Modeling and Simulation 建模与仿真, 2025, 14(5), 293-303
Published Online May 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.145394

Hans汉斯

基于选择性状态空间的时间序列预测模型

林琦淇

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2025年4月14日；录用日期：2025年5月7日；发布日期：2025年5月14日

摘　要

针对时间序列数据特征复杂、模式多样且长期依赖导致模型计算复杂度高等问题，文章提出了一种基于选择性状态空间与时间卷积网络的多尺度时间序列预测模型。该模型通过专家混合网络为不同类型的时间序列数据动态分配专家模块，并结合一维离散小波变换与时间卷积网络捕捉多尺度特征。具体而言，专家模块利用小波变换将时序数据分解为高频和低频分量，分别提取短期波动和长期趋势信息，并通过时间卷积网络进行特征提取。处理后的数据输入编码器，采用基于选择性状态空间的Mamba模型同时建模正向和逆向时序信息，增强时序建模能力。实验结果表明，该模型在交通、经济、天气和电力等领域的多变量时间序列预测任务中均显著优于现有基线模型，验证了其有效性与鲁棒性。

关键词

多变量时序预测，小波变换，状态空间模型

Time Series Prediction Model Based on Selective State Space

Qiqi Lin

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Apr. 14th, 2025; accepted: May 7th, 2025; published: May 14th, 2025

Abstract

To address the challenges of complex time series data characteristics, diverse patterns, and high computational complexity caused by long-term dependencies, this paper proposes a multi-scale time series prediction model based on selective state space and temporal convolutional networks. The model dynamically allocates expert modules to different types of time series data through a mixture of expert networks and captures multi-scale features by integrating one-dimensional

文章引用: 林琦淇. 基于选择性状态空间的时间序列预测模型[J]. 建模与仿真, 2025, 14(5): 293-303.
DOI: 10.12677/mos.2025.145394

discrete wavelet transform with temporal convolutional networks. Specifically, the expert modules decompose time series data into high-frequency and low-frequency components using wavelet transform, extracting short-term fluctuations and long-term trends, respectively, and performing feature extraction through temporal convolutional networks. The processed data is then fed into an encoder, which employs a selective state space-based Mamba model to simultaneously model forward and backward temporal information, enhancing temporal modeling capabilities. Experimental results demonstrate that the model significantly outperforms existing baseline models in multivariate time series prediction tasks across various domains, including transportation, economics, weather, and electricity, validating its effectiveness and robustness.

Keywords

Multivariate Time Series Forecasting, Wavelet Transform, State Space Model

1. 引言

随着深度学习技术的迅猛发展，时间序列预测在诸多领域得到了广泛应用。在金融投资领域[1]，时间序列预测为市场趋势研判与投资决策优化提供了量化支持；在气象科学领域[2]，高精度的预测结果为灾害预警机制构建与农业生产规划奠定了数据基础；在交通流量估计与能源消耗预测方面[3]，时间序列预测技术为城市规划以及能源管理决策提供了科学的解决方案。作为深度学习领域的重要应用之一，时间序列预测在多个学科领域展现出极高的研究价值与实践意义。

由于时间序列数据特征复杂，目前时间序列预测任务仍然面临着多种挑战。首要挑战在于时序数据中蕴含着多重潜在模式，包括趋势性、季节性、周期性以及随机噪声等。趋势性表现为数据的长期单调变化，季节性表现为固定时间间隔内的重复规律(如日、月或年周期)，而周期性则指非固定间隔的波动(如经济周期)。此外，外部因素(如突发事件)的干扰进一步增加了时间序列的复杂性。这种模式多样性在不同应用领域表现不同，例如，电力数据通常具有较强的季节性[4]，而金融市场数据则往往表现出高度的随机性和波动性[1]。这种异质性使得单一模型难以普适于各类时间序列数据。时间序列预测任务面临的另一个重要挑战是时序数据中的长期依赖问题，即当前时刻的数据可能依赖于远距离的历史信息。例如，气象预测中未来天气状况可能受数月甚至数年前气候模式的影响[2]。传统时序预测模型(如 RNN [5]、LSTM [6])在处理长期依赖时，常面临梯度消失或梯度爆炸的困境，难以有效捕捉远距离依赖关系。另外，随着序列长度的增加，模型的计算复杂度显著上升，尤其是在处理高维多变量时间序列时，计算资源的需求呈指数级增长，这进一步加剧了模型的训练和推理难度。因此，如何在保证模型性能的同时降低计算复杂度，是时间序列预测领域亟待解决的问题之一。

针对时间序列数据中复杂多样的特征提取问题，传统方法主要依赖于统计学原理。如 Elfkey 等人[7]采用自相关函数绘制自相关图并分析峰值，来识别时间序列中的周期性模式。然而，这种方法通常假设数据具有线性和平稳性，难以适应现实世界中复杂的非线性时序数据。为了克服这一局限性，研究者转向频域分析方法，如 Pfander 等人[8]结合了傅里叶变换和小波变换，有效检测了时间序列中的周期性特征，验证了频域分析在处理复杂时序数据中的优势。受此启发，本文提出了结合小波变换和时间卷积网络进行多尺度特征提取。此外，对于长序列建模中计算复杂度高的问题。尽管基于自注意力机制的模型

