

基于Transformer的CERN生态观测时序数据通用预测模型研究

郭学兵^{1,3}, 程翔², 唐新斋¹, 张黎^{1,3,4}, 何洪林^{1,3,4}

1. 中国科学院地理科学与资源研究所生态系统网络观测与模拟重点实验室, 北京 100101;
2. 中国科学院自动化研究所, 北京 100190;
3. 国家生态科学数据中心, 北京 100101;
4. 中国科学院大学资源与环境学院, 北京 100190

摘要

中国生态系统研究网络 (Chinese Ecosystem Research Network, CERN) 积累了多站点、多变量、时间跨度超20年的生态系统原位时序观测数据。挖掘利用这些数据开展时序预测建模, 可应用于站点数据质控分析、数据插补、实时预警等场景。选取其中两个站点6年的气象、太阳辐射及土壤温湿盐度共36个变量的小时级数据, 通过Transformer模型在时序窗口长度、全输入变量自注意力与变量可预测性、多变量与单变量输出模式、多站点融合等方面的研究, 实现了120 h长窗口多变量输入, 对后续24 h多变量的准确预测, 绝大多数可预测变量归一化后的均方根误差介于0.002与0.04之间, 决定系数为0.9以上, 为建立数据驱动的多站点、多变量时序数据通用预测模型提供了性能优异的新路径。

关键词

Transformer模型; 生态观测; 中国生态系统研究网络; 长窗口; 多站点; 多变量; 时序数据预测

中图分类号: Q149; TP183

文献标志码: A

doi: 10.11959/j. issn. 2096-0271.2026046

Research on a general prediction model for CERN ecological observational time series data based on Transformer

Guo Xuebing^{1,3}, Cheng Xiang², Tang Xinzhai¹, Zhang Li^{1,3,4}, He Honglin^{1,3,4}

1. Key Laboratory of Ecosystem Network Observation and Modeling, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China;
2. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China;
3. National Ecosystem Science Data Center, Beijing 100101, China;
4. College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100190, China

Abstract

The Chinese Ecosystem Research Network (CERN) has accumulated a massive amount of ecosystem in-situ time-series data from multiple sites with multiple variables, spanning over 20 years. Mining and utilizing these data to conduct

time-series predictive modeling was applied in scenarios such as data quality control analysis, data interpolation, and real-time pre-warning at the stations. Hourly data spanning six years, involving 36 observational variables including meteorological, solar radiation, soil temperature, soil moisture and soil salinity of CERN's two stations were selected. Based on the Transformer model, studies on various time window width, self-attention mechanism towards all input variables, diverse variables predictability and comparing the model performance of multivariate output with univariate output were conducted, as well as fusion modeling of multiple sites' data. The accurate prediction of meteorological-radiation-soil multivariate and multi-site which used 120-hour-width time window as input and multivariate output for the next 24 hours on hourly basis were realized, with root mean square error (RMSE) ranging between 0.002 and 0.04 and coefficient of determination (R^2) greater than 0.9 for most variables after normalization. This demonstrates that the Transformer model has enormous potential and pioneers a new data-driven approach to construct a general prediction model that involves multiple variables and multiple sites simultaneously.

Key words

Transformer model, ecological observation, China Ecosystem Research Network, long window, multi-site, multivariate, time series data prediction

0 引言

复杂时序预测和建模广泛存在于众多领域,也是生态系统领域最为基础性的问题。基于模型参数估计的传统方法严重依赖特定领域的专家知识,且要求研究人员对过程原理进行深刻理解与精准抽象^[1-2]。随着过程机理和数据特征的复杂程度增加,机器学习方法逐渐被引入时间序列数据的预测研究中^[3]。例如,齐焕东等^[4]提出一种基于规则集和多层感知机的质量控制方法,并对比分析了13种机器学习模型在南海区域散点温盐深数据集质控方面的性能,然而,该方法前期需要大量的数据标注工作,导致数据集的准备过程极为困难。近十年来,以神经网络为核心的深度学习技术在时间序列分析领域的应用日益广泛,特别是长短期记忆(long short-term memory, LSTM)网络在时序数据预测方面得到了深入的研究^[5-8]。例如,基于LSTM模型,李亚蒙等^[9]对全球范围、长期开放的海洋观测站点所在海区各季节性

代表月份的海面风速进行了24 h短期预测;席小青等^[10]利用美国夏威夷的气温、相对湿度、10 min平均风速等变量预测了太阳辐照度数据;赵书强等^[11]采用美国可再生能源实验室个别站点2007年日间每5 min内地表太阳辐照度、气溶胶光学厚度、相对湿度、站点平均风速、最大风速等作为模型输入,进行地表太阳辐照度的短期预测。

上述研究从多个侧面对时间序列预测模型的技术方法进行了验证、对比与分析,为本文的研究工作提供了宝贵的参考。然而,经过综合分析发现,这类研究仍存在诸多局限或不足:一是多数研究采用国际开源数据作为数据源,对数据的准备与分析偏少;二是研究大多局限于不同模型间预测效果的比较,数据-模型未进行本地化处理,导致模型对我国实际应用场景的适用性存在不确定性;三是大多数研究工作都是基于关联观测输入条件下的单一观测测量进行预测,在实际使用中遇到多变量预测时则需要部署大量这样的模型。

近年来人工智能模型层出不穷,Transformer不仅在自然语言处理和计算

机视觉等领域表现优异，学者们还利用其有效捕获长期依赖关系的能力，推动了时序数据预测任务的飞速发展^[12]。例如，在中期全球天气预报的复杂模式识别方面，华为盘古气象大模型展现出了良好效果，该模型基于Transformer的三维深度网络构建，较LSTM和卷积神经网络具有显著优势。该研究成果发表于Nature期刊^[13-14]。

对中国生态系统研究网络（Chinese Ecosystem Research Network, CERN）历史数据质量进行评价分析、对未来数据进行实时预测、为“双碳”等重大国家需求开展模拟分析，是CERN重要且紧迫的业务需求，亟须通过AI新技术来赋能。本文以CERN时间序列数据为研究对象，开展Transformer模型在生态领域时间序列数据预测方面的应用方法及技术研究，验证其有效性与适用性，试图建立一个能支持多站点、多变量、长窗口预测的通用预测模型。本文旨在为CERN未来实现长期观测数据（基础观测数据、加工数据）的高效实时预测与模拟分析，并最终达到实用化、业务化目标，开展前期探索性研究。

1 研究数据、任务与基础模型

1.1 数据来源与数据处理

CERN 经过 20 余年系统化、规范化、高质量的长期联网观测，积累了丰富的生态系统定位观测数据，涵盖农田、森林、草地等 10 多个典型生态系统类型，以及 40 余个生态站点的水分、土壤、气象、生物要素等数百个指标^[15-18]，这些指标的获取方式、观测频率各异。其中，自动观测设备（自动气象观测站、土壤温湿盐度时域反射法观测系统等）所采集的数据是其

中一个重要类别，这些数据呈现多生态站点、多变量、高频率（小时级）特征，且具有长时序、实时性强、相关性复杂等多源异构大数据的特征，是CERN最具特色的数据资源。

结合生态气象领域专业知识，本文选取了两个生态站点的重要生态地面观测数据作为研究对象，具体为青海海北高寒草地生态系统国家野外科学观测研究站（简称“海北站”，代码 HBG）和陕西长武农田生态系统国家野外科学观测研究站（简称“长武站”，代码 CWA）2018—2023 年逐小时的气象、太阳辐射^[19]和土壤温湿盐度观测数据。其中，气象数据包括气温 T1、大气压 P1、空气相对湿度 RH1、降雨 R1、地表温度 Tg01、风速 W10av1 的小时记录值；太阳辐射数据包括 3 种小时瞬时值（太阳辐照度 Eg_1、紫外辐射辐照度 UV_1 和净辐射辐照度 Estar_1）以及 3 种小时累积值（总辐射总量曝辐量 Eg_2、紫外辐射总量曝辐量 UV_2 和光合有效辐射总量通量密度 Par_2）；土壤温湿盐度数据包括多个剖面（10 cm、20 cm、30 cm、50 cm、80 cm、100 cm）的介电常数 RD、土壤水分 SM、电导率 EC，土壤温度 ST。此外，包括太阳高度角（Hs）数据，它是影响太阳辐射强度的重要因素。

