%	17.16%	-
ILI Japan	1	0.358	1.426	1.213	1.016	0.405	0.515	0.525	0.470	0.325	0.413	0.290
	2	0.772	1.635	1.458	1.294	0.666	0.855	1.151	0.755	0.586	0.698	0.535
	4	1.720	1.975	1.870	1.758	1.234	1.150	1.455	1.207	1.082	1.147	0.944
	8	2.981	2.373	2.365	2.285	1.688	1.866	2.012	1.810	1.706	1.708	1.650
	16	2.572	2.023	2.010	2.007	1.551	2.654	4.027	1.766	2.054	1.688	1.911
	Avg	1.680	1.886	1.783	1.672	1.109	1.408	1.834	1.202	1.151	1.131	**1.066**
	Δ (%)	36.55%	43.48%	40.21%	36.24%	3.88%	24.29%	41.88%	11.31%	7.38%	5.74%	-
Measles	1	0.071	0.182	0.143	0.133	0.066	0.203	0.321	0.085	0.113	0.094	0.083
	2	0.120	0.223	0.176	0.184	0.153	0.257	0.817	0.128	0.138	0.127	0.112
	4	0.225	0.310	0.258	0.296	0.288	0.331	0.226	0.213	0.186	0.205	0.161
	8	0.483	0.567	0.471	0.512	0.501	0.671	0.403	0.417	0.351	0.377	0.310
	16	1.052	1.110	1.013	1.088	0.904	1.115	0.754	0.806	0.818	0.722	0.752
	Avg	0.390	0.478	0.412	0.443	0.382	0.515	0.504	0.330	0.321	0.305	**0.269**
	Δ (%)	31.03%	43.72%	34.71%	39.28%	29.58%	47.77%	46.63%	18.49%	16.20%	11.80%	-
Dengue	1	0.244	0.250	0.261	0.224	0.255	0.525	0.521	0.225	0.420	0.240	0.223
	2	0.373	0.343	0.343	0.316	0.450	0.807	0.670	0.314	0.579	0.334	0.302
	4	0.696	0.564	0.579	0.560	0.798	0.957	0.766	0.571	0.661	0.586	0.561
	8	1.732	1.168	1.183	1.256	1.239	1.684	1.539	1.223	1.308	1.292	1.046
	16	4.082	3.876	3.315	3.109	2.659	3.364	2.934	3.376	2.532	2.537	2.509
	Avg	1.426	1.240	1.136	1.093	1.080	1.467	1.286	1.142	1.100	1.000	**0.892**
	Δ (%)	37.45%	28.06%	21.48%	18.39%	17.41%	39.20%	30.64%	21.89%	18.91%	10.80%	-
Covid	1	33.780	22.592	22.009	23.811	34.161	42.049	28.130	25.088	32.376	23.645	21.548
	2	33.193	23.460	22.542	24.809	24.883	30.631	28.059	23.123	35.418	25.047	22.224
	4	32.482	24.729	24.816	26.345	31.328	41.029	29.432	23.889	36.251	24.224	22.476
	8	36.573	31.019	33.934	33.081	35.964	55.812	41.791	31.217	40.429	31.548	28.403
	16	42.910	43.820	41.432	47.561	50.244	47.993	69.976	51.265	52.590	43.309	40.555
	Avg	35.787	29.124	28.947	31.121	35.316	43.503	39.478	30.917	39.413	29.555	**26.559**
	Δ (%)	25.79%	8.81%	8.25%	14.66%	24.80%	38.95%	32.72%	14.10%	32.61%	10.14%	-

Table 2: Few-shot learning results with horizons ranging from 1 to 16 future steps. The length of the lookback window is set to 36. Each model is evaluated after being trained on 20%, 40%, 60%, and 80% of the full training data. $\Delta(\%)$ stands for the relative improvement of the model after training with 20% more data in terms of average MSE over all horizons. The full result is shown in Appendix A.5.

