HierarNet: Independent Interactive Hierarchical Disease Outbreak Forecasting

Zichi Zhang*, Phi Hung Nguyen*, Ngoc Phu Doan*, Viet-Hung Tran*, Xuan Hoang Nguyen, Hui Wang, Hans Vandierendonck, Son Thai Mai*

Queen's University Belfast, UK
{zzhang54, h.wang, h.vandierendonck, thaison.mai}@qub.ac.uk

Abstract

Early warning systems for disease outbreaks play a crucial role in public health for management and contingency planning. However, most predictive modeling works focus on flat models that incorporate exogenous inputs (e.g. climate, demographics) to predict future outbreaks at different locations, but do not jointly model multiple spatial aggregation levels. In this paper, we introduce *HierarNet*, a unique *independent-interactive hierarchical* forecasting framework that aims to predict disease outbreaks at different levels of spatial resolution, such as provinces, regions, and nations. HierarNet consists of two main phases. In the *local* phase, we train *independent* forecasting models for all locations at all levels. In the *global* phase, all models *iteratively interact* with others across different levels via their hierarchical relationships under an *ensemble* fashion to maximize their agreements. This *global local hierarchical interactive* scheme makes HierarNet a highly effective and *flexible* method (i.e. it can work with an arbitrary base prediction model and available exogenous data for each location independently). Extensive experiments are conducted on various disease datasets (e.g., Dengue fever, flu, diarrhea, and Bluetongue) in different countries (e.g., France, Vietnam, and USA) to show the performance of HierarNet compared to 19 state-of-the-art (SOTA) methods such as MinT, DYCHEM, WITRAN, SegRNN, TSMixer, PatchTST, or iTransformer. We also illustrate the *generability* of HierarNet in other domains, e.g., web traffic forecasting.

1 Introduction

In public health, accurately forecasting the incidence of diseases is critical for effective planning and intervention to mitigate negative effects (Zhong et al. 2024; Bomfim et al. 2020; Ribeiro et al. 2020). However, most current predictive modeling works (e.g. for dengue or influenza) focus on building flat models that incorporate exogenous inputs (e.g. climate or demographics) to forecast future disease outbreaks for different locations, but do not jointly model multiple spatial aggregation levels, such as provinces, regions, and country (Hii et al. 2012; Barboza et al. 2022; Kimura et al. 2022). Such hierarchical structures are commonly exhibited in disease data. E.g., infection counts may be aggregated by city, region, and country, or by different disease categories. Generating coherent forecasts across all these aggregated levels not only might help to improve prediction accuracy, but it is also essential for public health agencies to effectively allocate medical resources and plan interventions at multiple scales (e.g. local, state, or national levels). However, there are limited works on hierarchical disease prediction, such as (Mellor et al. 2023; Wang et al. 2019).

Generally, for predicting outbreaks at arbitrary locations, any existing forecasting techniques for time series (or longitudinal) data can be employed. Traditional statistical methods (e.g. ARIMA, ETS) can capture temporal trends and seasonality in individual series, but they struggle to incorporate complex, multi-level dependency structures (Spiliotis et al. 2019). Recently, deep learning (DL) models (e.g. MLP-, RNN-, and Transformer-based architectures) have emerged as alternative solutions with SOTA performances in various generic benchmarks such as Informer (Zhou et al. 2021), PatchTST (Nie et al. 2023), or iTransformer (Liu et al. 2023). However, most of these models have been designed for single or flat series and do not natively enforce hierarchical coherence. If applied naively to each level independently, deep forecasts may yield inconsistent aggregates.

Recently, Hierarchical time series (HTS) models, in which lower-level forecasts must sum to higher-level totals, have emerged in public health as promising native approaches for hierarchical disease forecasting (Mohanty et al. 2025; Mellor et al. 2023; Kimura et al. 2022). For example, (Mohanty et al. 2025) uses ARIMA as a base forecast model and MinT (Wickramasuriya et al. 2019), a hierarchical reconciliation approach, to predict COVID-19 in the USA from county to nation levels. Similarly, (Mellor et al. 2023) forecasts influenza hospital admissions using hierarchical generalized additive models. So far, most SOTA HTS methods follow post-hoc approaches, where reconciliation processes are performed on base forecast outputs to generate final coherent results. However, these methods rely on strong assumptions, such as Gaussian errors or unbiased base forecasts (Wickramasuriya et al. 2019; Hyndman et al. 2016; Hyndman et al. 2011; Ben et al. 2019). Other works enforce coherence as a learning constraint during training steps rather than a post-processing step such as (Han et al. 2021; Han et al. 2022) with bottom-up and (Rangapuram et al. 2021; Wang et al. 2024) with end-to-end reconciliation strategies. These methods do not rely on assumptions re-

*These authors contributed equally.

quired by post-processing methods and have shown promising results in different HTS benchmarks. However, they all focus on improving coherence among nodes at adjacent levels (particularly on higher levels like (Han et al. 2021)), thus neglecting inter-level relationships and nonadjacent level coherence. Moreover, improving coherence at a node does not explicitly lead to performance improvement on its aggregated members at the lower level, especially bottom layer ones. Hence, methods like SHARQ (Han et al. 2021) may not perform well if base models are not accurate enough.

Our contributions. In this paper, we introduce *HierarNet*, a special DL-based *independent iterative interactive* framework specifically designed for HTS forecasting. Unlike prior methods, HierarNet aims to improve layer-wise coherence and base model performance simultaneously. It is built upon a few key concepts described below.

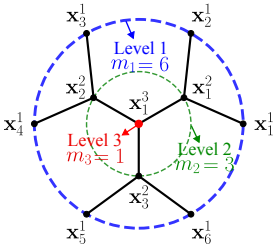
First, rather than using a single model to predict all nodes (locations) as in (Rangapuram et al. 2021; Cini et al. 2024), we train *independent* forecasting models for all nodes at all levels. Each node can employ any DL architectures and available local exogenous inputs to maximize its prediction accuracy. Due to the highly interdisciplinary nature of public health, relevant exogenous data (e.g. climate, demographics, or socio-economy) for diseases may be unavailable or inaccessible (Van Panhuis et al. 2014; Strongman et al. 2019). But this *independent local training* scheme allows us to flexibly utilize them for effective disease predictions.

Second, HierarNet introduces a special *iterative model interaction* training scheme, where each model from each node *globally interacts* with others from relevant nodes within a predefined interaction window in the hierarchy to maximize their agreements. This process is performed repeatedly by using *top-down* and *bottom-up ensemble* learning schemes to enforce both *inter* and *intra* level coherence at each round. This interactive training scheme is the key difference between HierarNet and existing DL HTS works described above. It helps to improve overall coherence among layers as well as prediction accuracy for all base models at the same time, as we will demonstrate in Section 4.