由于缺失数据会影响预测模型的有效训练样本数，进而降低模型的预测性能，因此数据治理的重要内容之一是对缺失情况，甚至是缺失数据分布情况进行统计与分析。本文通过数据库表查询检索，对两个站 6 年的逐小时气象指标进行了详细统计。统计结果表明，海北站的 Tg01 数据在 2021—2022 年缺失严重（分别缺失大约半年数据），直接使用会导致有效样本严重减少。采用皮尔逊相关系数分析指标间相关性发现，Tg01 与 T1 具有极强的相关

性 (0.903)。因此, 本文弃用 Tg01, 留用 T1。最终, 作为 Transformer 训练测试数据源的数据包含气象辐射指标共计 (12×52 584) h、土壤数据共计 (24×52 584) h。

进一步, 本文采用 Min-Max 归一化策略对保留的数据进行处理:

$$s_v^{\text{norm}} = \frac{s_v - n_v}{m_v - n_v}$$

(1)

其中, m_v 为大于等于特征变量 $s_v \in \mathbb{R}$ 最大值的预设上限, n_v 为小于等于特征变量 s_v 最小值的预设下限。各特征变量的归一化参数见表1。

1.2 多变量长时序预测任务

为验证深度学习模型对无监督时序数

据中多变量复杂动力学系统的表征能力, 本文以自回归任务框架为基础构建预测模型。该框架属于序列数据建模的核心方法, 其本质是通过历史观测序列的时空依赖关系实现未来状态的概率推断。从统计学习视角来看, 自回归任务将单步预测的联合概率分布 $P(s_1, s_2, \cdots, s_T)$ 分解为条件概率的链式乘积:

$$P(s_1, s_2, \cdots, s_T) = P(s_T | s_{<T}) \prod_{t=1}^{T-1} P(s_t | s_{<t})$$

(2)

其中, s_t 为 t 时刻的观测值, $s_{<t}$ 为 t 时刻之前的所有观测值。通过最大化目标序列的似然函数, 自回归模型可学习到数据生成过程的潜在动力学规律, 这种范式已被广泛应用于气候预测、股票波动建模等复杂系统中^[20-21]。经典自回归模型通常采用线性假设构建预测函数:

表1 各特征变量的归一化参数

特征变量	预设上限 m_v	预设下限 n_v
总辐射辐照度/(W·m ⁻²)	1 420	-15
紫外辐射辐照度/(W·m ⁻²)	80	0
净辐射辐照度/(W·m ⁻²)	1 000	-250
总辐射总量曝辐量/(MJ·m ⁻²)	4.2	-0.3
紫外辐射总量曝辐量/(MJ·m ⁻²)	0.3	0
光合有效辐射总量通量密度/(mol·m ⁻² ·h ⁻¹)	7.5	-4
太阳高度角是否大于5°	1	0
风速/(m·s ⁻¹)	17	0
降水/mm	65	0
大气压/hPa	950	650
相对湿度	100%	0
气温/°C	45	-35
介电常数实部	35	0
土壤体积含水量/(cm ³ ·cm ⁻³)	0.5	0
电导率/(S·m ⁻¹)	0.15	0
土壤温度/°C	40	-15

$$s_T = \sum_{i=1}^{T-1} W_i s_i + b_i \tag{3}$$

其中， W_i 为可学习参数矩阵， b_i 为白高斯噪声。该类模型虽具有解析解和可解释性优势，但处理气象数据时在捕捉非线性关系、建模长时程依赖、耦合多变量表征、预测极端事件等方面存在能力不足。深度自回归模型通过引入非线性变换层来突破上述限制，通过误差反向传播这一高效的学习算法，这类自回归模型表现出比经典自回归模型更强的表征能力。

本文在训练测试样本构建阶段采用滑动窗口法：以若干小时连续数据作为输入序列，其后续的24 h连续数据作为预测目标，完整保留了气象辐射和土壤变量的时间连续性。基于深度神经网络的特性，在实现长时序预测时，本文采用序列到序列的模式，同时生成目标序列的各个位置，而非传统自回归模型在长时序预测时采用的递归生成方法^[22]。

各生态观测变量的时间尺度相同（小时级），且时间戳完全对齐。考虑Transformer模型具有自注意力机制特性，本文仅对输入变量的观测数据以简单展平方式处理，未将时间信息显式编码并作为输入网络的单独通道^[23]。同时，由于长武站与海北站的气压数据波动范围（分别为860~905 hPa、670~705 hPa）完全无交叉，该特异性生态变量已在一定程度上起到了类似站点信息嵌入的作用。

由于数据有一定的残缺，不同长度的输入窗口下Transformer模型可以利用的样本数见表2。

1.3 模型结构与评价指标

Transformer模型是一种基于自注意力机制的深度学习架构，它突破了传统神经网络结构的局限，凭借长程依赖建模和

表2 不同长度的输入窗口下Transformer模型可以利用的样本数

输入窗口 长度	训练样本数(70%)/组		测试样本数(30%)/组	
	长武站	海北站	长武站	海北站
24 h	27 235	34 358	11 694	14 757
120 h	25 126	31 328	10 785	13 448
240 h	22 837	28 523	9 801	12 243

并行化计算能力，成为人工智能领域的革命性技术。完整的Transformer模型采用编码器-解码器的对称架构，其中编码器采用多头注意力层处理输入序列，解码器则采用掩码自注意力生成输出序列^[24]。

本文使用的多变量时间序列表征学习框架由两部分组成：一个堆叠三层Transformer编码器的主干特征提取模块、一个线性层输出头部。以输入窗口长度为120 h，输出窗口长度为24 h的情况为例，此时输入的维度为120×36。其中，时间维度的长度为120，对应逐小时的采样方式；特征维度的长度为36，包含12个气象辐射特征和6个深度下的4个土壤特征。120 h输入窗口、24 h输出窗口下的网络结构如图1所示，其中，特征提取模块的参数量与输入的特征变量数有关，输出头部的参数量与输入的时间维度数有关。Transformer模型的参数量（120 h输入、24 h输出、36维特征）见表3，总计参数量为13.7M，丢弃率为0。

本文使用的表征学习框架以均方根误差（root mean square error，RMSE）作为损失函数和评价指标（以归一化后的数据计算），并使用误差反向传播算法优化模型权重。RMSE是一种常用的测量数据间差异的指标，计算式为：

$$RMSE = \sqrt{\frac{1}{v} \sum_{i=1}^v (y_i - \hat{y}_i)^2} \tag{4}$$

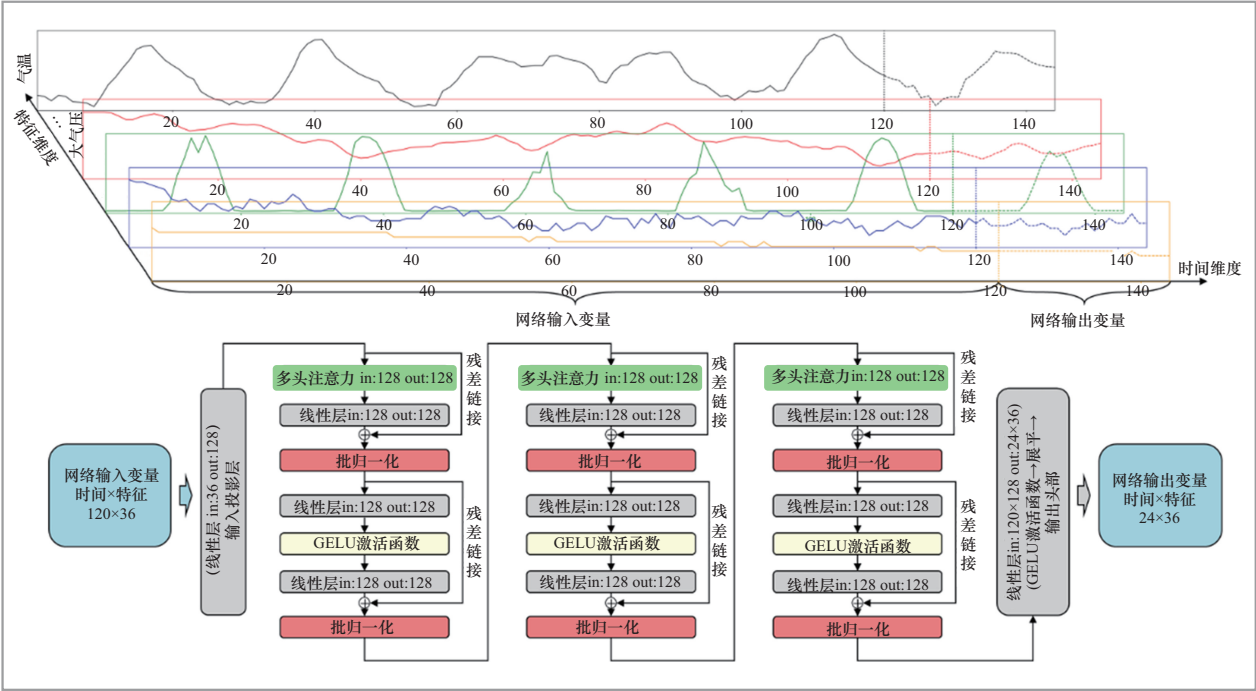


图1 120 h输入窗口、24 h输出窗口下的网络结构

表3 Transformer模型的参数量(120 h输入、24 h输出、36 维特征)