Dataset/Model	CAPE					PatchTST					Dlinear					MOMENT					PEM				
	20%	40%	60%	80%	100%	20%	40%	60%	80%	100%	20%	40%	60%	80%	100%	20%	40%	60%	80%	100%	20%	40%	60%	80%	100%
ILI USA	**2.121**	**1.400**	**0.760**	**0.369**	**0.309**	**2.114**	**1.219**	**0.677**	**0.401**	**0.373**	**2.822**	**1.594**	**0.816**	**0.412**	**0.346**	**3.990**	**1.847**	**0.913**	**0.459**	**0.381**	**2.143**	**1.261**	**0.681**	**0.419**	**0.353**
Δ(%)	-	33.99%	45.71%	51.45%	16.26%	-	42.34%	44.45%	40.77%	6.98%	-	43.53%	48.78%	49.51%	16.02%	-	53.69%	50.58%	49.72%	17.00%	-	41.13%	46.00%	38.33%	15.76%
Dengue	**13.335**	**6.386**	**2.356**	**1.511**	**0.892**	**13.712**	**7.304**	**2.771**	**1.678**	**0.984**	**15.828**	**8.420**	**2.850**	**1.748**	**1.080**	**15.697**	**7.536**	**2.816**	**1.733**	**1.358**	**12.90**	**7.055**	**2.745**	**1.707**	**0.964**
Δ(%)	-	52.07%	63.12%	35.87%	40.95%	-	46.72%	62.06%	39.43%	41.39%	-	46.81%	66.15%	38.64%	38.19%	-	52.00%	62.63%	38.45%	21.65%	-	45.32%	61.09%	37.79%	43.51%
Measles	**0.483**	**0.600**	**0.381**	**0.285**	**0.269**	**0.863**	**0.834**	**0.448**	**0.359**	**0.306**	**1.194**	**1.130**	**0.602**	**0.478**	**0.394**	**1.661**	**0.915**	**0.425**	**0.471**	**0.500**	**0.670**	**0.896**	**0.430**	**0.364**	**0.306**
Δ(%)	-	-24.22%	36.50%	25.20%	5.61%	-	3.36%	46.25%	19.91%	14.81%	-	5.36%	46.64%	20.63%	17.58%	-	44.91%	53.55%	-10.59%	-6.16%	-	-33.87%	51.91%	15.35%	15.93%

The detailed proof can be found in Appendix A.1. Theorem 3.2 indicates that the optimization of the model's predictive ability can be approximated by Eq. (8), which corresponds to the expectation of $\mathcal{L}_{recon}$. To further enhance robustness, the contrastive loss is combined to regularize the environment estimator. Therefore, the overall optimization objective becomes minimizing the final pre-training loss:

$$\Theta_{t+1} \leftarrow \arg\min_{\Theta} \left[\mathcal{L}_{\text{final}}(\mathbf{X}, \hat{\mathbf{X}}, \mathbf{E}^{t+1}) \right]. \tag{9}$$

4 Experiment

4.1 Experiment Setup

Datasets. For pre-training CAPE, PatchTST, and PEM, we manually collected 17 distinct weekly-sampled diseases from Project Tycho [37]. For evaluation, we utilize five downstream datasets covering various diseases and locations: ILI USA [38], ILI Japan [39], COVID-19 USA [40], Measles England [41], and Dengue across countries [42]. Additionally, RSV [43] and Monkey Pox [44] infections in the US are used to test zero-shot performance. More details can be found in Appendix A.2.

Table 3: Zero-shot performance with a lookback window length of 12. All results are averaged over 4 weeks or days in the future. $\Delta(\%)$ stands for the relative improvement of CAPE over the baselines.

Dataset	Δ (%)	CAPE	PatchTST	PEM	MOMENT
ILI USA	9.26%	**0.147**	**0.164**	**0.162**	0.549
ILI Japan	17.06%	**0.705**	**0.907**	**0.850**	2.062
Measles	3.97%	**0.145**	**0.167**	**0.159**	0.533
Monkey Pox	20.00%	**0.0004**	**0.0005**	**0.0005**	0.0013
Dengue (mixed)	10.17%	**0.371**	**0.427**	**0.413**	1.624
RSV	26.06%	**0.834**	**1.128**	**1.260**	1.849
Covid (daily interval)	13.80%	**5.173**	**6.001**	**6.320**	18.881

Baselines. For baselines, we leverage the models from the comprehensive toolkit *EpiLearn* [45]. To provide a comprehensive evaluation, we compare CAPE with two sets of models: *non-pretrained* and *pre-trained*. Non-pretrained models include statistical methods like ARIMA [46], RNN-based [18, 19] approaches such as LSTM and GRU, the linear model DLinear [20], and transformer-based methods [23, 24, 25]. For pre-trained models, we evaluate popular approaches including PatchTST [31], PEM [12], and a time series foundation model MOMENT [32]. More experimental details can be found in Appendix A.3.