Case studies. We conduct extensive experiments on 5 different diseases (including Dengue fever, influenza, diarrhea, COVID-19, and Bluetongue) in 3 countries (including France, Vietnam, and the USA) to demonstrate the performance of HierarNet compared to 19 state-of-the-art (SOTA) methods in DL-based time series forecasting, such as WITRAN, SegRNN, TSMixer, PatchTST, or iTransformer, and HTS forecasting, such as MinT, SHARQ, ERM, and DYCHEM. We also illustrate the *generability* of HierarNet on other domains, e.g., web traffic forecasting.

2 Problem Formulation

Hierarchical Time Series (HTS). Let $\mathbf{x} = \{x_t |\ t \in [1, n]\} \in \mathbb{R}^n$ be a disease time series, where n is the number of time steps. A HTS includes multiple aggregation levels, from the most disaggregated series (bottom level) up to aggregated totals. Let $\mathbf{X}_1 = \{\mathbf{x}_i^1 |\ i \in [1, m_1]\} \in \mathbb{R}^{m_1 \times n}$ be the bottom level of the hierarchy with m_1 nodes, where



$$\mathbf{H}_1^{3,2} = \begin{bmatrix}1 & 1 & 1\end{bmatrix}$$
$$\mathbf{H}_1^{3,1} = \begin{bmatrix}1 & 1 & 1 & 1 & 1 & 1\end{bmatrix}$$
$$\mathbf{H}_1^{2,1} = \begin{bmatrix}1 & 1 & 0 & 0 & 0 & 0\end{bmatrix}$$
$$\mathbf{H}_2^{2,1} = \begin{bmatrix}0 & 0 & 1 & 1 & 0 & 0\end{bmatrix}$$
$$\mathbf{H}_3^{2,1} = \begin{bmatrix}0 & 0 & 0 & 0 & 1 & 1\end{bmatrix}$$
$$\left((\mathbf{H}_1^{1,1})^\intercal, \ldots, (\mathbf{H}_6^{1,1})^\intercal\right)^\intercal = \mathbf{I}_6$$

Figure 1: The tree and the aggregation vectors of a hierarchical time series with 3 levels, 4 total aggregated series and 6 bottom-level series ($L = 3, m_1 = 6, m_2 = 3$ and $m_3 = 1$).

each contains a time series. Then, a set of HTS can be defined with multiple levels $L \in \mathbb{Z}_+$, the most disaggregated (bottom) level at level 1 is the original node and the other nodes are the aggregated nodes, e.g. nodes at level $2, \ldots, L$. Next, we introduce a binary vector $\mathbf{H}_i^{\ell,\ell-k} \in \{0,1\}^{1 \times m_{\ell-k}}$ encoding the hierarchical aggregation of node i at level ℓ aggregated from level $\ell - k$, where $\ell \in \mathbb{Z}_+$ and $\ell \leq L$, $k \in \mathbb{Z}_+$ and $k < \ell$, and $m_{\ell-k}$ is the number of nodes at level $\ell - k$. The j-th element that $\mathbf{H}_i^{\ell,\ell-k}(j) = 1$ if and only if the j-th node at level $\ell - k$ contributes to the i-th aggregated node at level ℓ. Then the aggregated node series at level ℓ is given by $\mathbf{X}_\ell = \mathbf{H}^{\ell,\ell-k}\mathbf{X}_{\ell-k} \in \mathbb{R}^{m_\ell \times n}$. We form the full HTS by stacking all the aggregated and bottom-level series as $\mathbf{X} = (\mathbf{X}_1^\intercal, \ldots, \mathbf{X}_L^\intercal)^\intercal \in \mathbb{R}^{m \times n}$, where $m = \sum_{\ell=1}^{L} m_\ell$ is the total number of nodes and $\mathbf{X}_\ell = \mathbf{H}^{\ell,\ell-1}\mathbf{X}_{\ell-1} \in \mathbb{R}^{m_\ell \times n}$ ($\ell \geq 2$). To avoid a pathological hierarchy, we assume $m_1 \geq 2$ (at least two bottom-level nodes) and $m_2 \geq 1$ (at least one aggregated node). Moreover, each aggregate node must include at least one bottom-level node, i.e. $\sum_{j=1}^{m_1} \mathbf{H}_i^{\ell,1}(j) \geq 1$ for each $i = 1, \ldots, m_\ell$, and all the upper nodes at level ℓ must include all the lower level nodes, i.e. $\sum_{i=1}^{m_\ell} \sum_{j=1}^{m_{\ell-1}} \mathbf{H}_i^{\ell,\ell-1}(j) = m_{\ell-1}$. Figure 1 shows the tree-like structure of a HTS with 3 levels, 4 total aggregated series, and 6 bottom-level series.

Hierarchical Forecasting. Let $\mathbf{X} \in \mathbb{R}^{m \times n}$ denote historical hierarchical time series with n time steps and m hierarchical nodes. Let $\mathbf{Y} \in \mathbb{R}^{m \times h}$ denote future data and $\hat{\mathbf{Y}} \in \mathbb{R}^{m \times h}$ denote forecasts, where h is the forecast horizon. Let $\mathbf{E} = \{\mathbf{e}_i | i \in [1, m]\} \in \mathbb{R}^{m \times n}$ denote arbitrary exogenous variables associated with m-th nodes. Let f_i^ℓ be a model of node i at level ℓ, the prediction of this model can be denoted as $\hat{\mathbf{y}}_i^\ell = f_i^\ell(\mathbf{x}_i^\ell, \mathbf{e}_i^\ell)$, where $\mathbf{x}_i^\ell \in \mathbb{R}^n$, $\mathbf{e}_i^\ell \in \mathbb{R}^n$ and $\mathbf{y}_i^\ell \in \mathbb{R}^h$ are the historical series, exogenous variable series and forecast series for this model, respectively. Our objective is to minimize the error between the prediction $\hat{\mathbf{y}}_i^\ell$ and the future $\mathbf{y}_i^\ell$ for each node in each level, i.e. $\underset{i \in [1,m]}{\text{minimise}}\, \text{Error}(\hat{\mathbf{y}}_i^\ell, \mathbf{y}_i^\ell)$. For simplicity, we omit the subscript i to denote the stacking vectors at level ℓ that $\mathbf{x}^\ell = \left((\mathbf{x}_1^\ell)^\intercal, \ldots, (\mathbf{x}_{m_\ell}^\ell)^\intercal\right)^\intercal \in \mathbb{R}^{m_\ell \times n}$, $\mathbf{e}^\ell = \left((\mathbf{e}_1^\ell)^\intercal, \ldots, (\mathbf{e}_{m_\ell}^\ell)^\intercal\right)^\intercal \in \mathbb{R}^{m_\ell \times n}$ and $\mathbf{y}^\ell = \left((\mathbf{y}_1^\ell)^\intercal, \ldots, (\mathbf{y}_{m_\ell}^\ell)^\intercal\right)^\intercal \in \mathbb{R}^{m_\ell \times h}$ (m_ℓ is the number of nodes at level ℓ). We omit ℓ for an arbitrary node $i \in [1, m]$.