Transformer 模型	输入维度	输出维度	参数量
输入投影层	36	128	4 736
3*编码层-多头注意力子层-键	128	128	49 536
3*编码层-多头注意力子层-查询	128	128	49 536
3*编码层-多头注意力子层-值	128	128	49 536
3*编码层-多头注意力子层-输出层	128	128	49 536
3*编码层-前馈网络子层-前馈层1	128	256	99 072
3*编码层-前馈网络子层-前馈层2	256	128	98 688
输出头部	120x128	24x36	13.3M

其中， y_i 为生态变量的真实值， $\hat{y}_i$ 为模型对应的预测值。RMSE 越低代表预测效果越好。

为了对生态变量的可预测性进行评估，本文使用决定系数 R^2 来衡量输入-输出变量间的因果关系，其计算式为：

$$R^2=1-\frac{\sum_{i=1}^v\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^v\left(y_i-\bar{y}\right)^2}$$

(5)

其中， $\bar{y}$ 为 y_i 的均值。输出变量的 R^2 越接近 1，表明模型对该输入变量具有越好的拟合效果。这意味着输出变量与输入变量

间的因果关系更强，输出变量更具有可预测性。

2 Transformer在多变量长
时序预测任务上的有效性研究

本节从时间维度和特征维度两方面验证具有注意力机制的Transformer模型在处理多变量时序数据时的优势。基于对Transformer模型强大时序预测及注意力机制等能力的理解，本节将单站点、全变量输入、全变量输出作为有效性基本模式。在此基础上，就不同时间窗口长度、当前生态变量范围内输入变量分组的预测效果，以及变量可预测性等进行比较和分析，为后续建立通用预测模型奠定实验基础。

2.1 Transformer模型的长时间序列
处理能力

多数生态变量具有昼夜周期性以及变化的时序滞后性，因此，研究窗口长度对生态变量长时序预测效果，具有重要意义。本文分别按24 h、120 h以及240 h设定输入窗口的长度，开展单站点、全变量输入、全变量输出的对比试验，在初步探究Transformer模型预测能力的同时，确定较优的实验设置。实验过程中，Transformer训练过程分别进行1 000个轮次，实验分别在5个不同的随机种子下各进行1次（下同）。不同长度输入窗口下辐射变量和气象变量预测效果对比如图2所示，不同长度输入窗口下土壤变量预测效果对比如图3所示。

图2展示了辐射变量和气象变量的预测结果。从辐射数据预测结果来看，输入窗口长度为120 h的模型在首尾2 h更具优势，输入窗口长度为24 h的模型在输出中段更具优势，这说明长窗口对具有昼夜周

期性的变量具有显著优势。从气象数据预测结果来看，输入窗口长度为120 h的模型在气压和风速指标上具有略微的预测优势。图3展示了50 cm、80 cm和100 cm深度的预测结果。从土壤数据预测结果来看，输入窗口长度为24 h的模型在除土壤体积含水量外的指标上具有预测优势。综上，输入窗口长度为240 h的模型预测效果落后于输入窗口长度为24 h和120 h的模型。相较于24 h输入下的模型，240 h输入下的模型参数量约为24 h输入下模型参数量的10倍，而训练数据量则减少了约16.15%，除周期性建模外，并没有本质性提升，因此，从计算效率以及性能角度来看，本文没有对240 h模型进行参数规模的优化。考虑输入窗口长度为24 h和120 h的模型在整体上具有相近的RMSE，本文后续实验选择120 h作为基础实验配置。

2.2 全输入特征的自主注意力能力

由于2.1节实验中的输入数据采用了展平方式，这可能引入变量间的复杂相关性导致模型性能下降的问题。为进一步探究弱相关性变量是否会影响模型预测质量，本节在输入窗口为120 h的数据上进行了输入变量为全变量或气象辐射变量的对比实验。基于合理性和灵敏反应输入对输出影响的需求，本节选择了对大气压、相对湿度和气温3个变量进行单变量输出预测。部分变量输入模式和全变量输入模式的预测结果对比如图4所示。

由图4可知，全变量输入的预测效果要普遍接近或好于只包含气象辐射输入的预测效果。这初步验证了Transformer模型可以在输入端引入相关性较低的变量，而不影响甚至获得更高预测准确性的假设，进一步证明了Transformer模型兼容更多

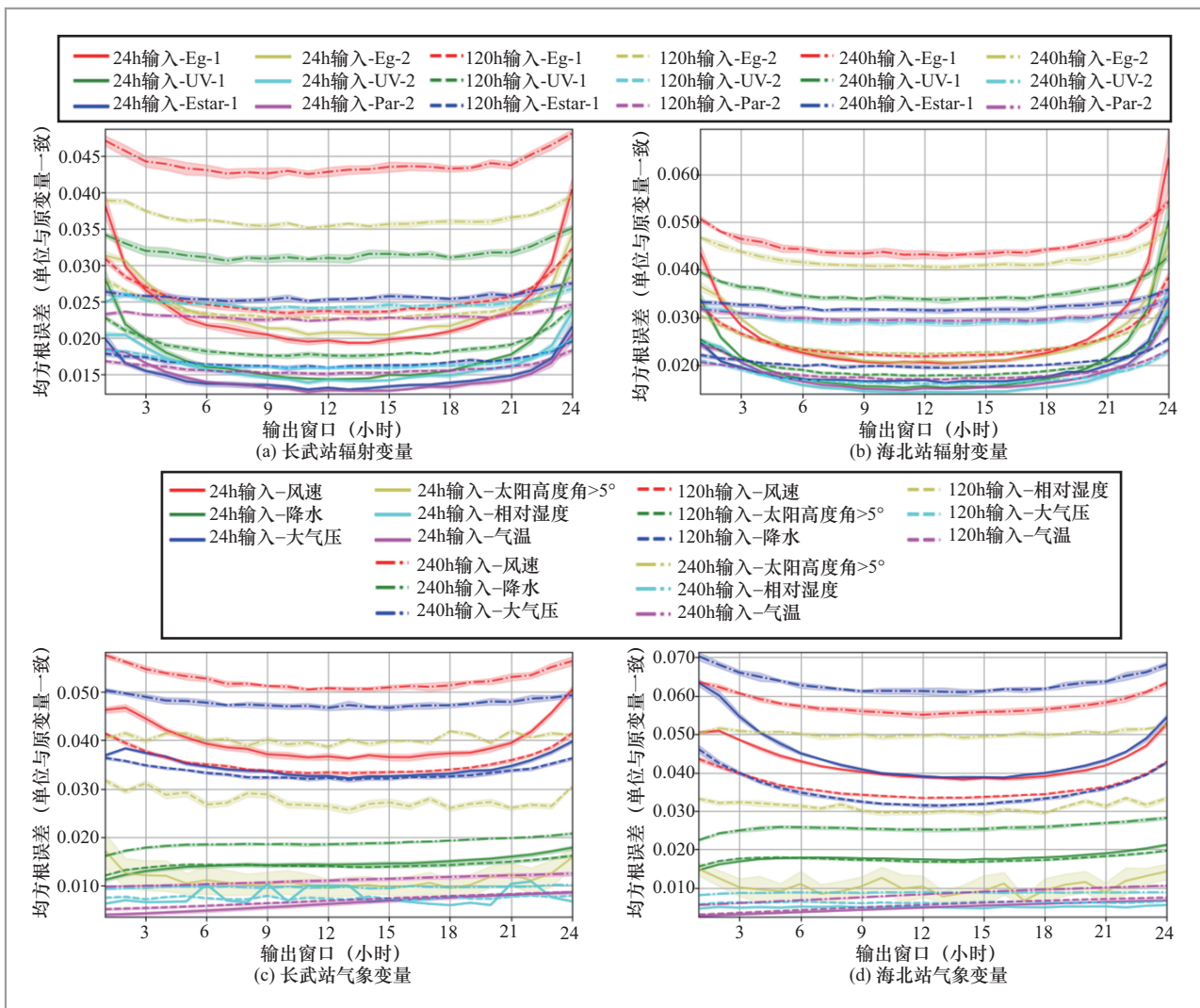


图2 不同长度输入窗口下辐射变量和气象变量预测效果对比

类型生态变量输入，不断迭代提升生态全变量预测效果的潜力。值得注意的是，Transformer这种全变量输入模式，使得其有能力兼容更多未知机理的弱关联和弱决定关系的生态变量类型，从而不断迭代提升模型对生态全变量的预测效果。