(如 Autoformer [9]、FedFormer [10]、Informer [11])能够捕捉长距离依赖关系，但其计算复杂度随序列长度呈二次方增长，限制了其在大规模时序数据中的应用。其中 Autoformer 设计了一种基于周期特征的自相关机制，揭示了子序列之间的内在依赖关系。尽管 Autoformer 的自相关机制有效优化了自注意力机制的计算复杂度，但它本质上是基于 Transformer 的。当处理长序列数据时，Autoformer 仍需要大量的计算资源。Informer 引入了 ProbSparse 注意力机制，通过 KL 散度度量 Query 向量的稀疏性，减少了计算量，支持更长的输入序列，然而 ProbSparse 注意力机制在数据噪声较大时表现不稳定。Fedformer 结合了频域增强和季节性趋势分解，将复杂度降至线性，支持极长序列输入，但是由于傅里叶变换假设序列平稳，对突变型序列(如突发事件冲击)建模能力较弱。这些方法虽然解决了长序列效率问题，但均面临两大共性局限：依赖注意力或频域分解的单一范式，难以同时适应趋势、周期与突发噪声的混合模式；另外现有方法多基于单向历史信息，忽略未来潜在模式对当前状态的逆向影响。最近，基于状态空间模型的 Mamba 模型[12]凭借其线性计算复杂度和高效的长序列建模能力，在序列任务中表现出色。基于此，本文在 Mamba 模型的基础上进行了改进，通过同时建模正向和逆向时序信息，增强了模型对长序列的建模能力。

针对上述不足，本文提出的模型使用了一种基于专家网络的多尺度特征提取框架。首先，通过门控机制将时间序列数据分配到不同的专家模块中，这一分配过程充分考虑了数据的季节性与趋势特征。在每个专家模块中，利用小波变换将数据分解为高频与低频分量，分别捕捉短期波动和长期趋势信息，并通过时间卷积网络进行特征提取。这种多尺度特征提取机制增强了对复杂时序模式的适应性。随后，经过专家网络处理的数据被输入到编码器中进一步建模。编码器采用双向建模策略，集成了两个 Mamba 模型：一个处理正向时序数据，另一个处理逆向时序数据。这种双向建模方法同时考虑了过去与未来的信息，显著提升了模型对时序依赖关系的建模能力。此外，Mamba 模型基于选择性状态空间模型实现，具有线性计算复杂度，能够高效处理长序列数据。最后，实验结果表明，本文模型在多变量时间序列预测任务中表现优异。在交通、经济、天气和电力四个领域的公开数据集上，模型均显著优于现有基线方法，验证了其在处理复杂时序数据中的有效性与鲁棒性。

本文的结构安排如下：第 2 节给出了问题定义和相关技术介绍，第 3 节介绍了本文模型框架并描述了模型中每个子模块的构造及其功能；第 4 节展示了实验结果和分析；最后，在第 5 节中，总结了全文工作并进一步展望了未来可能的研究方向。

2. 预备知识

本节在明确问题定义后，首先对模型整体框架进行了概述，随后详细阐述了各个子模块的设计，包括利用小波变换和时间卷积网络捕获时间序列周期特征的混合专家模块，采用选择状态空间模型对特征数据建模的编码器模块。

2.1. 问题定义

多变量时间序列预测任务利用过去的时间序列数据建模并预测未来时间点的值。通过给定历史观测值 $X=\{x_1,\cdots,x_T\}\in\mathbb{R}^{B\times T\times N}$，预测未来的时间序列 $Y=\{x_{T+1},\cdots,x_{T+S}\}\in\mathbb{R}^{S\times N}$。其中，$B$ 表示批量大小，T 表示历史时间步长，S 表示需要预测的未来时间步长，N 表示每个时间步的变量数量。

2.2. 小波变换(DWT)

一维离散小波变换(Discrete Wavelet Transform, DWT) [13]是一种信号处理工具，能够将时序信号分解为高频分量和低频分量。高频分量通常反应信号的细节部分(如噪声或快速变化部分)，而低频分量则对应信号的平滑部分(如趋势或缓慢变化部分)。DWT 通过使用一组小波基函数实现信号分解，这些基函数由尺度函数和小波函数组成，前者用于提取信号的低频分量，后者则用于捕捉信号的高频分量。小波基函

数通过对母小波进行平移和缩放生成。假设输入信号为 $x[n]$，长度为 N，则 DWT 的分解过程表示如下：

$$A_j[k]=\sum_n x[n]\cdot\phi_{j,k}[n] \tag{1}$$

$$D_j[k]=\sum_n x[n]\cdot\psi_{j,k}[n] \tag{2}$$

其中 A_j 和 D_j 分别表示低频分量和高频分量，$\phi_{j,k}$ 和 $\psi_{j,k}$ 分别表示尺度函数和小波函数在尺度 j 和位置 k 处的平移和缩放形式。

2.3. 时间卷积网络(TCN)