4.2 Baseline Comparison

We now evaluate the CAPE model under three settings: *fine-tuning*, *few-shot fine-tuning*, and *zero-shot forecasting*.

4.2.1 Fine-Tuning (Full-Shot Setting)

For non-pre-trained models, we train the entire model on the training split, while for pre-trained models, we fine-tune on downstream datasets by transferring the task-specific head h_ψ from pre-training to the forecasting task. We evaluate short-term and long-term performance by reporting MSE across horizons from 1 to 16. From Table 1, we observe: (a) CAPE achieves the best average MSE across all downstream datasets. It outperforms the best baseline by 9.91% on average and up to 14.85%. On the COVID dataset, CAPE performs best across all horizons, showing effectiveness on novel diseases. (b) Models like PEM, PatchTST, and MOMENT consistently rank second or third on 4 out of 5 downstream datasets. The best pre-trained model (excluding CAPE) outperforms the best non-pre-trained model by 6.223% on average. Among them, PatchTST has the highest average performance, surpassing PEM by 5.51% and MOMENT by 10.45%. Additionally, PEM outperforms MOMENT by 4.86%, indicating the importance of epidemic-specific pre-training. (c) Informer consistently outperforms Autoformer and Fedformer by 24.40% and 17.90% respectively, due to its sparse attention mechanism that reduces overfitting. Informer also surpasses Dlinear by 1.90%, suggesting that careful selection of model size and parameters is crucial for optimal performance. (d) Furthermore, environment modeling proves valuable, as CAPE consistently outperforms PatchTST, which shares a similar design. While both models are pre-trained on the epidemic-specific datasets, CAPE surpasses PatchTST by 11.13%.

4.2.2 Few-Shot and Zero-Shot Performance

Few-Shot Forecasting. In real-world scenarios, predicting outbreaks of diseases unknown or in new locations is challenging for purely data-driven models due to limited initial data. Thus, few-shot or zero-shot forecasting capabilities are essential for epidemic models. To simulate a few-shot scenario, we reduce the original training data from 100% to [20%, 40%, 60%, 80%]. We report the average MSE across 1 to 16 time steps. From Table 8, we make the following observations: (a) With an increasing volume of training materials, the performance of all models consistently improves. (b) CAPE achieves the best performance in most scenarios, demonstrating the superior few-shot ability. (c) Compared with models pre-trained on epidemic-specific datasets, Dlinear failed to achieve better performance when only 20% of training data is available. However, Dlinear is able to outperform MOMENT on ILI USA and Measles datasets when both models are trained or fine-tuned using 20% training data, which indicates the importance of pre-training. (d) Though CAPE achieves the best average performance on the ILI USA dataset when the training material is reduced, it achieves a good performance in short-term forecasting from 1 to 4 weeks (see Appendix A.5).

Table 4: Ablation study of removing components from CAPE.

Dataset	Model	H=1	H=2	H=4	H=8	H=16	Avg
ILI USA	CAPE	0.155	0.200	0.270	0.404	0.516	**0.309**
	w/o Env	0.326	0.448	0.508	0.642	0.735	0.532
	w/o Contrast	0.174	0.241	0.335	0.492	0.570	0.363
	w/o Pretrain	0.158	0.202	0.283	0.408	0.545	0.319
Measles	CAPE	0.069	0.096	0.155	0.280	0.743	**0.269**
	w/o Env	0.083	0.111	0.168	0.407	0.755	0.304
	w/o Contrast	0.090	0.124	0.276	0.431	0.801	0.344
	w/o Pretrain	0.074	0.113	0.223	0.402	0.816	0.326
Dengue	CAPE	0.218	0.301	0.540	1.193	2.210	**0.892**
	w/o Env	0.232	0.316	0.484	1.089	3.622	1.149
	w/o Contrast	0.198	0.273	0.460	1.128	3.329	1.078
	w/o Pretrain	0.210	0.276	0.449	1.115	3.759	1.162

Zero-Shot Forecasting. To further demonstrate the potential of our model, we evaluate CAPE in a zero-shot setting. Specifically, for transformer-based models, we retain the pre-training head and freeze all parameters during testing. All models are provided with a short input sequence of 12 time steps and tasked with predicting infections for the next 4 time steps. From Table 3, we make the following observations: (a) CAPE outperforms baselines across all downstream datasets, showing superior zero-shot forecasting ability. (b) Models pre-trained on epidemic-specific datasets achieve better performance compared to those pre-trained without epidemic-specific data (MOMENT). This indicates the necessity of choosing domain-specific materials for pre-training.