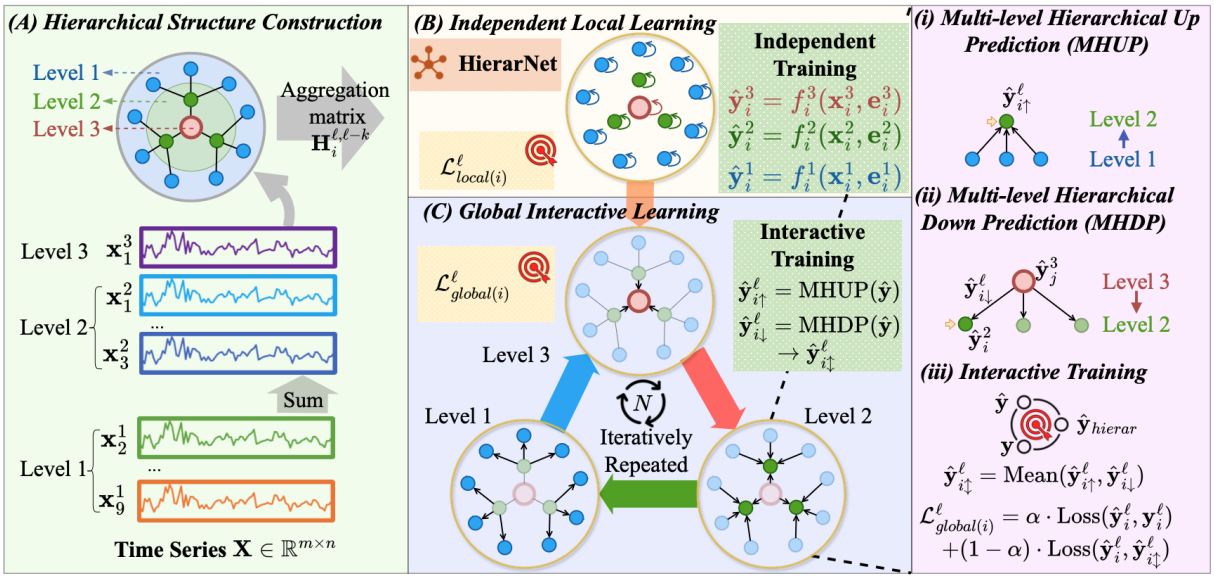


Figure 2: The overview pipeline of HierarNet with 3 main phases: (A) **Hierarchical Structure Construction** to construct the hierarchical time series from the bottom-level series $\{\mathbf{x}_i^1 | i \in [1, m_1]\}$; (B) **Independent Local Learning** to capture the temporal dependencies at each node locally via independent model trainings; (C) **Global Interactive Learning** to maximize *local* performance and *global* consistency among nodes via 3 specific modules MHUP, MHDP and IT.

3 Our Proposed Method HierarNet

Figure 2 illustrates the overall pipeline of HierarNet, which is built upon 3 key concepts: (i) *locally independent* DL-based model constructions for all hierarchy nodes with flexible exogenous inputs for different nodes; (ii) *globally model interactions* for performance and coherence improvements across all nodes and levels via *top-down* and *bottom-up ensemble* learning and *inter-intra level coherence* enforcement; (iii) *iterative* model refinements to improve *local* performance and *global* coherence among nodes rather than a single training round like all other works. In Phase A, we construct a hierarchical structure from bottom-level time series as described in Section 2 (e.g. via geographic information or categories). All models are independently initialized in Phase B locally before interacting with others globally via the hierarchy in Phase C, as detailed below.

3.1 Independent Local Learning

Unlike other HTS methods with a single model to predict all nodes, such as (Rangapuram et al. 2021; Cini et al. 2024), HierarNet approaches the problem following its own philosophy of *multiple small interactive models*, i.e. independent models at different nodes interact and update themselves to maximize their own performance and their coherence. For the interaction to be effective, each model first needs to capture well the unique characteristics of its local time series.

Concretely, we construct a set of models $F = \{f_i^\ell\}$ $(\forall \ell \in [1, L] \wedge i \in [1, m_\ell])$ for all nodes in the hierarchy independently, where f_i^ℓ is an arbitrary base prediction model (e.g. LSTM) using input disease data $\mathbf{x}_i^\ell$ and arbitrary exogenous data $\mathbf{e}_i^\ell$ (e.g. rainfall) at node i of level ℓ. Let $\hat{\mathbf{y}}_i^\ell = f_i^\ell(\mathbf{x}_i^\ell, \mathbf{e}_i^\ell)$ be the prediction outcome and $\mathbf{y}_i^\ell$ be the ground truth, f_i^ℓ is independently trained using its local loss:

$$\mathcal{L}_{local(i)}^{\ell} = \mathcal{L}_{pred(i)}^{\ell} = \text{Loss}(\mathbf{y}_i^\ell, \hat{\mathbf{y}}_i^\ell) \quad (1)$$

where Loss is an arbitrary loss function (e.g. MSE). This strategy is very *flexible* since each node can use a different model and data that are most suitable for it for efficacy.

3.2 Iterative Global Interactive Learning

The key idea of this Phase is that each model f_i^ℓ obtains prediction outcomes from its relevant nodes within K levels in the hierarchy, where K is a predefined interaction window, and uses them to adjust itself via its data aggregation relationships to enhance its own local prediction accuracy and global coherence. The refinement process is repeated in a predefined number of rounds (N) to maximize correction chances for all models. This phase is the main difference between HieararNet and other HTS methods, particularly recent DL-based ante-hoc methods, which only aim to improve coherence at higher levels such as SHARQ (Han et al. 2021). The interaction process is performed via a *top-down* and *bottom-up ensemble* learning strategy on the hierarchy.

Multi-level Hierarchical Up Prediction (MHUP). Following the property of HTS, we aggregate prediction outcomes at a level $\ell - k$ to create an expected output for level ℓ, where $k \in [1, min(K, \ell - 1)]$. We ensemble all results from lower levels to ℓ to form a bottom-up expected result. Let $\hat{\mathbf{y}}^\ell = \left((\hat{\mathbf{y}}_1^\ell)^\intercal, \ldots, (\hat{\mathbf{y}}_{m_\ell}^\ell)^\intercal\right)^\intercal \in \mathbb{R}^{m_\ell \times h}$ be a concatenated vector of all prediction outcomes for models f_i^ℓ at level ℓ.