时间卷积网络(Temporal Convolutional Network, TCN) [14]是一种处理时序数据的卷积神经网络架构。TCN 利用一维卷积和扩张卷积捕捉时序数据中的长期依赖关系。扩张卷积通过引入“空洞”扩展感受野，从而在不增加参数量的情况下捕捉远距离依赖关系。假设输入序列为 $x=(x_1,x_2,\cdots,x_T)$，卷积核为 $w=(w_1,w_2,\cdots,w_k)$，扩张率为 d，则一维卷积和扩张卷积可以表示为：

$$y_t=\sum_{i=1}^{k} w_i\cdot x_{t+i-1} \tag{3}$$

$$y_t=\sum_{i=1}^{k} w_i\cdot x_{t-d\cdot(i-1)} \tag{4}$$

TCN 的典型结构包括多层扩张卷积，每层的扩张率逐渐增加，通常使用 ReLU 作为激活函数，并在卷积层后使用批量归一化(Batch Norm [15])。第 l 层 TCN 的计算公式与通过残差连接得到的最终输出如下所示：

$$y_t^{(l)}=\mathrm{ReLU}\left(\mathrm{BatchNorm}\left(\sum_{i=1}^{k} w_i\cdot x_{t+i-1}\right)\right) \tag{5}$$

$$\mathrm{output}^{(l)}=y^{(l)}+x^{(l-1)} \tag{6}$$

3. 建立模型

3.1. 整体框架

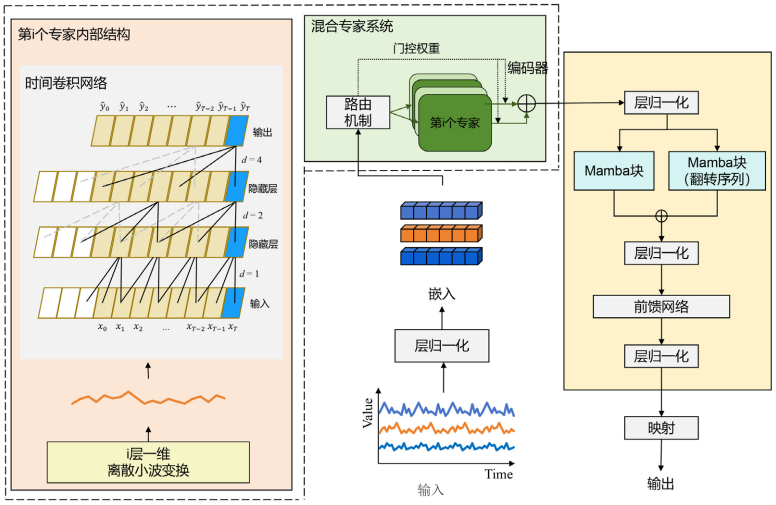

Figure 1. Framework of model
图 1. 模型的整体架构

图 1 展示了模型的整体架构，由层归一化、专家混合网络、编码器、映射层堆叠而成。其中，层归一化技术[16]可以削弱原始时间序列数据中的非平稳性影响；映射层[17]则是一个专为长序列预测任务定制的全连接神经网络。

模型的核心组件是专家混合网络和编码器模块。由于时间序列数据的内在周期性表现出高度多样化，传统基于数学统计原理的周期特征提取方法已难以有效应对。为此，本文采用了专家混合网络，通过门控机制将具有相似趋势性和季节性特征的时序数据分配到同一专家模块中进行处理。每个专家模块包括小波变换和时间卷积网络，其中小波变换层数因不同专家模块而有所差异。时间序列经过小波变换分解为高频与低频分量，然后利用时间卷积网络提取多尺度周期特征。随后数据流进入编码器进行建模，编码器中含有两个基于选择状态空间实现的 Mamba 模型[12]，分别处理正向与逆向时序数据，增强时序建模的能力。

3.2. 周期特征提取模块

不同应用场景中，时间序列数据具有显著差异化的周期特征。因此，选择适当的小波变换分解层数成为关键挑战。分解层数的设定直接影响多尺度分析的粒度。一般来说，复杂的时间序列需要更大的分解层数以全面捕捉其多维特征，但过小的层数可能导致深层次周期特征被遗漏，而过大的层数则可能引入冗余信息，甚至导致模型过拟合。为解决这一问题，本文引入了专家混合网络技术，通过其智能决策机制，能够根据时间序列数据的特性动态调整小波变换的分解层数，从而在捕捉多尺度特征的同时避免冗余信息的产生，显著提升了模型的建模能力和预测效率。

如图 1 所示，专家混合网络由一个路由函数、多个专家模块以及一个聚合函数组成。输入数据 $X_{in} \in \mathbb{R}^{B\times T\times N}$ 首先通过嵌入模块的处理，将数据转换为高维空间中的表现形式 $X_{emb} \in \mathbb{R}^{B\times N\times d_m}$，其中，$B$ 表示批量大小，N 表示每个时间步的变量数量，d_m 表示嵌入维度大小。X_{emb} 作为混合专家网络的输入，首先通过路由函数计算并分配权重给每个专家模块。这一过程基于输入数据的季节趋势性特征，目的是为不同特性的数据确定合适的小波变换分解层级。路由函数的具体公式如(7)所示。

$$R\left(X_{emb}\right)=\mathrm{Softmax}\left(\mathrm{TopK}\left(\left(X_{emb}W_r\right)+\epsilon\cdot\mathrm{Softplus}\left(X_{emb}W_{noise}\right),k\right)\right) \tag{7}$$