4.3 Ablation Study

We conducted an ablation study to assess CAPE's components (Table 4). Replacing environment estimators with non-disentangling self-attention layers consistently worsened performance across all datasets, notably increasing ILI USA's MSE from 0.309 to 0.532, underscoring the importance of environmental factors. Similarly, removing contrastive loss while retaining environment estimators raised Measles' MSE from 0.269 to 0.344, with smaller increases for ILI USA and Dengue. Training CAPE directly on downstream datasets without pre-training also decreased performance, with MSE rising to 0.319 (ILI USA), 0.326 (Measles), and 1.162 (Dengue), though less than removing environment estimation. These results indicate that all CAPE components are essential for optimal forecasting and that tailoring component emphasis to dataset characteristics can further enhance performance.

4.4 Transferability

Cross-Location. We include measles data from the USA in the pre-training dataset. To evaluate our model's ability to adapt to cross-region data, we incorporate measles outbreak data from the UK into the downstream datasets. As shown in Table 4, the pre-trained CAPE outperforms the non-pre-trained version by 17.48%. While we pre-train our model with influenza data from the USA, the zero-shot evaluation on the influenza outbreak in Japan also shows superior performance, underscoring the crucial role of pre-training in enabling generalization to novel regions.

Cross-Disease. While we include various types of diseases in our pre-training dataset, novel diseases including Dengue (non-respiratory) and COVID-19 that are unseen in the pre-training stage are incorporated during the downstream evaluation. The ability of our model to adapt to novel diseases is proven compared to the version not pre-trained on the Dengue dataset (Table 4), improving which by 23.24%, as well as the superior zero-shot performance on the COVID dataset (Table 3), which surpasses the MOMENT that is not pre-trained on other diseases by 72.60%.

Cross-Interval. While we only pre-train using weekly-sampled data, our model outperformed the non-pre-trained version on the irregularly sampled Dengue dataset, demonstrating robustness to different time intervals. Additionally, on the daily-sampled COVID-19 dataset, our model maintained

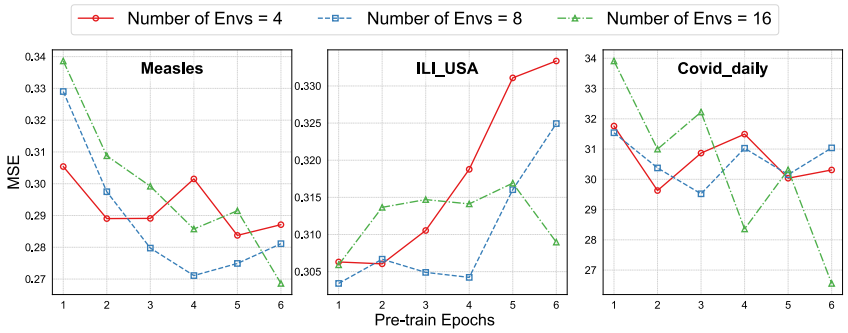

Figure 3: Downstream performance with different numbers of environments and pre-training epochs.

strong zero-shot performance, further illustrating its ability to generalize across varying temporal resolutions.

4.5 Deeper Analysis

Impact of Pre-Training Epochs. Evaluating four downstream datasets (Figure 3), we find that increasing pre-training epochs consistently improves performance on Measles and COVID datasets but degrades it for ILI USA. Additionally, models with more environment states K perform better as pre-training epochs increase.

Impact of Pre-Training Materials. We examine potential biases in our pre-training dataset by splitting it into respiratory and non-respiratory diseases. As shown in Figure 4, with similar volumes of pre-training data, the model performs better on downstream datasets when their disease types align with the pre-training data (e.g., respiratory diseases). However, the size of the pre-training material has a more significant impact on downstream performance.