Definition 1 (Bottom-up Ensemble)**.** *For each node i at level ℓ, its bottom-up prediction expectation within an in-*

Dataset	Dengue Fever				Diarrhoea				Influenza				Bluetongue			
Metric	RMSE	MAE	RMSE Rank	MAE Rank	RMSE	MAE	RMSE Rank	MAE Rank	RMSE	MAE	RMSE Rank	MAE Rank	RMSE	MAE	RMSE Rank	MAE Rank
DLinear	126.6	81.9	12.06	12.92	262.63	207.05	15.6	15.66	581.93	454.48	15.42	15.45	2.49	1.56	11.36	11.65
RLinear	108.99	61.6	5.68	5.3	164.78	116.57	7.49	7.33	362.96	255.63	7.4	7.15	1.97	1	6.09	5.9
TiDE	128.78	83.4	12.02	12.77	287.52	235.66	15.55	15.96	608.5	489.94	15.42	15.57	2.34	1.47	10.22	10.59
iTransformer	110.12	66.05	9.97	10.41	175.23	128.95	10.3	10.59	360.5	255.44	8.35	8.21	2.47	1.49	13.11	12.95
PatchTST	109.01	61.69	7.66	7.44	159.34	112.43	7.08	6.93	358.51	256.31	8.25	8.08	2.26	1.32	11.02	11.16
MICN	115.21	66.56	9.15	8.31	178.19	129.09	8.82	8.62	351.26	245.69	6.52	6.4	2.47	1.44	11.41	11.27
TSMixer	255.9	188.23	19.16	19.35	593.94	505.4	19.48	19.68	1062.21	907.94	19.2	19.39	10.42	8.46	18.68	18.83
TimeXer	112.91	65.53	8.96	9.12	161.04	115.93	7.41	7.61	359.13	255.72	7.81	7.84	2.62	1.68	12.54	12.89
FreTS	133.19	86.31	12	12.51	271.27	219.87	14.53	14.76	494.41	394.24	13.2	13.59	2.31	1.41	9.35	9.6
LightTS	117.09	70.51	8.44	8.96	169.2	123.76	7.72	7.86	388.85	283.38	9.6	9.67	2.42	1.35	7.83	7.44
FiLM	119.2	67.18	8.05	7.59	158.53	110.64	5.36	5.3	341.89	246.13	6.27	6.43	2.42	1.33	8.26	8.39
LSTM	117.48	69.38	9.24	9.34	157.67	111.26	6.19	6.27	355.9	257.09	7.32	7.55	2.4	1.39	8.4	8.34
GRU	119.07	70.85	10.81	10.92	158.04	112.8	6.58	6.69	390.35	281.46	9.48	9.46	2.95	1.8	10.36	10.22
SegRNN	133.18	81.97	13.08	13.2	195.81	142.23	12.22	11.63	464.67	344.51	13.17	12.75	3.2	2.19	14.12	14.6
WITRAN	164.28	109.85	16.32	16.9	246.84	182.13	15.83	15.54	601.43	448.87	16.23	15.96	4.94	3.45	16.34	16.59
MinT	117.28	69.84	6.93	7.12	175.22	127	8.95	8.87	365.57	272.42	6.31	6.64	2.3	1.5	4.28	5.16
ERM	145.93	78	9.76	7.7	215.65	149.4	12.64	12	371.2	256.58	9.33	8.34	45.45	9.36	10.6	9.18
DYCHEM	131.38	73.01	7.61	6.92	228.37	183.77	10.36	10.67	421.54	322.83	10.36	10.7	1.82	**0.68**	**2.95**	**2.03**
SHARQ	255.34	190.76	18.38	18.59	412.95	350.4	13.76	14.09	854.82	709.48	15.38	15.73	22.87	19.8	19.56	19.75
HierarNet	**106.78**	**59.62**	**4.75**	**4.63**	**150.95**	**104.5**	**4.15**	**3.94**	**331.79**	**235.89**	**4.97**	**5.08**	**1.64**	0.72	3.36	3.32

Table 1: Full results of Dengue Fever, Diarrhoea, Influenza and Bluetongue on RMSE, MAE, RMSE Rank and MAE Rank averaged from all provinces and 6 different forecasting horizon settings $h \in \{1, 2, 3, 4, 5, 6\}$ on look-back window length 12. The best values are marked in bold, the second best values are marked in underline.

teraction window K, denoted as $\hat{y}^{\ell}_{i\uparrow}$, is defined as follows:

$$\hat{\mathbf{y}}^{\ell}_{i\uparrow} = \frac{1}{K_b} \sum_{k=1}^{K_b} \left(\mathbf{H}_i^{\ell,\ell-k} \hat{\mathbf{y}}^{\ell-k} \right) \tag{2}$$

where $1 \leq \ell \leq L$, $K_b = min(K, \ell - 1)$, and $1 \leq i \leq m_\ell$.

Multi-level Hierarchical Down Prediction (MHDP). Similarly, we can calculate the expected prediction outcome for each node at level ℓ via the prediction outcomes of its ancestor node at a higher level $\ell + k$ and its sibling nodes at the same level ℓ, where $k \in [1, min(K, L - \ell)]$.

Definition 2 (Top-down Ensemble)**.** *For each node i at level ℓ and its ancestor node j at a level $\ell+k$ in the hierarchy (i.e. $\mathbf{H}_j^{\ell+k,\ell}(i) = 1$), the prediction expectation for node i from higher-levels, denoted as $\hat{\mathbf{y}}^{\ell}_{i\downarrow}$, is defined as follows:*

$$\hat{\mathbf{y}}^{\ell}_{i\downarrow} = \frac{1}{K_b} \sum_{k=1}^{K_b} \left(\hat{\mathbf{y}}_j^{\ell+k} - \mathbf{H}_j^{\ell+k,\ell} \hat{\mathbf{y}}^{\ell} + \hat{\mathbf{y}}_i^{\ell} \right) \tag{3}$$

where $1 \leq \ell \leq L$, $K_b = min(K, L - \ell)$, and $1 \leq i \leq m_\ell$.

Concretely, for each node i at level ℓ, the down prediction from node j at level $\ell + k$, where j is an ancestor of i ($H_j^{\ell+k,\ell}(i) = 1$), is calculated by prediction results of j minus the sum of all prediction results from all sibling nodes p of i ($p \neq i$ and $H_j^{\ell+k,\ell}(p) = 1$). The expectation of i is the averaged expectation from K_b levels above l.

Interactive Training (IT). Given an arbitrary node i at a level ℓ, our target is to update the model f_i^{ℓ} to maximize consistency from its own prediction $\hat{\mathbf{y}}_i^{\ell}$, its bottom-up expectation $\hat{\mathbf{y}}^{\ell}_{\uparrow i}$ and its top-down expectation $\hat{\mathbf{y}}^{\ell}_{\downarrow i}$. To do so, we first define a hierarchical consistent expectation as:

$$\hat{\mathbf{y}}^{\ell}_{i\updownarrow} = \text{Mean}(\hat{\mathbf{y}}^{\ell}_{i\uparrow}, \hat{\mathbf{y}}^{\ell}_{i\downarrow}) \tag{4}$$

For each model f_i^{ℓ}, we update it in each interaction round using a consistency loss function as follows:

$$\mathcal{L}^{\ell}_{global(i)} = \alpha \cdot \mathcal{L}^{\ell}_{pred(i)} + (1 - \alpha) \cdot \mathcal{L}^{\ell}_{hierar(i)} \tag{5}$$

$$= \alpha \cdot \text{Loss}(\hat{\mathbf{y}}_i^{\ell}, \mathbf{y}_i^{\ell}) + (1 - \alpha) \cdot \text{Loss}(\hat{\mathbf{y}}_i^{\ell}, \hat{\mathbf{y}}^{\ell}_{i\updownarrow}) \tag{6}$$

where α is a regulation parameter. Let lr be the initial learning rate of each model. In the global phase, we use a learning rate $lr \cdot \beta$, where $\beta \in [0, 1]$ is a predefined parameter.