2.3 基于Transformer模型的变量可预测性分析

基于2.1节的预测模型，本节通过计算

120 h输入窗口下各生态变量的 R^2 对其可预测性进行分析。在两个站点上，6类辐射变量的 R^2 均为0.927以上，相对湿度($R^2 > 0.963$)和气温($R^2 > 0.977$)的预测值与实测值吻合度极高。预测大气压的 R^2 分别为0.822~0.937(长武站)和0.659~0.943(海北站)。在预测初期，大气压的预测表现较好但随延时增加快速降低，这与误差所反映的变化一致。对于预测风速的 R^2 ，虽然其在海北站有可以接受的预测结果(0.830~0.921)，但其在长武站的

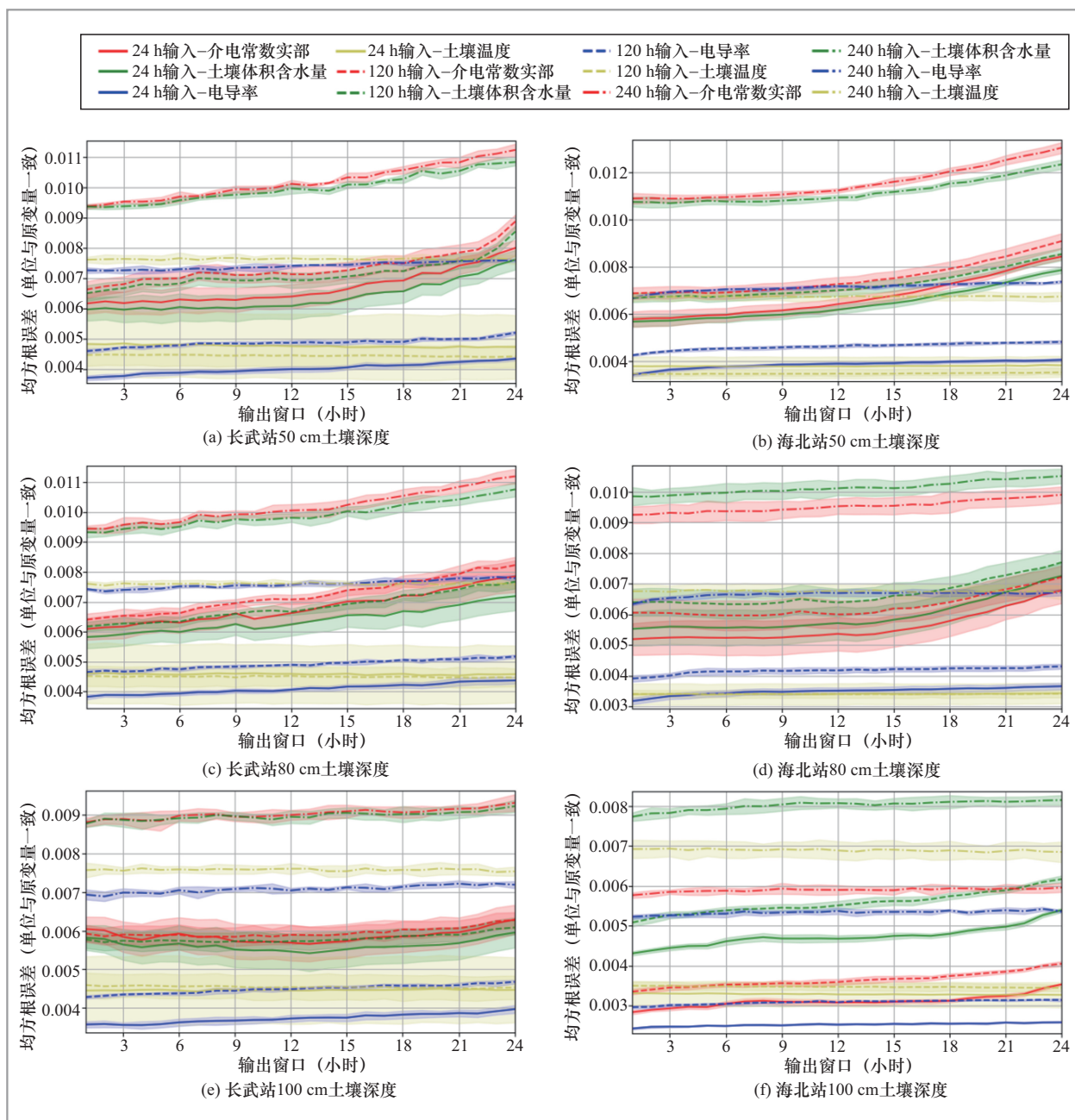


图3 不同长度输入窗口下土壤变量预测效果对比

预测结果较低 (0.705 ~ 0.782)。降水预测的 R^2 则全部为负值, 表明其预测效果甚至低于使用均值进行预测, 属于不可预测变量。对4类土壤变量, 模型表现出精确的预测效果 ($R^2 > 0.911$)。

根据模型预测的 R^2 判断, 在目前的输

入变量范围内, 风速和降水属于难以预测的变量。这与生态学中“这两个变量具有较高随机性, 目前的输入变量不足以对这两个变量进行预测”的观点一致。总体而言, 在缺乏先验机理的支撑下, Transformer模型可以从全变量中实现对