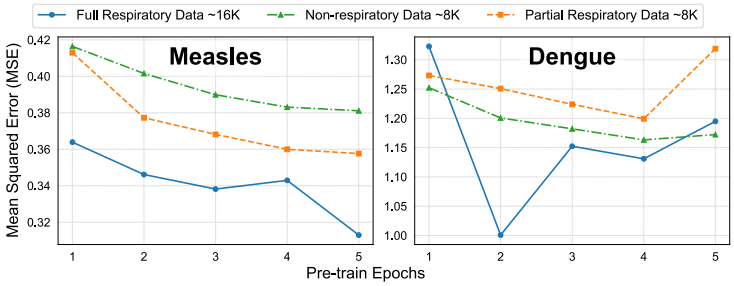

Figure 4: Downstream performance variation when the model is pre-trained with either respiratory or non-respiratory diseases only.

Impact of Pre-Training Material Scale. To explore how the pre-training material scale affects downstream performance, we scaled the original pre-training dataset and test on downstream datasets. As shown in Figure 5, a sudden performance boost is observed at around a 60% reduction for both Measles and Dengue datasets.

Tackling Distribution Shift. In this study, distribution shifts refer to changes in infection patterns observed from the training set to the test set. To evaluate distribution shifts, we compute the Central Moment Discrepancy (CMD) score [47] between training and test distributions for each disease (see Appendix A.8). Figure 6 shows that our model with environment estimation achieves the lowest CMD score, demonstrating its effectiveness in mitigating the impact of temporal distribution shifts.

Disentangling Disease Dynamics. We validate our model's ability to capture intrinsic disease dynamics by extracting latent embeddings from various datasets and computing the Davies-Bouldin Index (DBI) for each pair. As shown in Figure 7, CAPE consistently achieves lower DBI scores than

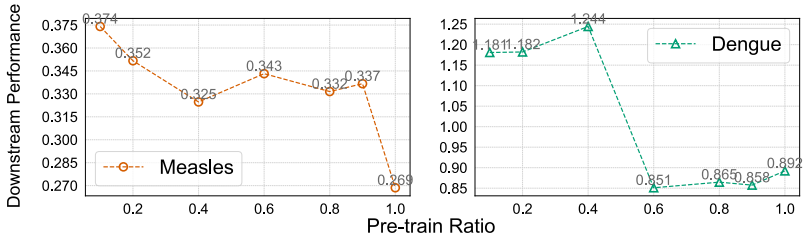

Figure 5: Downstream performance across pre-training ratios.

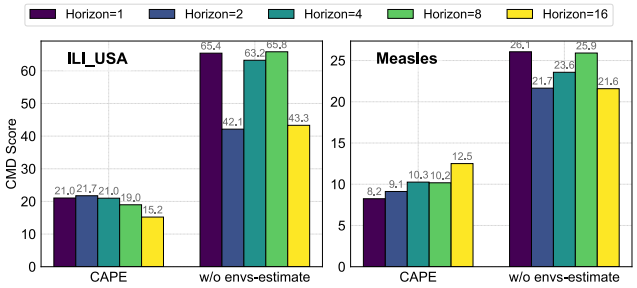

Figure 6: We report the CMD scores of the embeddings produced by CAPE with and without environment estimation, which quantify distributional differences between the training and test sets.

PatchTST across all pairs, demonstrating its superior effectiveness in distinguishing diseases and separating disease-specific patterns from environmental influences.

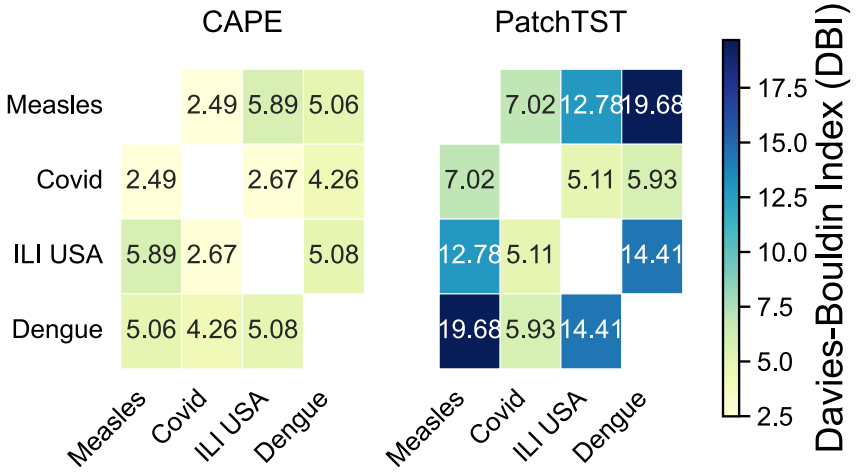

Figure 7: Davies-Bouldin Index score between the embeddings of each pair of downstream datasets, output by the pre-trained model without fine-tuning. A visualization is shown in Appendix A.9.