3.3 Base Model Selection

HierarNet can be used with any existing time series prediction model. Unless otherwise stated, HierarNet uses only one Linear layer to capture the temporal dependencies, which balances the high performance and efficiency. Moreover, we remove the mean and standard deviation values from the input time series and add them back after the Linear layer to avoid non-stationary effects (Kim et al. 2022):

$$\hat{\mathbf{y}} = \text{Linear}(\frac{\mathbf{x} - \mu}{\sqrt{\sigma + \epsilon}}) \times \sqrt{\sigma + \epsilon} + \mu \tag{7}$$

where μ and σ are the mean and std of the input time series $\mathbf{x}$, and ϵ is a small constant for numerical stability.

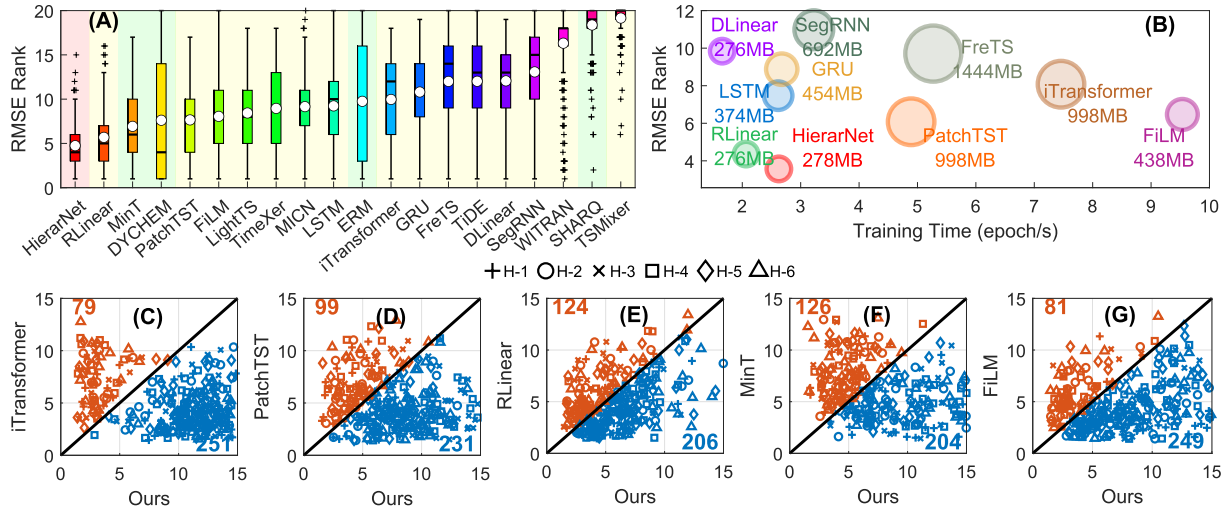


Figure 3: Performances on 55 provinces in Vietnam on the Dengue fever dataset. (A) Distributions of RMSE ranks for all methods (circles denote mean values). (B) Computation costs for some selected methods. (C-F) Direct RMSE rank comparisons between HierarNet and others on all 330 combinations of 55 provinces and 6 forecasting horizons. A point below a diagonal line means that HierarNet has a better RMSE rank.

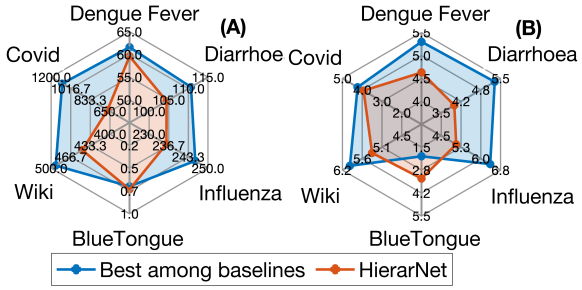


Figure 4: Averaged MAEs (A) and MAE rankings (B) of HierarNet over all prediction horizons compared to best values from all baselines for all studied datasets.

4 Experiments

Benchmarks. We demonstrate the performance of HierarNet on 5 disease datasets (incl. Dengue fever, influenza, diarrhea, COVID-19, and Bluetongue collected monthly or daily) in 3 countries (incl. France, Vietnam, and the USA) with administrative country geographic hierachies for predicting nodes in both long and short-terms. We also study the generality of HierarNet on other domains incl. web traffic forecasting (c.f. Supplementary for details).

Baselines. We compare HierarNet with 19 state-of-the-art (SOTA) baselines including: general Long-term time series forecasting (LTSF) models (incl. DLinear (Zeng et al. 2023), RLinear (Li et al. 2023), TiDE (Das et al. 2023), iTransformer (Liu et al. 2023), PatchTST (Nie et al. 2023), MICN (Wang et al. 2023), TSMixer (Chen et al. 2023), TimeXer (Wang et al. 2024), FreTS (Yi et al. 2023), LightTS (Zhang et al. 2022), and FiLM (Zhou et al. 2022)), RNN-based models (incl. LSTM (Hochreiter and Schmidhuber 1997), GRU (Chung et al. 2014), SegRNN (Lin et al. 2023), WITRAN (Jia et al. 2023)), HTSF-based models (incl. MinT (Wickramasuriya et al. 2019), ERM (Ben et al. 2019), DYCHEM (Han et al. 2022) and SHARQ (Han et al. 2021)).

Evaluations. We evaluate the performance of all methods on 6 prediction horizons $h = \{1, 2, 3, 4, 5, 6\}$, which is equivalent to 1-6 months/days ahead with a lookback window 12, using RMSE and MAE. Since RMSE and MAE can vary significantly among nodes, we propose to use their rank values among all baselines to have another view on their performance. Smaller RMSEs, MAEs, and rankings are better (c.f. Supplementary for more details).

4.1 Main Results

Accuracy analysis. Due to space limitations, Table 1 only shows the averaged results from 6 prediction horizons for 4 datasets (c.f. Supplementary for full results). The results show that HierarNet achieves the best results in 4 datasets and 4 metrics. E.g., HierarNet surpasses the SOTA Transformer-based method iTransformer with 7.76% and 7.65% reductions in averaged RMSEs and MAEs on the Influenza dataset, respectively. In the Dengue fever dataset, it outperforms the SOTA HTSF methods MinT with 2.18 and 2.49 point differences in RMSE and MAE rankings, respectively. These demonstrate that superior performances can be achieved even with a simple Linear architecture rather than large complex architectures like iTransformer, thus revealing the efficiency of our hierarchical interactive modeling approach for long-term and short-term disease forecasting.