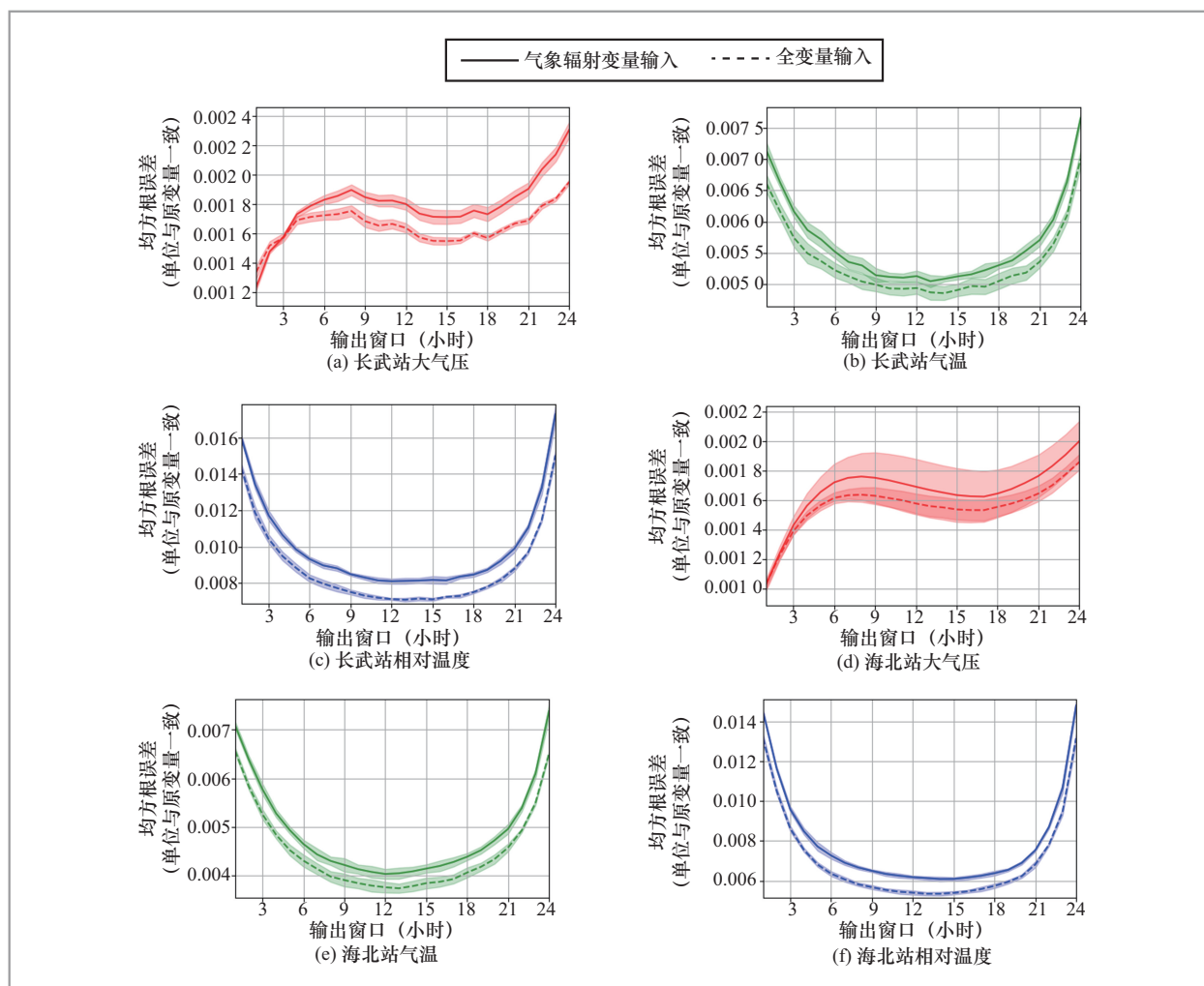


图4 部分变量输入模式和全变量输入模式的预测结果对比

可预测变量的自主筛选。

3 Transformer在多变量长时序预测任务上的通用性研究

本文第2节通过对比输入窗口长度(120 h)、变量可预测性(降水、风速难以预测)和输入模式(全变量输入),初步证明了Transformer模型在多变量长时序预测任务上的有效性。本节则进一步探讨分析输出模式、站点融合和泛化应用等问题,对模型的通用性预测进行研究。

3.1 输出变量模式对预测效果的影响

在输入窗口为120 h的全变量数据上,本节对比了两种输出模式下大气压、相对湿度和气温的单站点预测效果。第一种模式为大气压、相对湿度和气温的单变量输出预测;第二种模式为去掉降水风速等不可预测变量后的可预测变量输出预测。单变量输出模式和可预测变量输出模式的预测结果对比如图5所示。由图5可知,两种模式预测精度都在可用范畴, RMSE介于0.002与0.04之间。

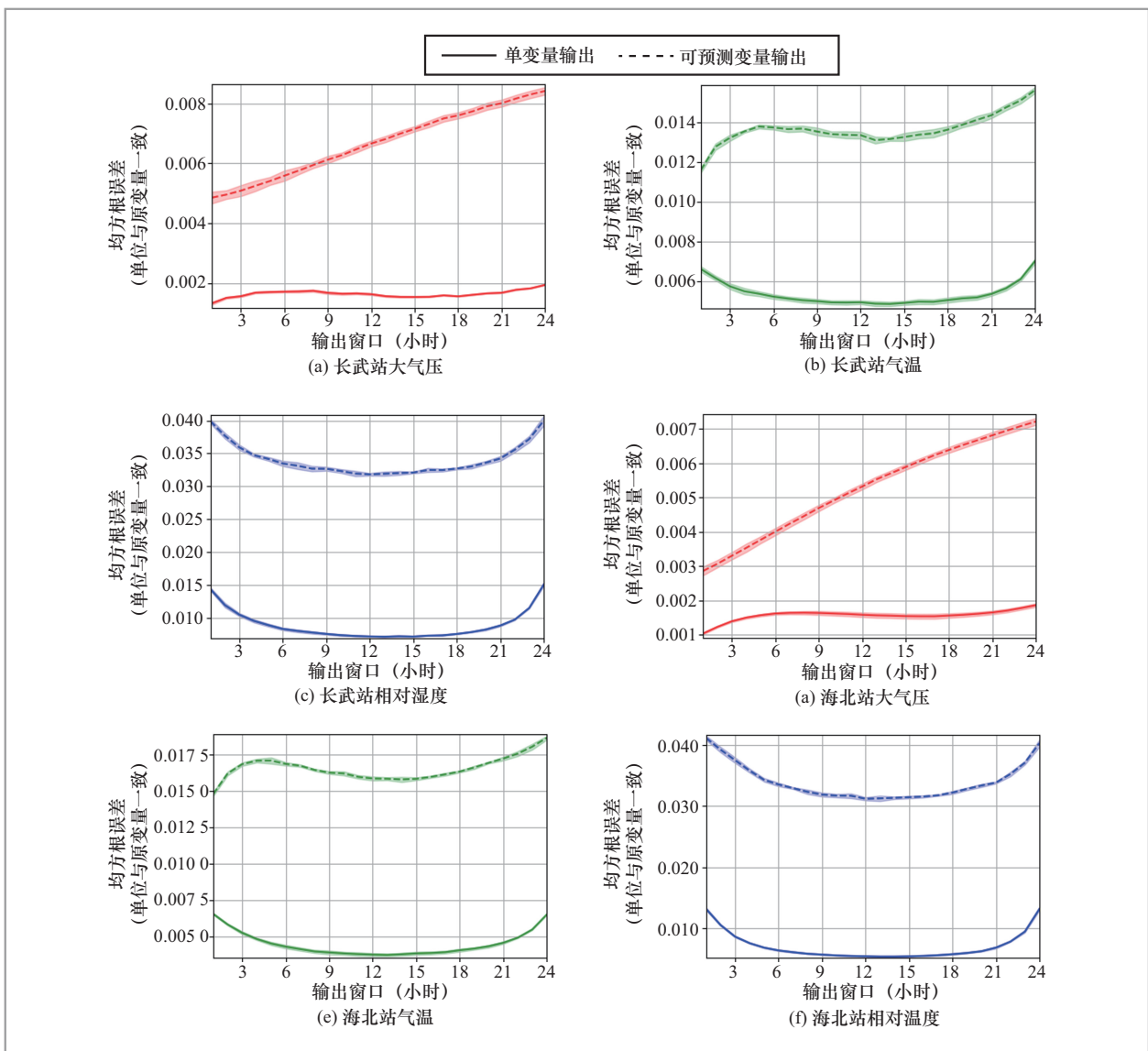


图5 单变量输出模式和可预测变量输出模式的预测结果对比

比较两种模式可以发现，单变量预测大气压、相对湿度和气温的RMSE低于可预测变量预测情况下的RMSE。这是因为Transformer模型在可预测变量预测的模式下，各个变量的误差均对损失函数产生贡献，并可能在特征提取层的参数学习中产生互相拮抗的影响。这一现象反映了模型在追求通用性和专用性时的矛盾，在实际应用时则可以根据对预测精度和系统复

杂度的需求，具体判断模型应该采用何种输出模式。

3.2 Transformer模型支持多站点预测建模

单个Transformer模型能否进行多站点建模，对于充分利用不同站点的数据，减少应用系统的复杂度具有很大意义。本

节讨论在输入窗口为 120 h 的数据上, 单个 Transformer 模型能否同时在双站点数据实现泛化应用。模型采用全变量输入模式和不包含降水风速的可预测变量输出模式。双站点数据集的训练集和测试集分别为单站点训练集和测试集的合并, 各包含 56 454 组和 24 233 组样本。单站点和双站点气象辐射变量预测效果对比如图 6 所示。

由图 6 可知, 双站点数据训练的模型对太阳高度角是否大于 5° 、相对湿度、气温以及 6 类辐射变量都具有比单站点数据训练的模型更好或接近的预测结果, 只有在大气压预测方面相对落后于单站点数据训练的模型。类似的结论出现在各深度的

土壤变量预测结果中, 单站点和双站点土壤变量预测效果对比如图 7 所示。图 7 为 50 cm、80 cm 和 100 cm 深度下的可视化结果, 唯一的特例是双站点数据训练的模型在海北站 100 cm 深度下的介电常数实部和电导率预测落后于单站点数据训练的模型。总体而言, 两种模式的预测精度都在可用范畴, RMSE 介于 0.002 与 0.04 之间。

不同气候类型下的生态观测数据可类比于深度学习中的“域”(domain)概念^[25-26]: 尽管各生态站(如黄土高原地区的长武站、青藏高原高寒草甸地区的海北站)的生态变量在数据结构上具有共性