5 Conclusion

We present Covariate-Adjusted Pre-Training for Epidemic time series forecasting, showcasing the benefits of pre-training and environment modeling. While leveraging pre-training materials, CAPE explicitly learns latent representations of the environment and performs backdoor adjustment. Extensive experiments validate CAPE's effectiveness in various settings, including few-shot and zero-shot.

References

[1] Maria Nicola, Zaid Alsafi, Catrin Sohrabi, Ahmed Kerwan, Ahmed Al-Jabir, Christos Iosifidis, Maliha Agha, and Riaz Agha. The socio-economic implications of the coronavirus pandemic (covid-19): A review. *International journal of surgery*, 78:185–193, 2020.

[2] Zewen Liu, Guancheng Wan, B Aditya Prakash, Max SY Lau, and Wei Jin. A review of graph neural networks in epidemic modeling. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6577–6587, 2024.

[3] Guancheng Wan, Zewen Liu, Max SY Lau, B Aditya Prakash, and Wei Jin. Epidemiology-aware neural ode with continuous disease transmission graph. *arXiv preprint arXiv:2410.00049*, 2024.

[4] Bijaya Adhikari, Xinfeng Xu, Naren Ramakrishnan, and B Aditya Prakash. Epideep: Exploiting embeddings for epidemic forecasting. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 577–586, 2019.

[5] Ian Cooper, Argha Mondal, and Chris G Antonopoulos. A sir model assumption for the spread of covid-19 in different communities. *Chaos, Solitons & Fractals*, 139:110057, 2020.

[6] Alok Kumar Sahai, Namita Rath, Vishal Sood, and Manvendra Pratap Singh. Arima modelling & forecasting of covid-19 in top five affected countries. *Diabetes & metabolic syndrome: clinical research & reviews*, 14(5):1419–1427, 2020.

[7] Vaia I Kontopoulou, Athanasios D Panagopoulos, Ioannis Kakkos, and George K Matsopoulos. A review of arima vs. machine learning approaches for time series forecasting in data driven networks. *Future Internet*, 15(8):255, 2023.

[8] Farah Shahid, Aneela Zameer, and Muhammad Muneeb. Predictions for covid-19 with deep learning models of lstm, gru and bi-lstm. *Chaos, Solitons & Fractals*, 140:110212, 2020.

[9] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[10] Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6555–6565, 2024.

[11] Qianli Ma, Zhen Liu, Zhenjing Zheng, Ziyang Huang, Siying Zhu, Zhongzhong Yu, and James T Kwok. A survey on time-series pre-trained models. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[12] Harshavardhan Kamarthi and B Aditya Prakash. Pems: Pre-trained epidmic time-series models. *arXiv preprint arXiv:2311.07841*, 2023.

[13] Max SY Lau, Bryan Grenfell, Michael Thomas, Michael Bryan, Kristin Nelson, and Ben Lopman. Characterizing superspreading events and age-specific infectiousness of sars-cov-2 transmission in georgia, usa. *Proceedings of the National Academy of Sciences*, 117(36):22430–22435, 2020.

[14] Szu-Chieh Chen and Meng-Huan Hsieh. Modeling the transmission dynamics of dengue fever: implications of temperature effects. *Science of the total environment*, 431:385–391, 2012.

[15] Jakob Runge, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 4(7):487–505, 2023.

[16] Shaobo He, Yuexi Peng, and Kehui Sun. Seir modeling of the covid-19 and its dynamics. *Nonlinear dynamics*, 101:1667–1680, 2020.

[17] Aaron C Shang, Kristen E Galow, and Gary G Galow. Regional forecasting of covid-19 caseload by non-parametric regression: a var epidemiological model. *AIMS public health*, 8(1):124, 2021.