Figure 3 further illustrates forecasting results for 55 provinces in Vietnam on the Dengue fever dataset. In (A), HierarNet acquires the best and stable overall RMSE rankings. In (C), HieararNet outperforms RLinear, an SOTA LTSF, on 206 (62.3%) of 330 cases across 55 provinces and 6 prediction horizons. It also surpasses MinT, a SOTA HTFS

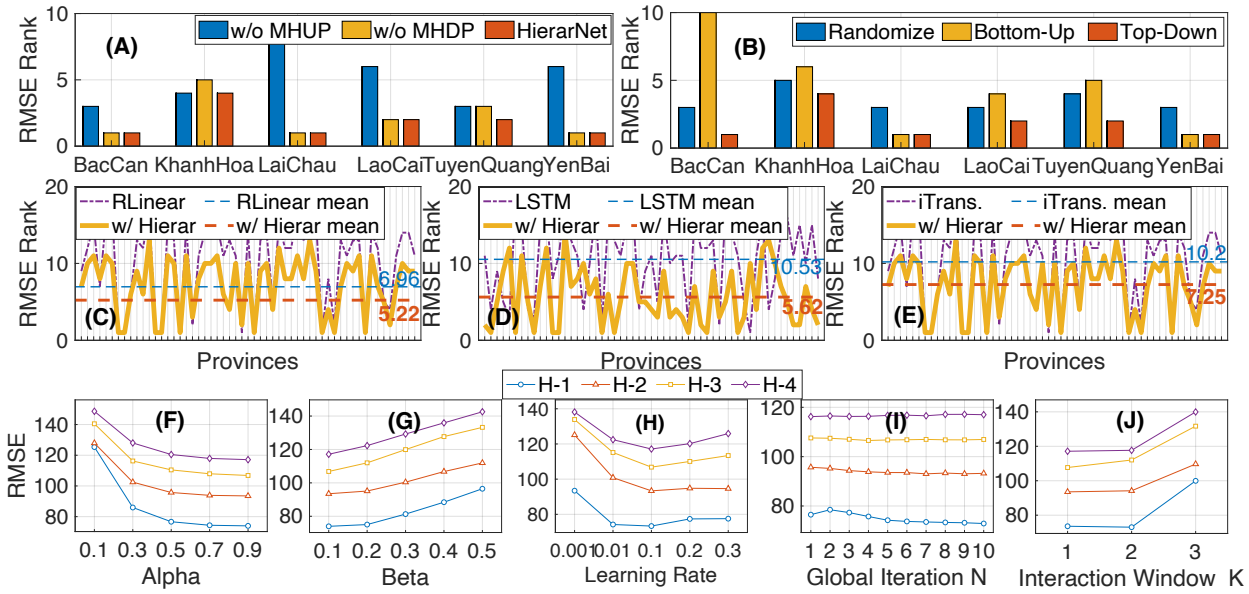


Figure 5: Ablation studies on the Dengue fever dataset. (A) Effects of MHUP and MHDP components in the global learning phase on 6 provinces. (B) Effects of different updating order strategies in the global learning phase on 6 provinces. (F-J) Hyperparameter sensitivity analysis on averaged RMSE of all provinces, including the loss regulation factor α, the number of global phase iterations N and the learning rate reduction factor β, the initial learning rate (lr), and the interaction window K.

method, in 204 (61.8%) cases in (F).

Figure 4 summarizes the performance of HierarNet compared to the best of all baselines over 6 prediction horizons and all datasets. HierarNet wins most of the cases, except on the Bluetongue and COVID datasets. On Bluetongue, DYCHEM performs the best, and HierarNet is the second-best method in MAE rankings, on COVID, MinTrace is the best on MAEs, and HierarNet is the best on MAE rankings.

Computational analysis. The integration of a backbone model of a single-layer Linear with the HierarNet framework, while introducing only marginal computational overhead compared to the original model, achieves a substantial performance leap as shown in Figure 3 (B). HierarNet requires multiple levels in the interactive training step for hierarchical dependencies modeling, which increases the learnable parameters and memory usage. E.g., it requires 3.58x less memory and runs 2.83x faster than iTransformer, while being more accurate in forecasting as shown in Figure 3 (A).

4.2 Ablation Studies and Analysis

Effects of the MHUP and MHDP schemes. We conduct ablation experiments on 6 provinces: BacCan, KhanhHoa, LaiChau, LaoCai, TuyenQuang, and YenBai of the Dengue fever dataset, as shown in Figure 5 (A). HierarNet can benefit significantly by doing both MHUP and MHDP compared to one-way updates. The reason is simple: using only MHUP or MHDP will lead to the accumulation of prediction errors.

Effects of the IT scheme. In Figure 5 (B), we study 3 different ways for interactive training: bottom-up, top-down, and randomized. The bottom-up way updates nodes from the bottom level to the final top level, and reversely for the top-down way. Randomized means that we update nodes in a random order. The results show that the top-down updating scheme is the most effective way. It makes sense since higher levels suffer from accumulation errors more than lower levels and should be updated first.

Effects of different backbone models. Since HierarNet is a *generic* framework, any backbone model can get benefit from it. In Figure 5 (C-E), we employ 3 DL architectures as backbone models for HierarNet, including: RLinear (MLP-based), LSTM (RNN-based), and iTransformer (Transformer-based). Our HierarNet framework helps to significantly boost the prediction accuracy of all backbone models in most studied provinces, proving the effectiveness of our iterative interaction training scheme.

Hyperparameter sensitivity. HierarNet has 3 parameters including: the loss regulation parameter α, the number of global iteraction N and the interaction window K. We study their sensitivity in Figure 5 (F-J). Increasing parameters α and N means models will be updated more, thus increasing overall accuracy. However, too large K decreases the performance since too far away nodes will have looser relationships, thus causing harm when interacting together. The best value for lr is around 0.1, while small values are preferred for β to limit updating speeds in the interaction phase.