其中，$R(\cdot)$表示整个路由函数，$W_r, W_{noise} \in \mathbb{R}^{d_m\times N_e}$ 是用于可学习的权重矩阵参数，N_e 表示专家的数量，TopK(·)表示保留前 k 个最大值。

时序数据通过路由机制被分配至对应的专家模块处理后，各个专家模块的输出 $X^{i}_{\exp\_out} \in \mathbb{R}^{B_i\times N\times d_m}$ 通过聚合函数加权聚合。其中，B_i 表示第 i 个专家的批量大小。聚合函数中利用了 $R(\cdot)$路由函数确保仅聚焦于前 k 个专家模块的输出结果。另外，在加权聚合前，为了保证所有专家的输出结果在维度上保持一致，使用 $T_i(\cdot)$对专家输出的结果进行维度调整。聚合函数具体公式如(8)所示。

$$X_{MoE\_out}=\sum_{i=1}^{N_e}\mathcal{L}\left(\bar{R}\left(X_{emb}\right)_i>0\right)R\left(X_{emb}\right)_i T_i\left(X^{i}_{\exp\_out}\right) \tag{8}$$

其中，$\mathcal{L}\left(\bar{R}\left(X_{emb}\right)_i>0\right)$表示当 $\bar{R}\left(X_{emb}\right)_i>0$ 时输出 1，否则输出 0。

专家模块内集成了小波变换技术和时间卷积网络(Temporal Convolutional Network, TCN)。如图 1 所示，$X^{i}_{\mathrm{expert}\_in} \in \mathbb{R}^{B_i\times N\times d_m}$ 表示第 i 个专家模块的输入作为小波变换的输入，通过一维小波变换(1-D Discrete Wavelet Transform, DWT)被分解为低频分量 $L \in \mathbb{R}^{B_i\times N\times \frac{d_m}{2}}$ 和高频分量 $H \in \mathbb{R}^{B_i\times N\times \frac{d_m}{2}}$。每个分量再利用时间卷积网络进行特征提取，其中扩张卷积的引入使得模型能够同时处理不同时间尺度上的信息，其公式表示如式(9)所示。提取的高频特征和低频特征经过全局平均池化后，通过全连接和拼接进行特征融合。具

体公式如下所示：

$$H_{\mathrm{GAP}}^{l}, L_{\mathrm{GAP}}^{l} = \mathrm{TCN}\left(\mathrm{DWT}\left(X_{\mathrm{expert\_in}}\right)\right) \tag{9}$$

$$H_{\mathrm{GAP}}^{l}, L_{\mathrm{GAP}}^{l} = \mathrm{TCN}\left(\mathrm{DWT}\left(X_{\mathrm{expert\_in}}\right)\right) \tag{10}$$

3.3. 编码器模块

本文提出的编码器模块由两个选择性状态空间 Mamba 模型组成。如图 1 所示，编码器模块内部交替排列层归一化模块(Layer Norm, LN)、Mamba 模块以及前馈网络(Feed Forward, FW)。在编码器模块中，前馈网络[18]通过深度神经网络的映射能力，使得模型捕捉到输入序列中的长依赖关系；层归一化[19]将数据归一化为高斯分布，有效缓解了因输入数据分布差异而导致的训练难题。本文设计的多层编码器模块，假设其包含 L 层，第 l 层编码器模块的整体方程总结为 $X_{en}^{l} = \mathrm{Encoder}\left(X_{en}^{l-1}\right)$，详情如下：

$$S_{m}^{l} = \mathrm{Mamba}\left(\mathrm{LN}\left(X_{en}^{l-1}\right)\right) + \mathrm{Mamba}\left(\mathrm{LN}\left(X_{en\_reverse}^{l-1}\right)\right) \tag{11}$$

$$S_{out}^{l} = \mathrm{LN}\left(\mathrm{FF}\left(\mathrm{LN}\left(S_{m}^{l}\right)\right)\right) \tag{12}$$

其中，$X_{en}^{l} = S_{out}^{l-1}$，$l \in \{1, \cdots, L\}$ 表示第 l 层编码器的输出。$X_{en\_reverse}^{l-1}$ 表示将时间序列 X_{en}^{l} 进行了翻转。当 $l = 0$ 时，编码器的输入 X_{en}^{0} 是专家网络的输出 X_{MoE_out}。$S_{m}^{l}, S_{out}^{l} \in \mathbb{R}^{B \times N \times d_m}$ 分别指的是 Mamba 模块的输出，以及编码器模块的输出。其中 B 表示批量大小，N 表示每个时间步变量的数量，d_m 表示特征维度大小。