[18] Peipei Wang, Xinqi Zheng, Gang Ai, Dongya Liu, and Bangren Zhu. Time series prediction for the epidemic trends of covid-19 using the improved lstm deep learning method: Case studies in russia, peru and iran. *Chaos, Solitons & Fractals*, 140:110214, 2020.

[19] Sathish Natarajan, Mohit Kumar, Sai Kiran Kumar Gadde, and Vijay Venugopal. Outbreak prediction of covid-19 using recurrent neural network with gated recurrent units. *Materials Today: Proceedings*, 80:3433–3437, 2023.

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

[21] Pedro Henrique Borghi, Oleksandr Zakordonets, and João Paulo Teixeira. A covid-19 time series forecasting model based on mlp ann. *Procedia Computer Science*, 181:940–947, 2021.

[22] Wyatt G Madden, Wei Jin, Benjamin Lopman, Andreas Zufle, Benjamin Dalziel, C Jessica E. Metcalf, Bryan T Grenfell, and Max SY Lau. Deep neural networks for endemic measles dynamics: Comparative analysis and integration with mechanistic models. *PLOS Computational Biology*, 20(11):e1012616, 2024.

[23] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

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

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

[26] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 2114–2124, 2021.

[27] Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Biloš, Hena Ghonia, Nadhir Hassen, Anderson Schneider, et al. Lag-llama: Towards foundation models for time series forecasting. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023.

[28] Archibald Fraikin, Adrien Bennetot, and Stéphanie Allassonnière. T-rep: Representation learning for time series using time-embeddings. *arXiv preprint arXiv:2310.04486*, 2023.

[29] Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in Neural Information Processing Systems*, 35:3988–4003, 2022.

[30] Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8980–8987, 2022.

[31] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

[32] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

[33] Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4227–4241, 2023.

[34] Licheng Jiao, Yuhan Wang, Xu Liu, Lingling Li, Fang Liu, Wenping Ma, Yuwei Guo, Puhua Chen, Shuyuan Yang, and Biao Hou. Causal inference meets deep learning: A comprehensive survey. *Research*, 7:0467, 2024.

[35] Chengzhi Mao, Kevin Xia, James Wang, Hao Wang, Junfeng Yang, Elias Bareinboim, and Carl Vondrick. Causal transportability for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7521–7531, 2022.

[36] Shiliang Sun et al. Caudits: Causal disentangled domain adaptation of multivariate time series. In *Forty-first International Conference on Machine Learning*.

[37] Willem G van Panhuis, Anne Cross, and Donald S Burke. Project tycho 2.0: a repository to improve the integration and reuse of data for global population health. *Journal of the American Medical Informatics Association*, 25(12):1608–1617, 2018.

[38] Centers for Disease Control and Prevention. Influenza-like illness (ili) data - usa. https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html, 2023.

[39] National Institute of Infectious Diseases. Infectious diseases weekly report (idwr) - japan. https://www.niid.go.jp/niid/en/idwr-e.html, 2023.

[40] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

[41] Max SY Lau, Alexander D Becker, Hannah M Korevaar, Quentin Caudron, Darren J Shaw, C Jessica E Metcalf, Ottar N Bjørnstad, and Bryan T Grenfell. A competing-risks model explains hierarchical spatial coupling of measles epidemics en route to national elimination. *Nature Ecology & Evolution*, 4(7):934–939, 2020.

[42] OpenDengue. Dengue data across countries. https://opendengue.org/, 2023.

[43] Centers for Disease Control and Prevention. Rsv surveillance data. https://www.cdc.gov/rsv/php/surveillance/rsv-net.html, 2023.

[44] Centers for Disease Control and Prevention. Monkey pox cases data. https://www.cdc.gov/mpox/data-research/cases/index.html, 2023.

[45] Zewen Liu, Yunxiao Li, Mingyang Wei, Guancheng Wan, Max SY Lau, and Wei Jin. Epilearn: A python library for machine learning in epidemic modeling. *arXiv preprint arXiv:2406.06016*, 2024.

[46] George Panagopoulos, Giannis Nikolentzos, and Michalis Vazirgiannis. Transfer graph neural networks for pandemic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4838–4845, 2021.

[47] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (cmd) for domain-invariant representation learning. *arXiv preprint arXiv:1702.08811*, 2017.

[48] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

[49] Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark. 2024.

[50] Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark. *arXiv preprint arXiv:2407.13278*, 2024.