Interaction analysis. In Figure 6 (A), the global interaction phase helps to improve Dengue prediction accuracy on 32/55 provinces compared to the local phase, especially in the Red River Delta area. In (B), we look deeper into some improvement and non-improvement cases during the interaction process. E.g., in BacCan and HaiPhong, the test loss generally decreases during the interaction phase, and

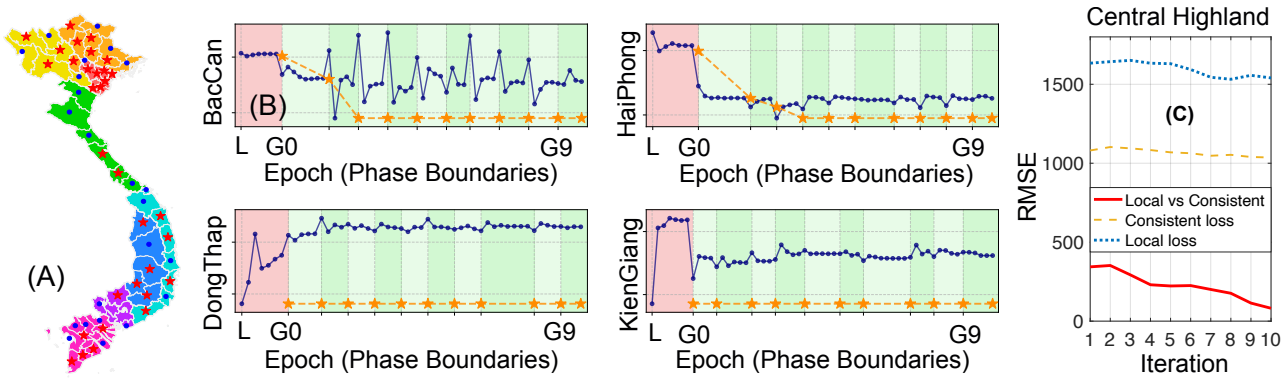


Figure 6: Effects of the global interaction learning on the Dengue fever dataset ($h = 1$). (A) The geographic maps of Vietnam, where red stars mark provinces with performance improvements after the global interaction phase and blue circles mark no-improvement cases. (B) The blue circle line is the loss value at each epoch (test loss) of the current model, and the starred yellow dash line is the loss value of the best-so-far model (selected from validation data) in some provinces. L denotes the independent phase, and G0 to G9 denote 10 interactive rounds. (C) Performance of a high-level node during the global interaction.

the best loss also decreases and remains unchanged after 2-3 global interaction rounds, indicating effective interactions and performance improvements. On the other hand, in DongThap and KienGiang, the interactions are ineffective with no change in the best loss and overall accuracy. Moreover, performance degradations are rarely seen in our experiments. In (C), we study the behaviors of a high-level node, Central Highland, during interactions. As we see, the local loss (i.e. its own predictions vs ground truths) decreases at each iteration, showing performance improvements. The consistency loss (i.e. aggregated predictions from other nodes vs ground truths) also decreases. At the same time, the local vs consistent loss (i.e. its own predictions vs aggregated predictions) reduces significantly. These indicate better global coherence among nodes in the hierarchy at each iteration of HierarNet.

5 Related Works

DL-based time series (TS) forecasting. The broader field of DL for TS forecasting has produced powerful architectures that could be applied to HTS forecasting. For example, MLP-based models such as DLinear (Zeng et al. 2023) capture the seasonality and trend patterns with a lightweight architecture, while Transformers-based models (e.g. Informer (Zhou et al. 2021), PatchTST (Nie et al. 2023), iTransformer (Liu et al. 2023), etc.) leverage attention to handle long-range dependencies. Neural forecasting models such as N-BEATS (Oreshkin et al. 2019) and DeepAR (Salinas et al. 2020) stack interpretable forecasting blocks or probabilistic auto-regressive components. However, these models are generally applied to flat multivariate series and do not account for aggregation constraints. Incorporating hierarchical structure into such deep models remains an open challenge.

HTS forecasting. These techniques can be divided into post-hoc and ante-hoc methods. Post-hoc ones provide different ways to combine base forecasting results in a post-processing reconciliation step to generate coherent results either via statistical methods (Wickramasuriya et al. 2019; Hyndman et al. 2016; Hyndman et al. 2011; Ben et al. 2019) or deep learning (Wang et al. 2024; Rangapuram et al. 2023). In contrast, our method HierarNet is an ante-hoc method, which aim to enforce coherence during model training. However, instead of using a single large model to predict all HTS nodes like (Rangapuram et al. 2021; Cini et al. 2024), HierarNet follows a multiple model approach like (Han et al. 2021; Han et al. 2022). But while these methods aim to enhance coherence at higher levels and highly rely on base models' initial performances. HierarNet introduces a unique iterative interaction scheme that allows all base models to propagate their outcomes to others across the hierarchy and update themselves to improve both their own local performances and global coherence among levels.

HTS forecasting for diseases. As mentioned in the Introduction, due to its potential impacts on public health, hierarchical disease forecasting has been emerged recently but with very limited works such as (Mohanty et al. 2025; Mellor et al. 2023; Wang et al. 2019). HierarNet, with its flexibility in designs, light computation cost, and high forecasting accuracy, would be a useful method for disease forecasting.

6 Conclusion

HierarNet is specifically designed for HTS forecasting with some distingushed key characteristics: (i) locally independent DL-based model constructions for all nodes; (ii) globally model interactions across all levels to improve both forecasting accuracy and hieararchical coherence; (iii) iterative model designs, and (iv) flexibility and generality wrt. models, exogenous variables and application domains. Experiments on 5 real-world epidemiological datasets demonstrate the effectiveness of our framework in modeling the complex temporal dependencies of diseases across different administrative levels. Our model consistently outperforms SOTA baselines, highlighting the benefits of incorporating hierarchical structure during interactive training.

Acknowledgments

This research is part-funded by the European Union (Horizon Europe 2021-2027 Framework Program Grant Agreement number 10107245. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them) and by the Engineering and Physical Sciences Research Council under grant number EP/X029174/1. In addition, Zichi Zhang is supported by the China Scholarship Council (CSC) under Grant No. 202410080008.

References

Barboza, M. F. X.; Monteiro, K. H. d. C.; Rodrigues, I. R.; Santos, G. L.; Monteiro, W. M.; Figueira, E. A. G.; Sampaio, V. d. S.; Lynn, T.; and Endo, P. T. 2022. Prediction of malaria using deep learning models: A case study on city clusters in the state of Amazonas, Brazil, from 2003 to 2018. *Rev. Soc. Bras. Med. Trop.*, 55: e0420–2021.

Ben Taieb, S.; and Koo, B. 2019. Regularized regression for hierarchical forecasting without unbiasedness conditions. In *ACML*, 1337–1347.

Bomfim, R.; Pei, S.; Shaman, J.; Yamana, T.; Makse, H. A.; Andrade Jr, J. S.; Lima Neto, A. S.; and Furtado, V. 2020. Predicting dengue outbreaks at neighbourhood level using human mobility in urban areas. *J. R. Soc. Interface*, 17(171): 20200691.

Chen, S.-A.; Li, C.-L.; Yoder, N.; Arik, S. O.; and Pfister, T. 2023. Tsmixer: An all-mlp architecture for time series forecasting. *arXiv preprint arXiv:2303.06053*.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Cini, A.; Mandic, D. P.; and Alippi, C. 2024. Graph-based Time Series Clustering for End-to-End Hierarchical Forecasting. In *ICML*. OpenReview.net.

Das, A.; Kong, W.; Leach, A.; Mathur, S.; Sen, R.; and Yu, R. 2023. Long-term forecasting with tide: Time-series dense encoder. *arXiv preprint arXiv:2304.08424*.

Han, X.; Dasgupta, S.; and Ghosh, J. 2021. Simultaneously reconciled quantile forecasting of hierarchically related time series. In *AISTAT*, 190–198. PMLR.