Mamba 模块采用了双分支并行处理结构，如图 2 所示。输入特征 $X_{in} \in \mathbb{R}^{B \times N \times d_m}$ 被分配到两个独立的处理路径中。在第一个分支中，输入特征 S_{attn} 首先通过线性变换层 Linear(·)进行维度扩展，扩展维度由扩展因子 α 决定。随后，扩展后的特征经深度卷积网络 Conv1d(·)处理。为进一步提升模型适应性，引入了选择性状态空间模型(Selective SSM)，其参数基于输入数据动态调整，从而增强模型对不同输入特征的响应能力。其动态调整机制如下：

$$X_{1} = \mathrm{Silu}\left(\mathrm{Linear}\left(X\right)\right) \tag{13}$$

$$X_{2} = \mathrm{SelectiveSSM}\left(\mathrm{Silu}\left(\mathrm{Conv1d}\left(\mathrm{Linear}\left(x\right)\right)\right)\right) \tag{14}$$

$$\mathrm{output} = \mathrm{Linear}\left(X_{1} \cdot X_{2}\right) \tag{15}$$

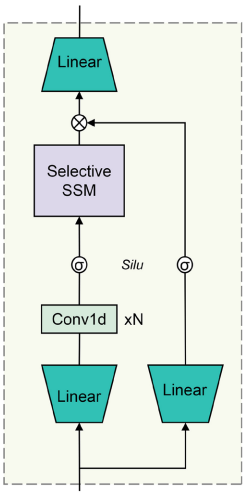

Figure 2. Structure of Mamba
图 2. Mamba 模型结构图

在第二个分支中，输入特征 S_{attn} 同样经过线性变换层 Linear(·)的扩展与 Silu(·)激活函数的非线性处理。两个分支的输出特征通过逐元素乘积融合，这一操作不仅保留了各分支的独特信息，还促进了特征间的互补与增强。最后，为了保持与输入特征维度的一致性，融合后的特征通过线性层 Linear(·)进行维度还原。

选择性状态空间模型(Selective SSM)通过动态调整状态空间模型参数矩阵，增强了对不同输入特征的响应能力。其核心是动态调整参数矩阵，使得模型可以对输入序列进行动态调整。Selective SSM 的状态转移方程如式(16)和式(17)所示，A_t, B_t, C_t, D_t 分别表示可根据输入特征动态调整的参数矩阵。

$$h_t = A_t h_{t-1} + B_t x_t \tag{16}$$

$$y_t = C_t h_t + D_t x_t \tag{17}$$

4. 实验结果

本小节将介绍实验中使用的数据集、实验细节、基线模型、评估指标，并通过在 4 个数据集(包括能源、天气和交通等)上进行实验来评估模型的性能。数据集相关参数描述如表 1 所示，实验相关设置描述如表 2 所示。

为了验证模型的性能，选择了五个时间序列预测基线模型进行比较：Informer、LogTrans、LSSL、TiDE。

Informer [11]：提出了 ProbSparse 自注意力机制，优化了注意力分布计算方式的同时有效降低了计算复杂度，通过独特的并行生成解码器机制显著提高了长时间序列预测的推理速度。

LogTrans [19]：引入卷积自注意力机制，利用因果卷积生成查询和键，从而更好地将局部上下文信息整合到注意力机制中。

LSSL [20]：提出了一种混合架构，将递归神经网络(RNN)、卷积网络(CNN)和连续时间模型与线性状态空间模型相结合，融合了各类模型的优势对时间序列数据进行预测。

TiDE [21]：采用密集的多层感知机(MLP)对过去的时间序列数据和协变量进行编码，并使用类似的 MLP 结构解码未来的时间序列数据和协变量。

Table 1. Description of datasets
表 1. 数据集的详细描述

数据集	变量数量	数据集划分	频率
Traffic	862	(12089, 1661, 3413)	1 小时
Weather	21	(3660, 5079, 10348)	10 分钟
Electricity	321	(18125, 2441, 5069)	1 小时
Exchange	8	(5120, 665, 1422)	1 天

Table 2. Experimental setup
表 2. 实验设置

设置项	具体值
训练设备	NVIDIA Quadro RTX 8000 GPUs
训练框架	PyTorch
优化器	ADAM
初始学习率	$1e^{-4}$
损失函数	L2 范式
训练 epochs	10
数据集训练输入长度	T = 96
数据集训练预测长度	F ∈ {96, 192, 336, 720}

指标(Metrics)：在时间序列预测实验中，选择了两个常用指标用于评估模型性能，分别是：平均绝对误差(MAE)和均方误差(MSE)，其公式如下所示：

$$\mathrm{MSE}\left(r,Y_f\right)=\frac{1}{n}\sum_{i=1}^{n}\left(r_i-Y_{f_i}\right)^2 \tag{18}$$

$$\mathrm{MAE}\left(r,Y_f\right)=\frac{1}{n}\sum_{i=1}^{n}\left|\left(r_i-Y_{f_i}\right)\right| \tag{19}$$

其中，n 表示样本数量，r_i 表示第 i 个样本的真实值，Y_{f_i} 表示第 i 个样本的预测值。

4.1. 时间序列预测结果

为评估本文提出的模型的性能优势，进行了多变量时间序列实验。结果如表 3 所示，其中加粗标注的为最优结果。预测指标 MSE/MAE 值越低，预测结果越准确。实验表明，本文模型在所有数据集上实现了最好的性能。