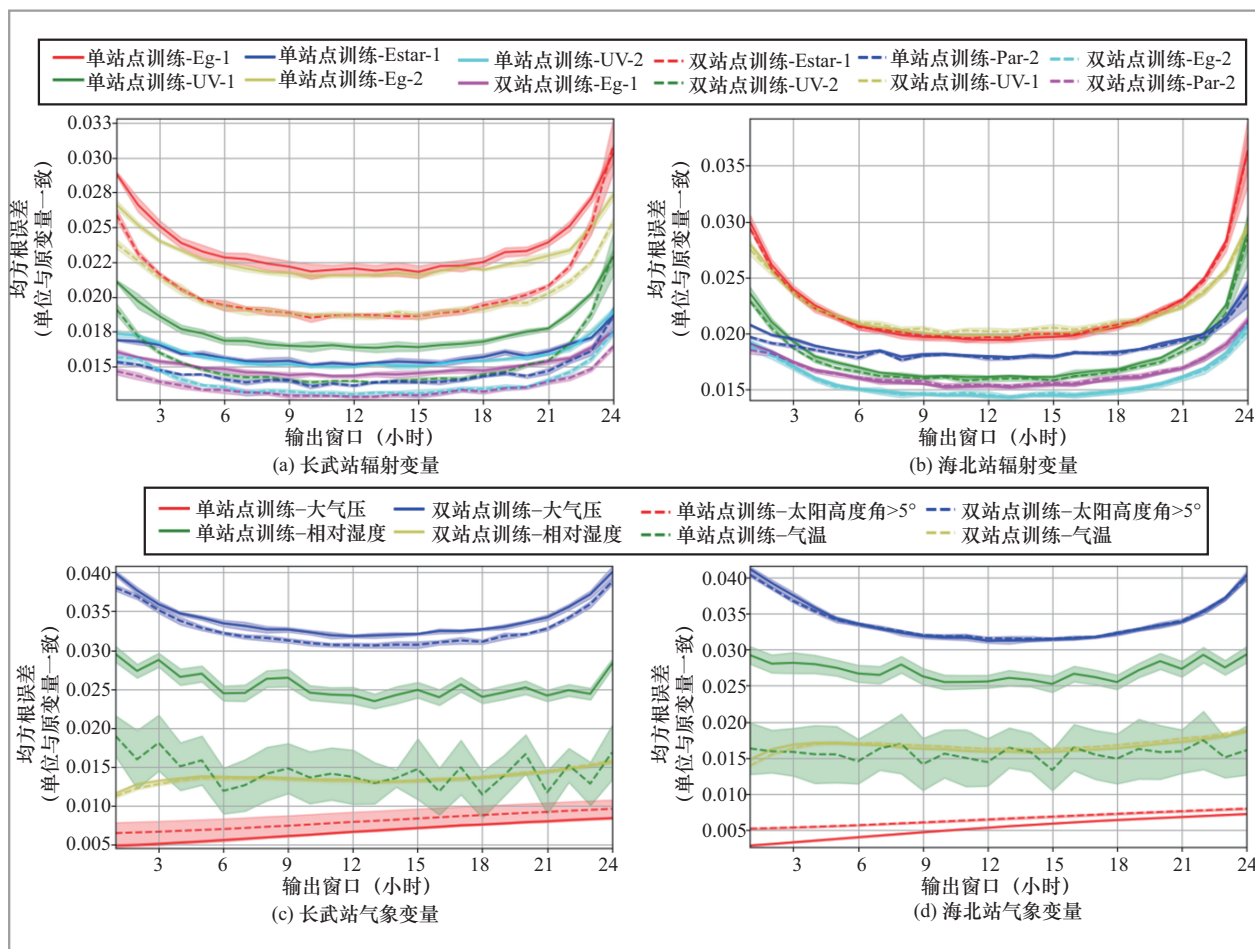


图6 单站点和双站点气象辐射变量预测效果对比

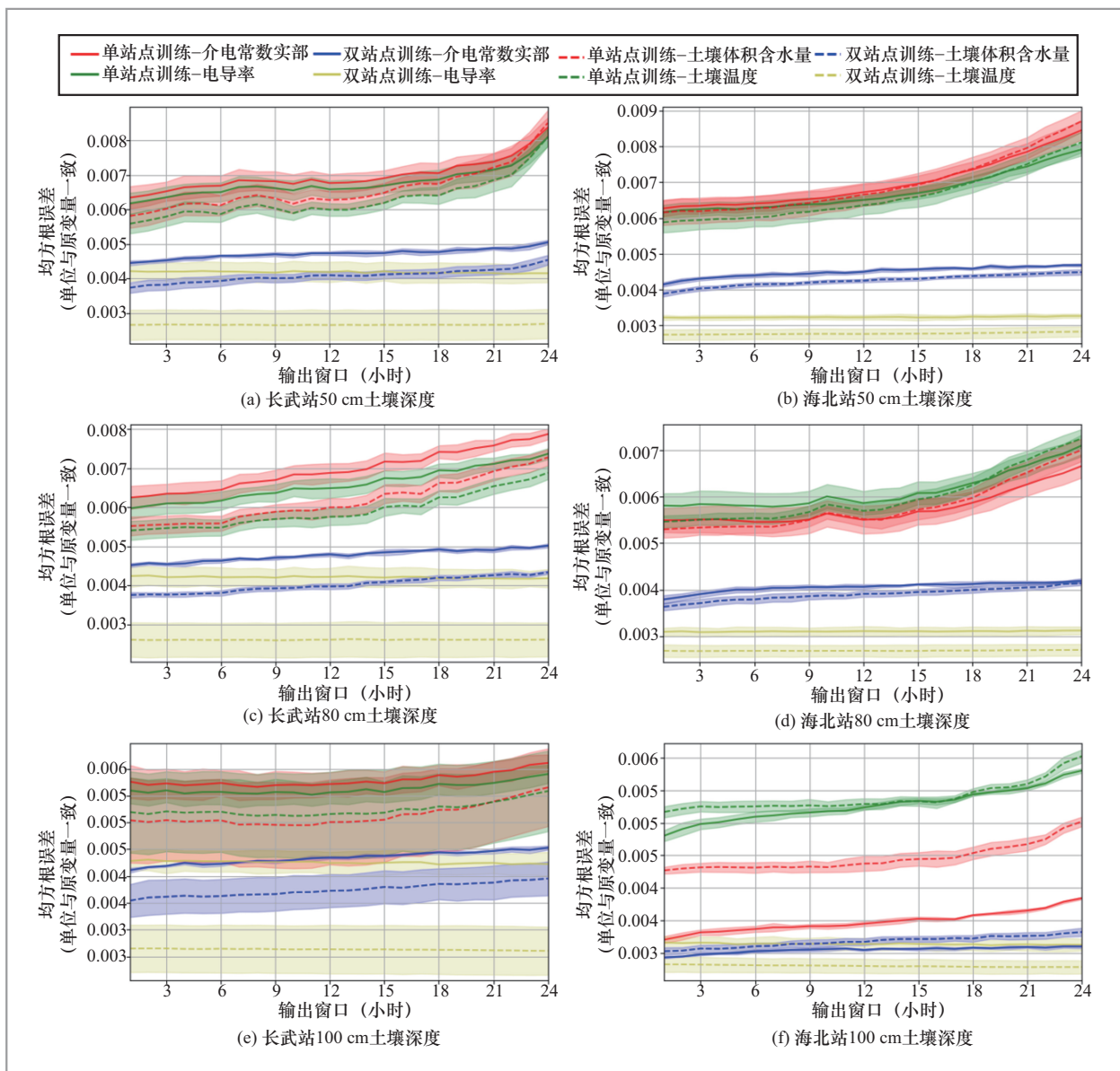


图7 单站点和双站点土壤变量预测效果对比

(如均包含温度、降水、大气压等核心要素),但其概率分布却呈现显著差异。这种“同构异分布”特性与跨域学习场景高度契合。在包含多个域的深度学习任务中,深度学习模型普遍能够动态捕捉不同数据分布下的跨域不变特征,进而构建兼具域适应性与泛化性的拟合模型。本节的实验结果初步表明,Transformer模型能够发掘