Han, X.; Hu, J.; and Ghosh, J. 2022. Dynamic combination of heterogeneous models for hierarchical time series. In *ICDMW*, 1207–1216. IEEE.

Hii, Y. L.; Zhu, H.; Ng, N.; Ng, L. C.; and Rocklöv, J. 2012. Forecast of dengue incidence using temperature and rainfall. *PLoS Negl. Trop. Dis.*, 6(11): e1908.

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

Hyndman, R. J.; Ahmed, R. A.; Athanasopoulos, G.; and Shang, H. L. 2011. Optimal combination forecasts for hierarchical time series. *Comput. Stat. Data Anal.*, 55(9): 2579–2589.

Hyndman, R. J.; Lee, A. J.; and Wang, E. 2016. Fast computation of reconciled forecasts for hierarchical and grouped time series. *Comput. Stat. Data Anal.*, 97: 16–32.

Jia, Y.; Lin, Y.; Hao, X.; Lin, Y.; Guo, S.; and Wan, H. 2023. Witran: Water-wave information transmission and recurrent acceleration network for long-range time series forecasting. In *NeurIPS*, volume 36, 12389–12456.

Kim, T.; Kim, J.; Tae, Y.; Park, C.; Choi, J.; and Choo, J. 2022. Reversible Instance Normalization for Accurate Time-Series Forecasting against Distribution Shift. In *ICLR*.

Kimura, T.; Matsubara, Y.; Kawabata, K.; and Sakurai, Y. 2022. Fast mining and forecasting of co-evolving epidemiological data streams. In *ACML*, 3157–3167.

Li, Z.; Qi, S.; Li, Y.; and Xu, Z. 2023. Revisiting Long-term Time Series Forecasting: An Investigation on Linear Mapping. *CoRR*, abs/2305.10721.

Lin, S.; Lin, W.; Wu, W.; Zhao, F.; Mo, R.; and Zhang, H. 2023. Segrnn: Segment recurrent neural network for long-term time series forecasting. *arXiv preprint arXiv:2308.11200*.

Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; and Long, M. 2023. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. *arXiv preprint arXiv:2310.06625*.

Mellor, J.; Christie, R.; Overton, C. E.; Paton, R. S.; Leslie, R.; Tang, M.; Deeny, S.; and Ward, T. 2023. Forecasting influenza hospital admissions within English sub-regions using hierarchical generalised additive models. *Commun. Med.*, 3(1): 190.

Mohanty, S.; Shimamura, A.; Nicholson, C. D.; González, A. D.; and Razzaghi, T. 2025. Hierarchical Time Series Forecasting of COVID-19 Cases Using County-Level Clustering Data. In *ORFO*, volume 6, 28. Springer.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *ICLR*.

Oreshkin, B. N.; Carpov, D.; Chapados, N.; and Bengio, Y. 2019. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437*.

Rangapuram, S. S.; Kapoor, S.; Nirwan, R. S.; Mercado, P.; Januschowski, T.; Wang, Y.; and Bohlke-Schneider, M. 2023. Coherent probabilistic forecasting of temporal hierarchies. In *AISTATS*, 9362–9376. PMLR.

Rangapuram, S. S.; Werner, L. D.; Benidis, K.; Mercado, P.; Gasthaus, J.; and Januschowski, T. 2021. End-to-end learning of coherent probabilistic forecasts for hierarchical time series. In *ICML*, 8832–8843. PMLR.

Ribeiro, M. H. D. M.; da Silva, R. G.; Mariani, V. C.; and dos Santos Coelho, L. 2020. Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. *Chaos Solit. Fractals*, 135: 109853.

Salinas, D.; Flunkert, V.; Gasthaus, J.; and Januschowski, T. 2020. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *Int. J. Forecast.*, 36(3): 1181–1191.

Spiliotis, E.; Petropoulos, F.; and Assimakopoulos, V. 2019. Improving the forecasting performance of temporal hierarchies. *Plos one*, 14(10): e0223422.

Strongman, H.; Williams, R.; Meeraus, W.; Murray-Thomas, T.; Campbell, J.; Carty, L.; Dedman, D.; Gallagher, A. M.; Oyinlola, J.; Kousoulis, A.; et al. 2019. Limitations for health research with restricted data collection from UK primary care. *Pharmacoepidemiol. Drug Saf.*, 28(6): 777–787.

Van Panhuis, W. G.; Paul, P.; Emerson, C.; Grefenstette, J.; Wilder, R.; Herbst, A. J.; Heymann, D.; and Burke, D. S. 2014. A systematic review of barriers to data sharing in public health. *BMC public health*, 14(1): 1144.

Wang, H.; Peng, J.; Huang, F.; Wang, J.; Chen, J.; and Xiao, Y. 2023. MICN: Multi-scale Local and Global Context Modeling for Long-term Series Forecasting. In *ICLR*.

Wang, L.; Chen, J.; and Marathe, M. 2019. DEFSI: Deep learning based epidemic forecasting with synthetic information. In *AAAI*, 9607–9612.

Wang, S. 2024. Neuralreconciler for hierarchical time series forecasting. In *WSDM '24*, 731–739.

Wang, Y.; Wu, H.; Dong, J.; Qin, G.; Zhang, H.; Liu, Y.; Qiu, Y.; Wang, J.; and Long, M. 2024. Timexer: Empowering transformers for time series forecasting with exogenous variables. *arXiv preprint arXiv:2402.19072*.

Wickramasuriya, S. L.; Athanasopoulos, G.; and Hyndman, R. J. 2019. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *J. Am. Stat. Assoc.*, 114(526): 804–819.

Yi, K.; Zhang, Q.; Fan, W.; Wang, S.; Wang, P.; He, H.; An, N.; Lian, D.; Cao, L.; and Niu, Z. 2023. Frequency-domain mlps are more effective learners in time series forecasting. In *NeurIPS*, volume 36, 76656–76679.

Zeng, A.; Chen, M.; Zhang, L.; and Xu, Q. 2023. Are Transformers Effective for Time Series Forecasting? In *AAAI*, 11121–11128.

Zhang, T.; Zhang, Y.; Cao, W.; Bian, J.; Yi, X.; Zheng, S.; and Li, J. 2022. Less Is More: Fast Multivariate Time Series Forecasting with Light Sampling-oriented MLP Structures. arXiv. *arXiv preprint arXiv:2207.01186*.

Zhong, J.; Shu, E.; Zhang, S.; Yang, Q.; Chen, Q.; and Niu, B. 2024. Prediction and transmission analysis of bluetongue disease in China. *Prev. Vet. Med.*, 230: 106290.

Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI*, 11106–11115.

Zhou, T.; Ma, Z.; Wen, Q.; Sun, L.; Yao, T.; Yin, W.; Jin, R.; et al. 2022. Film: Frequency improved legendre memory model for long-term time series forecasting. *Adv. Neural Inf. Process. Syst.*, 35: 12677–12690.