Table 3. Multivariate time series forecast result
表 3. 多变量时序预测结果

Models		**MSTCN**		Informer		LogTrans		LSSL		TiDE	
Metric		**MSE**	**MAE**	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity	96	**0.140**	**0.237**	0.274	0.368	0.258	0.357	0.300	0.392	0.237	0.329
	192	**0.160**	**0.266**	0.296	0.386	0.266	0.368	0.297	0.390	0.236	0.330
	336	**0.160**	**0.271**	0.300	0.394	0.280	0.380	0.317	0.403	0.249	0.344
	720	**0.203**	**0.289**	0.373	0.439	0.283	0.376	0.338	0.417	0.284	0.373
	Avg	**0.166**	**0.266**	0.311	0.397	0.272	0.370	0.313	0.401	0.251	0.344
Exchange	96	**0.084**	**0.200**	0.847	0.752	0.968	0.812	0.395	0.474	0.094	0.218
	192	**0.177**	**0.303**	1.204	0.895	1.040	0.851	0.776	0.698	0.184	0.307
	336	**0.329**	**0.418**	1.672	1.036	1.659	1.081	1.029	0.797	0.349	0.431
	720	**0.875**	**0.702**	2.478	1.310	1.941	1.127	2.283	1.222	0.852	0.698
	Avg	**0.366**	**0.406**	1.550	0.998	1.402	0.968	1.121	0.798	0.370	0.413
Traffic	96	**0.381**	**0.256**	0.719	0.391	0.684	0.384	0.798	0.436	0.805	0.493
	192	**0.388**	**0.267**	0.696	0.379	0.685	0.390	0.849	0.481	0.756	0.474
	336	**0.450**	**0.294**	0.777	0.420	0.734	0.408	0.828	0.476	0.762	0.477
	720	**0.472**	**0.235**	0.864	0.472	0.717	0.396	0.584	0.489	0.719	0.449
	Avg	**0.423**	**0.263**	0.764	0.416	0.705	0.395	0.832	0.471	0.760	0.473
Weather	96	**0.165**	**0.213**	0.300	0.384	0.458	0.490	0.174	0.252	0.202	0.261
	192	**0.211**	**0.252**	0.698	0.544	0.658	0.589	0.238	0.313	0.242	0.298
	336	**0.269**	**0.299**	0.578	0.523	0.797	0.652	0.287	0.355	0.287	0.335
	720	**0.350**	**0.353**	1.059	0.741	0.869	0.675	0.384	0.415	0.351	0.386
	Avg	**0.249**	**0.279**	0.634	0.548	0.696	0.602	0.271	0.334	0.271	0.320

在 Electricity 数据集上，本文模型的平均 MSE 达到了 0.166，较先前的最佳结果(0.193)提高了 14%。

在 Traffic 数据集上，本文提出的模型 MSE 平均值达到了 0.423，较先前最佳结果(0.617)提升了 31.4%。进一步分析表明，模型在中等预测长度(96、192 和 336)上表现尤为突出：在 Exchange 数据集中，预测长度为 96、192 和 336 时，较次优模型平均性能分别提升了 21.5%、21.7%、10.4%；在 Electricity 数据集中，相应预测长度下的平均性能分别提升了 16.7%、13.0%和 19.2%。然而，当预测长度延长至 720 时，性能提升幅度有所下降，分别为 9.2%和 7.7%，表明在极端长时预测任务中，本文设计的模型性能优势可能受到噪声的影响。总体而言，实验结果验证了本文所设计的模型在处理不同周期特征的时间序列数据时具有鲁棒性和优越性。

4.2. 模型消融实验结果

为了验证本文提出的各个模块对时间序列预测性能提升的有效性，本节做了消融实验，将 MSTCN 拆分为以下三种变体：MSTCN w/o MoE、MSTCN w/o TCN、MSTCN Replace BM with Mamba，实验结果如表 4 所示。其中 MSTCN w/o MoE 表示从 MSTCN 中移除混合专家网络(MoE)，验证了其对小波变换分解层数自适应调整的帮助。MSTCN w/o TCN 表示从 MSTCN 中移除时间卷积网络(TCN)，评估了时间卷积网络在多尺度提取特征中对预测结果的贡献。MSTCN Replace BM with Mamba 表示将 MSTCN 中编码器里的双向 Mamba 模块(BM)替换成单一 Mamba 模块，以验证双向建模方法在时序依赖关系建模中的有效性。

从表 4 展示的实验结果来看，MSTCN w/o MoE 展现出了最明显的性能提升，这验证了专家网络能够灵活地根据时间序列数据集的季节趋势特征动态选择合适的小波变换分解层数，从而增强模型的适应性。此外，小波变换与时间卷积网络(TCN)的结合有效捕获了时间序列的多尺度特征，实现了对多周期特征的提取。最后，从 MSTCN Replace BM with Mamba 的实验结果中可以看出双向 Mamba 模块通过同时建模历史与未来信息，在捕捉时间序列内部依赖关系方面发挥了关键作用。总的来说，各个模块都发挥了其各自的优势，提升了模型的整体预测性能。

Table 4. Ablation studies
表 4. 模型消融实验结果

Models		**MSTCN**		MSTCN w/o MoE		MSTCN w/o TCN		MSTCN replace BM with Mamba	
Metric		**MSE**	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity	96	**0.140**	**0.237**	0.151	0.249	0.146	0.243	0.147	0.244
	192	**0.160**	**0.266**	0.169	0.273	0.164	0.270	0.163	0.270
	336	**0.160**	**0.271**	0.166	0.279	0.163	0.274	0.165	0.274
	720	**0.203**	**0.289**	0.207	0.294	0.205	0.290	0.204	0.291
Weather	96	**0.165**	**0.213**	0.178	0.223	0.202	0.220	0.171	0.217
	192	**0.165**	**0.252**	0.219	0.259	0.219	0.260	0.215	0.255
	336	**0.269**	**0.299**	0.273	0.304	0.273	0.302	0.274	0.302
	720	**0.350**	**0.353**	0.353	0.357	0.351	0.355	0.353	0.353