多个不同生态变量数据中蕴含的共通机理,形成在多场合下泛用的预测模型。

3.3 多站点可预测变量预测验证

前述研究已经初步证明了Transformer模型在生态领域的有效性和通用性。为了进一步验证模型在未知数据

分布上的泛化能力，本节基于海北站和长武站 2024 年上半年的生态数据，对 3.2 节中训练的 120 h 输入的多站点、全变量输入、可预测变量输出（无降水风速）预测模型（使用 2018—2023 年 70% 的生态数据训练）进行测试。剔除缺失与异常数据后，得到长武站 4 345 条、海北站 3 816 条测试样本。2024 年上半年的生态数据测试结果见表 4。表 4 展示了模型输出序列第 1 个小时中各个变量的 RMSE。由表 4 可知，

表4 2024年上半年的生态数据测试结果

特征变量	长武站 RMSE	海北站 RMSE
总辐射辐照度	0.097 2±0.002 6	0.117 7±0.001 7
紫外辐射辐照度	0.081 6±0.002 7	0.103 9±0.001 2
净辐射辐照度	0.047 3±0.001 4	0.055 9±0.000 5
总辐射总量曝辐量	0.079 1±0.003 0	0.090 4±0.001 6
紫外辐射总量曝辐量	0.056 8±0.001 8	0.069 0±0.001 4
光合有效辐射总量通量密度	0.051 4±0.001 7	0.059 5±0.001 1
太阳高度角是否大于5°	0.037 1±0.003 0	0.044 1±0.003 3
大气压	0.009 9±0.001 2	0.009 5±0.000 8
相对湿度	0.111 7±0.002 0	0.117 1±0.002 0
气温	0.022 5±0.000 6	0.026 6±0.000 7
介电常数实部 10 cm	0.010 3±0.000 5	0.013 3±0.000 5
介电常数实部 20 cm	0.009 3±0.000 7	0.012 6±0.000 9
介电常数实部 30 cm	0.008 3±0.0011	0.010 2±0.001 1
介电常数实部 50 cm	0.009 0±0.001 1	0.010 0±0.001 1
介电常数实部 80 cm	0.007 4±0.000 7	0.008 5±0.000 8
介电常数实部 100 cm	0.006 4±0.0007	0.006 9±0.000 8
土壤体积含水量 10 cm	0.010 5±0.000 5	0.013 6±0.000 8
土壤体积含水量 20 cm	0.009 0±0.000 7	0.010 8±0.000 6
土壤体积含水量 30 cm	0.008 6±0.001 2	0.009 1±0.000 6
土壤体积含水量 50 cm	0.008 7±0.000 8	0.010 0±0.001 0
土壤体积含水量 80 cm	0.007 5±0.0007	0.009 5±0.000 9
土壤体积含水量 100 cm	0.006 6±0.000 8	0.008 0±0.000 9
电导率 10 cm	0.005 0±0.000 2	0.005 5±0.000 1
电导率 20 cm	0.004 1±0.000 1	0.005 6±0.000 4
电导率 30 cm	0.004 0±0.000 3	0.004 8±0.000 3
电导率 50 cm	0.004 4±0.000 4	0.004 6±0.000 4
电导率 80 cm	0.004 2±0.000 2	0.005 2±0.000 4
电导率 100 cm	0.004 1±0.000 4	0.004 0±0.000 4
土壤温度 10 cm	0.007 3±0.001 1	0.008 5±0.000 7
土壤温度 20 cm	0.005 5±0.000 7	0.006 9±0.000 7
土壤温度 30 cm	0.005 4±0.001 1	0.006 5±0.001 1
土壤温度 50 cm	0.004 9±0.001 0	0.005 9±0.001 3
土壤温度 80 cm	0.004 9±0.000 7	0.005 6±0.000 8
土壤温度 100 cm	0.004 7±0.000 5	0.005 4±0.000 7

Transformer 模型在各个生态指标上均表现出较高的预测精度，这表明 Transformer 模型在未知数据分布的实际生态应用场景中也具有适用性。

4 结束语

通过本文的探索性实验与研究可以看出，Transformer 模型在 CERN 长窗口、多变量及多站点生态数据预测中展现出了强大的能力，其可以改变依赖专用规则、专有经验或专用模型等手段进行预测的现状，为通用化的未来数据或临近站点时序数据预测提供了新的途径，进而支撑多站点、多变量场景下的数据质控、插补、实时预警及模拟推演等高阶应用。为实现此目标，需要从生态数据采集和预测流程设计两方面进行优化提升。

一方面，需要综合分析 CERN 多样化观测变量数据采集方式、频度、数据类型和生态学特点，纳入更多高相关性辅助变量，扩大数据覆盖的时空范围，为构建以地面观测、无人机观测或卫星遥感等作为模型输入的多模态监测模型提供有效性强的数据，经过时序数据的清洗、集成、规约、变换等处理工作^[27-28]，形成高质量的 AI ready 数据用于 Transformer 模型训练，支撑数据挖掘和数据预测等任务，从而进一步提高模型的质量、可靠性和鲁棒性。

另一方面，面对生态数据天-空-地一体化采集与复杂化预测需求，需要引入更多新的策略来改进预测流程的各阶段：在构建输入数据的过程中，从通道角度区分生态变量，结合高维特征空间对齐技术和时间编码技术，实现模型在多时间尺度变量融合、异步时序信息整合及缺失通道处

理等场景下的应用需求和目标^[23]；运用嵌入的方式融合空间位置等静态信息，结合域自适应以及元学习等技术增强模型对地理异质性的泛化能力；引入跨特征注意力、损失函数加权以及混合专家知识等策略对模型的网络结构和优化算法进行调整改进^[29-33]。

总之，充分利用人工智能最新技术，构建覆盖 CERN 野外观测站的多样化数据处理智能体，涵盖站点数据质量控制、数据产品加工、数据尺度扩展、多模态数据融合等工具，进而推动从生态站点工具到区域性、全国性工具平台的跨越和提升，有助于提高 CERN 信息化和智能化水平。构建 CERN 野外站生态观测应用智能体存在广泛深入的研究空间，且具有广阔的发展和应用前景。

致谢

感谢国家生态系统观测研究网络/中国生态系统研究网络科技资源服务平台, 特别感谢青海海北高寒草地生态系统生态站和陕西长武农田生态系统生态站等提供数据支撑。

参考文献：

[1] 韩海涛, 李仲龙. 地面实时气象数据质量控制方法研究进展[J]. 干旱气象, 2012, 30(2): 261-265.
Han H T, Li Z L. Review of quality control methods of surface real time meteorological data[J]. Journal of Arid Meteorology, 2012, 30(2): 261-265.

[2] Liu H, Hu B, Zhang L, et al. Ultraviolet radiation over China: Spatial distribution and trends[J]. Renewable and Sustainable Energy Reviews, 2017, 76: 1371-1383.

- [3] 杨剑锋, 乔佩蕊, 李永梅, 等. 机器学习分类问题及算法研究综述[J]. 统计与决策, 2019, 35(6): 36-40.
Yang J F, Qiao P R, Li Y M, et al. A review of machine-learning classification and algorithms[J]. Statistics and Decision, 2019, 35(6): 36-40.
- [4] 齐焕东, 朱程, 李序春, 等. 基于规则集和多层感知机的 Argo 温度数据质量控制方法[J]. 热带海洋学报, 2024, 43(5): 190-202.
Qi H D, Zhu C, Li X C, et al. Rule set and multilayer perceptron based quality control method for Argo temperature data[J]. Journal of Tropical Oceanography, 2024, 43(5): 190-202.
- [5] 李文, 邓升, 段妍, 等. 时间序列预测与深度学习: 文献综述与应用实例[J]. 计算机应用与软件, 2020, 37(10): 64-70, 84.
Li W, Deng S, Duan Y, et al. Time series forecasting and deep learning: literature review and empirical example[J]. Computer Applications and Software, 2020, 37(10): 64-70, 84.
- [6] 沈旭东. 基于深度学习的时间序列算法综述[J]. 信息技术与信息化, 2019(1): 71-76.
Shen X D. A survey of time series algorithms based on deep learning[J]. Information Technology and Informatization, 2019(1): 71-76.
- [7] 王慧健, 刘峥, 李云, 等. 基于神经网络语言模型的时间序列趋势预测方法[J]. 计算机工程, 2019, 45(7): 13-19, 25.
Wang H J, Liu Z, Li Y, et al. Trend prediction method of time series trends based on neural network language model[J]. Computer Engineering, 2019, 45(7): 13-19, 25.
- [8] 陈亮, 王震, 王刚. 深度学习框架下 LSTM 网络在短期电力负荷预测中的应用[J]. 电力信息与通信技术, 2017, 15(5): 8-11.
Chen L, Wang Z, Wang G. Application of LSTM networks in short-term power load forecasting under the deep learning framework[J]. Electric Power Information and Communication Technology, 2017, 15(5): 8-11.
- [9] 李亚蒙, 孙宝楠, 丁军航, 等. 基于 LSTM 的海洋表面短期风速预测研究[J]. 海洋科学, 2022, 46(11): 55-66.
Li Y M, Sun B N, Ding J H, et al. Short-term wind-speed prediction of ocean surface based on LSTM[J]. Marine Sciences, 2022, 46(11): 55-66.
- [10] 席小青, 彭河山, 黄家骥. 基于气象数据的太阳辐照度预测方法[J]. 江西电力职业技术学院学报, 2023, 36(7): 4-7.
Xi X Q, Peng H S, Huang J J. Prediction method of solar irradiance based on meteorological data[J]. Journal of Jiangxi Vocational and Technical College of Electricity, 2023, 36(7): 4-7.
- [11] 赵书强, 尚煜东, 杨燕燕, 等. 基于长短期记忆神经网络的地表太阳辐照度预测[J]. 太阳能学报, 2021, 42(3): 383-388.
Zhao S Q, Shang Y D, Yang Y Y, et al. Short-term solar irradiation forecast based on LSTM neural network[J]. Acta Energetica Solaris Sinica, 2021, 42(3): 383-388.
- [12] 孟祥福, 石皓源. 基于 Transformer 模型的时间序列数据预测方法综述[J]. 计算机科学与探索, 2025, 19(1): 45-64.
Meng X F, Shi H Y. Survey of Transformer-based model for time series forecasting[J]. Journal of Frontiers of Computer Science & Technology, 2025, 19(1): 45-64.
- [13] Bi K F, Xie L X, Zhang H H, et al. Accurate medium-range global weather forecasting with 3D neural networks[J]. Nature, 2023, 619(7970): 533-538.
- [14] 姚凯, 朱晓彤, 云天, 等. 盘古气象大模型在东北地区适用性的初步评估[J]. 沙漠与绿洲气象, 2025, 19(1): 157-164.
Yao K, Zhu X T, Yun T, et al. Applicability assessment of the pangu weather

forecast model in Northeast China[J]. Desert and Oasis Meteorology, 2025, 19 (1): 157-164.

[15] 杨萍, 于秀波, 庄绪亮, 等. 中国科学院中国生态系统研究网络(CERN)的现状与未来发 展思路[J]. 中国科学院院刊, 2008, 23(6): 555-561.

Yang P, Yu X B, Zhuang X L, et al. Present status and train of thought of future development of Chinese Ecosys - tem Research Network(CERN) of CAS[J]. Bulletin of the Chinese Academy of Sci - ences, 2008, 23(6): 555-561.

[16] 廖小罕, 封志明, 高星, 等. 野外科学观测研 究台站(网络)和科学数据中心建设发展[J]. 地 理学报, 2020, 75(12): 2669-2683.

Liao X H, Feng Z M, Gao X, et al. The development of field scientific observa - tion and research stations (networks) and scientific data centers[J]. Acta Geo - graphica Sinica, 2020, 75(12): 2669- 2683.

[17] 何洪林, 于贵瑞. 《中国生态系统研究网络 (CERN) 专题》卷首语[J]. 中国科学数据, 2020, 5(1): 1.

He H L, Yu G R. Preface of China Eco - system Research Network (CERN) [J]. China Scientific Data, 2020, 5(1): 1.

[18] 国家生态系统观测研究网络, 中国生态系统研 究网络. 国家生态系统观测研究网络科技资源 服务系统[DB]. 2015.

National Ecosystem Research Network of China, Chinese Ecosystem Research Network. National Ecosystem Research Network Scientific and Technological Resources Service System [DB]. 2015.

[19] 刘广仁. 生态系统大气环境观测规范[M]. 北 京: 中国环境科学出版社, 2007.

[20] Kaur J, Parmar K S, Singh S. Autore - gressive models in environmental fore - casting time series: a theoretical and application review[J]. Environmental Science and Pollution Research, 2023, 30(8): 19617-19641.

[21] Hansen B E. Threshold autoregression in economics[J]. Statistics and Its Inter - face, 2011, 4(2): 123-127.

[22] Sun Z Q, Li Z H, Wang H Q, et al. Fast structured decoding for sequence models[C]// Proceedings of the 33rd International Conference on Neural Information Pro - cessing Systems. New York: ACM, 2019: 3016-3026.

[23] 易杰, 曹腾飞, 黄明峰, 等. 基于时间编码 LSTM的高校舆情热点趋势预测研究[J]. 大数 据, 2022, 8(5): 124-138.

Yi J, Cao T F, Huang M F, et al. Re - search on trend prediction of time - coded LSTM based public opinion hot spots in universities[J]. Big Data Re - search, 2022, 8(5): 124-138.

[24] Zerveas G, Jayaraman S, Patel D, et al. A transformer-based framework for multivariate time series representation learning[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. New York: ACM, 2021: 2114-2124.

[25] Glorot X, Bordes A, Bengio Y. Domain ad - aptation for large-scale sentiment classi - fication: a deep learning approach[C]// Proceedings of the 28th international conference on machine learning (ICML- 11). Madison: Omnipress, 2011: 513-520.

[26] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Sys - tems 30. San Diego: Neural Information Processing Systems Foundation, 2017: 5998-6008.

[27] 钱力, 刘志博, 胡懋地, 等. AI就绪的科技情 报数据资源建设模式研究[J]. 农业图书情报学 报, 2024, 36(3): 32-45.

Qian L, Liu Z B, Hu M D, et al. Con - struction model of AI-ready for scien - tific and technological intelligence data

resources[J]. Journal of Library and Information Science in Agriculture, 2024, 36(3): 32–45.

[28] 邝晓飞, 万里平, 连嘉茜, 等. 基于Transformer和多序列特征的土壤含水率预测[J]. 农业机械学报, 2025, 56(8): 120–127.

Kuang X F, Wan L P, Lian J Q, et al. Prediction of soil moisture content based on Transformer and multi sequence feature[J]. Transactions of the Chinese Society for Agricultural Machinery, 2025, 56(8): 120–127.

[29] Zhang Y H, Yan J C. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting[C]//Proceedings of the Eleventh International Conference on Learning Representations. San Diego: International Conference on Learning Representations, 2023.

[30] Finzi M, Gruver N, Qiu S K, et al. Large language models are zero-shot time series forecasters[C]//Proceedings of the Advances in Neural Information Processing Systems 36. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 19622–19635.

[31] Jin M, Wang S Y, Ma L T, et al. Time-LLM: time series forecasting by reprogramming large language models[C]//Proceedings of the 12th International Conference on Learning Representations. San Diego: International Conference on Learning Representations, 2024: 1–20.

[32] Rasul K, Ashok A, Williams A R, et al. Lag-Llama: towards foundation models for time series forecasting[C]//R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models, Co-located with the 37th Conference on Neural Information Processing Systems. San Diego: Neural Information Processing Systems Foundation, 2023.

[33] Zhou T, Ma Z Q, Wen Q S, et al. Fed-former: frequency enhanced decomposed Transformer for long-term series forecasting[C]//Proceedings of the 39th International Conference on Machine Learning. Cambridge: PMLR, 2022: 27268–27286.

作者简介



郭学兵（1967–），女，中国科学院地理科学与资源研究所生态系统网络观测与模拟重点实验室高级工程师，主要研究方向为生态信息学。



程翔（1998–），男，博士，中国科学院自动化研究所助理研究员，主要研究方向为人工智能驱动的科学。



唐新斋（1976–），男，中国科学院地理科学与资源研究所生态系统网络观测与模拟重点实验室高级工程师，主要研究方向为遥感信息处理。



张黎（1982–），女，博士，中国科学院地理科学与资源研究所生态系统网络观测与模拟重点实验室副研究员、国家生态科学数据中心副主任，主要研究方向为生态系统与全球变化。



何洪林（1971–），男，博士，中国科学院地理科学与资源研究所生态系统网络观测与模拟重点实验室研究员、国家生态科学数据中心主任，主要研究方向为生态信息学、生态遥感、碳循环模型模拟。

收稿日期: 2026-01-04
通信作者: 程翔,chengxiang2018@ja.ac.cn
基金项目: 国家重点研发计划(No.2022YFF1300100)
Foundation Items: The National Key Research and Development Program of China (No.2022YFF1300100)