4.3. 时间序列预测可视化结果

图 3 展示了模型在 Electricity 数据集上预测结果的可视化。其中，四张子图分别对应预测长度为 96、192、336、720 个时间步长的实验结果，横向坐标轴表示时间步长，纵向坐标轴的单位为千瓦时。可以看

出，预测曲线(Prediction)与真实值(Ground Truth)曲线高度吻合，充分证明了模型在预测任务中的高效性能。

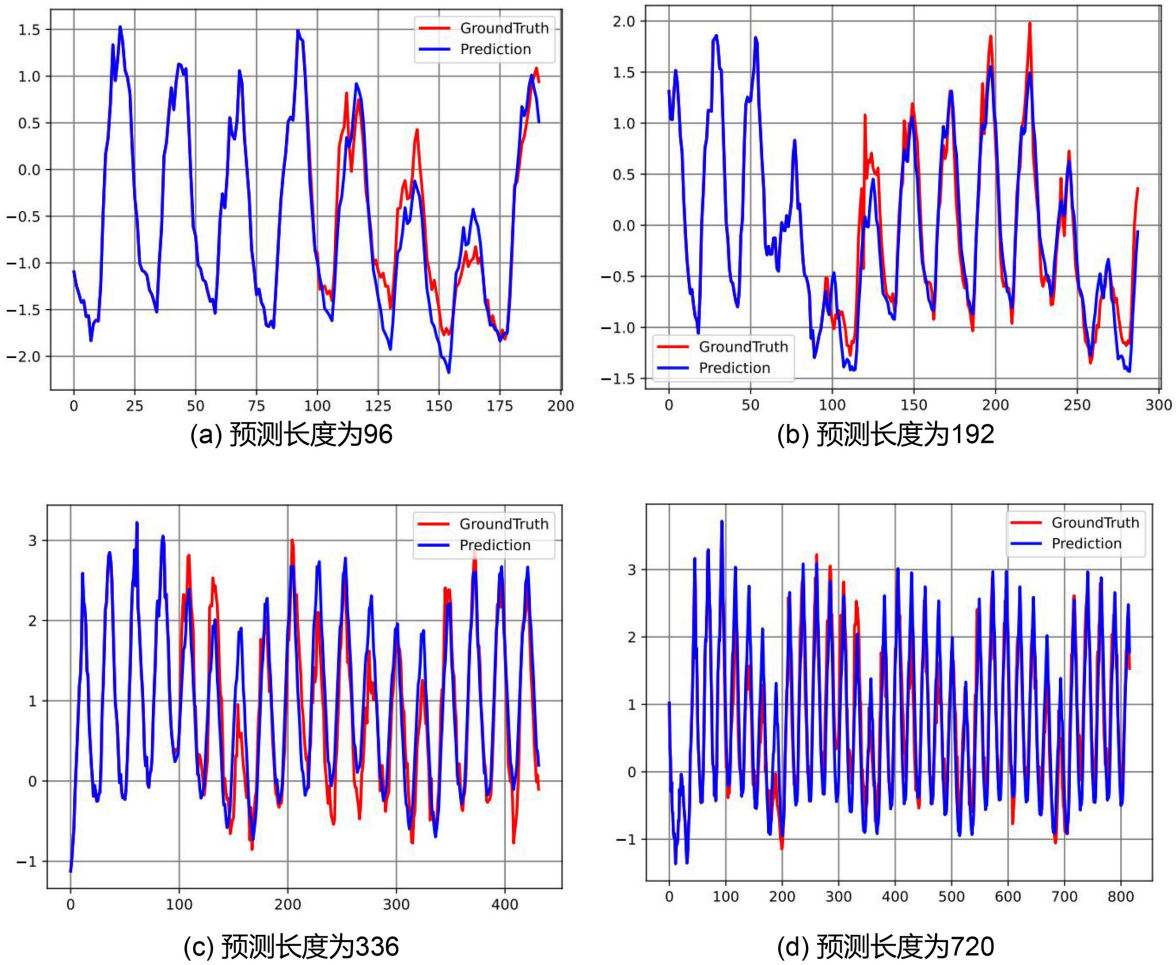

Figure 3. Visualization results on electricity dataset
图 3. Electricity 数据集的预测结果可视化

5. 结论

本文提出了一种创新的时序预测模型，融合了专家混合网络的灵活性、小波变换和时间卷积网络的多尺度特征提取能力，以及选择性状态空间模型的分析优势。通过专家混合网络，模型能够根据不同时间序列的特性动态分配不同的专家来处理。在周期特征提取方面，首先利用了小波变换时序数据分解为高频与低频分量，再通过时间卷积网络实现多尺度特征提取，最后使用 Mamba 模型捕捉数据中的长期依赖性。实验结果表明，模型在多变量时间序列预测任务上表现出色，与基线模型相比，在所有数据集上均取得了最优结果。未来的研究方向包括探索选择性状态空间模型与注意力机制的结合，以进一步提升模型性能。

参考文献

[1] Fjellstrom, C. (2022) Long Short-Term Memory Neural Network for Financial Time Series. 2022 *IEEE International Conference on Big Data* (*Big Data*), Osaka, 17-20 December 2022, 3496-3504.

https://doi.org/10.1109/bigdata55660.2022.10020784

[2] Angryk, R.A., Martens, P.C., Aydin, B., Kempton, D., Mahajan, S.S., Basodi, S., *et al*. (2020) Multivariate Time Series Dataset for Space Weather Data Analytics. *Scientific Data*, **7**, Article No. 227. https://doi.org/10.1038/s41597-020-0548-x

[3] Chen, C., Petty, K., Skabardonis, A., Varaiya, P. and Jia, Z. (2001) Freeway Performance Measurement System: Mining Loop Detector Data. *Transportation Research Record*: *Journal of the Transportation Research Board*, **1748**, 96-102. https://doi.org/10.3141/1748-12

[4] Lai, G., Chang, W., Yang, Y. and Liu, H. (2018) Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks. *The* 41*st International ACM SIGIR Conference on Research & Development in Information Retrieval*, Ann Arbor, 8-12 July 2018, 95-104. https://doi.org/10.1145/3209978.3210006

[5] Zaremba, W., Sutskever, I. and Vinyals, O. (2014) Recurrent Neural Network Regularization. arXiv: 1409.2329. https://doi.org/10.48550/arXiv.1409.2329

[6] Zhao, Z., Chen, W., Wu, X., Chen, P.C.Y. and Liu, J. (2017) LSTM Network: A Deep Learning Approach for Short-term Traffic Forecast. *IET Intelligent Transport Systems*, **11**, 68-75. https://doi.org/10.1049/iet-its.2016.0208

[7] Elfeky, M.G., Aref, W.G. and Elmagarmid, A.K. (2005) Periodicity Detection in Time Series Databases. *IEEE Transactions on Knowledge and Data Engineering*, **17**, 875-887. https://doi.org/10.1109/tkde.2005.114

[8] Pfander, G.E. and Benedetto, J.J. (2002) Periodic Wavelet Transforms and Periodicity Detection. *SIAM Journal on Applied Mathematics*, **62**, 1329-1368. https://doi.org/10.1137/s0036139900379638

[9] Wu, H., Xu, J., Wang, J. and Long, M. (2021) Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. *Advances in Neural Information Processing Systems*, **34**, 22419-22430.

[10] Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L. and Jin, R. (2022) FEDformer: Frequency Enhanced Decomposed Transformer for Long-Term Series Forecasting. *Proceedings of the* 39*th International Conference on Machine Learning*, Baltimore, 17-23 July 2022, 27268-27286.

[11] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., *et al*. (2021) Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 11106-11115. https://doi.org/10.1609/aaai.v35i12.17325

[12] Gu, A. and Dao, T. (2023) Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv: 2312.00752. https://doi.org/10.48550/arXiv.2312.00752

[13] Mallat, S.G. (1989) A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 674-693. https://doi.org/10.1109/34.192463

[14] Hewage, P., Behera, A., Trovati, M., Pereira, E., Ghahremani, M., Palmieri, F., *et al*. (2020) Temporal Convolutional Neural (TCN) Network for an Effective Weather Forecasting Using Time-Series Data from the Local Weather Station. *Soft Computing*, **24**, 16453-16482. https://doi.org/10.1007/s00500-020-04954-0

[15] Ioffe, S. and Szegedy, C. (2015) Batch Normalization: Accelerating Deep Network Training by Reducing Internal Co-Variate Shift. *Proceedings of the* 32*nd International Conference on International Conference on Machine Learning*, Lille, 6-11 July 2015, 448-456.

[16] Ba, J.L., Kiros, J.R. and Hinton, G.E. (2016) Layer Normalization. arXiv: 1607.06450. https://doi.org/10.48550/arXiv.1607.06450

[17] Li, Z., Qi, S., Li, Y. and Xu, Z. (2023) Revisiting Long-Term Time Series Forecasting: An Investigation on Linear Mapping. arXiv:2305.10721. https://doi.org/10.48550/arXiv.2305.10721

[18] Bebis, G. and Georgiopoulos, M. (1994) Feed-Forward Neural Networks. *IEEE Potentials*, **13**, 27-31. https://doi.org/10.1109/45.329294

[19] Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.X. and Yan, X. (2019) Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting. *Proceedings of the* 33*rd International Conference on Neural Information Processing Systems*, Vancouver, 8-14 December 2019, 5243-5253.

[20] Gu, A., Johnson, I., Goel, K., Saab, K., Dao, T., Rudra, A. and Ré, C. (2021) Combining Recurrent, Convolutional, and Continuous-Time Models with Linear State Space Layers. *Proceedings of the* 35*th International Conference on Neural Information Processing Systems*, Online, 6-14 December 2021, 572-585.

[21] Das, A., Kong, W., Leach, A., Mathur, S., Sen, R. and Yu, R. (2023) Long-Term Forecasting with Tide: Time-Series Dense Encoder. arXiv: 2304.08424. https://doi.org/10.48550/arXiv.2304.08424