

用于碳通量长时间缺失值插补的深度学习模型 ARformer

齐建东^{1,2}, 吴 鹏¹, 查天山³

(1. 北京林业大学信息学院, 北京 100083; 2. 国家林业和草原局林业智能信息处理工程技术研究中心, 北京 100083;
3. 北京林业大学水土保持学院, 北京 100083)

摘 要: 为提高净生态系统交换量 (net ecosystem exchange, NEE) 在长期缺失下的插补精度, 为地球碳循环的后续研究提供关键数据, 该研究提出了一种用于 NEE 长时间缺失值插补的模型 Adapter-Reverseformer (ARformer)。首先, 为了更好地捕捉 NEE 数据的时序特征, 重新设计和优化了层归一化 (layer normalization)、前馈网络 (feed-forward network) 和自注意力 (self-attention) 3 个模块的功能, 模型能够使用更长的回望窗口, 从更长的时间序列中捕获时序特征; 其次, 设计了特征融合模块 Attention-MLP, 以拟合 NEE 与环境因子的即时响应关系, 提高模型在不同土地利用类型 NEE 数据上的插补精度。在全球长期通量观测网络 (FLUXNET) 65 个站点、10 种土地利用类型的数据上进行试验, 结果显示, ARformer 模型的 R^2 在 0.762~0.913 之间, 均方根误差 (root mean square error, RMSE)、平均绝对误差 (mean absolute error, MAE) 和偏差 (bias) 在 0.668~2.724、0.410~1.751 和 -0.024~0.067 $\mu\text{mol}/(\text{m}^2\cdot\text{s})$ 之间。在不同缺失长度和土地利用类型下插补精度均高于边际分布采样法 (marginal distribution sampling, MDS)、随机森林 (random forest, RF)、DLinear、PatchTST 和 iTransformer。该研究结果可为 NEE 数据在长时间缺失场景下的插补提供参考。

关键词: 碳通量; Adapter-Reverseformer; 深度学习; 长缺失插补

doi: 10.11975/j.issn.1002-6819.202407071

中图分类号: S24

文献标志码: A

文章编号: 1002-6819(2024)-24-0127-10

齐建东, 吴鹏, 查天山. 用于碳通量长时间缺失值插补的深度学习模型 ARformer[J]. 农业工程学报, 2024, 40(24): 127-136. doi: 10.11975/j.issn.1002-6819.202407071 <http://www.tcsae.org>

QI Jiandong, WU Peng, ZHA Tianshan. Constructing ARformer: A deep learning model for the long-term gap-filling of carbon flux data[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2024, 40(24): 127-136. (in Chinese with English abstract) doi: 10.11975/j.issn.1002-6819.202407071 <http://www.tcsae.org>

0 引 言

涡度相关法 (eddy covariance, EC), 是一种用于直接并长期观测陆地生态系统与大气之间物质和能量通量的技术, 是众多气象、生态站点观测通量的重要工具。该方法目前在测定净生态系统交换量 (net ecosystem exchange, NEE) 方面最为直接^[1-2], 为地球碳循环的后续研究提供了关键数据^[3]。但受极端恶劣天气以及仪器故障的影响, 原始数据经常会出现缺失, 有时缺失长度可能长达半个月甚至一个月; 并且受观测原理与仪器构造等影响, 为了保证数据的准确性, 在得到最终的通量计算结果之前会对原始数据进行一系列必要的质量控制, 也会剔除一些原始观测数据, 造成数据缺失。因此为了获得完整的 NEE 数据集, 必须选择合适的插补方法来插补数据中的缺失。

传统常用的 NEE 插补方法包括非线性回归法 (non-linear regression, NLR)、平均昼夜变化法 (mean diurnal variation, MDV)、查表法 (look-up table, LUT) 以及其

组合而成的边际分布采样法 (marginal distribution sampling, MDS) 等^[4-6]。这些方法各有特点, 选择不同的插补方法可能会产生不同的结果。MDV 方法使用相邻时间段的观测平均值来填补数据空缺, 这种方法结果较为可靠。然而, 由于没有利用环境因子, 在极端晴天或阴天的情况下插补结果可能出现偏差。相比之下, NLR 和 LUT 方法利用了 NEE 数据与环境因子的响应关系。NLR 利用生态系统呼吸方程和光响应曲线估算 NEE, 能够更准确地拟合 NEE 的变化特征。然而, NLR 方法难以拟合 NEE 数据的低频以及高频特征, 从而影响了其 NEE 数据插补的精度和稳定性。LUT 方法则是根据已有观测数据制成索引表, 数据缺失时则在索引表中查找对应环境因子下的观测值来插补缺失数据, 日间最适合以空气温度作为索引, 夜间则适合使用土壤温度, 其插补结果通常精度高于 MDV 方法, 但会受到索引表制成数据所选的气象条件的影响。MDS 方法将 MDV 和 LUT 方法结合, 可以自动延长插补窗口, 在插补 NEE 数据的短期缺失时十分有效, 已成为国际上许多通量网络插补 NEE 缺失数据的标准方法之一。然而, MDS 更适用于 NEE 数据连续缺失小于 15 d 的情况, 对于 15 d 以上的长时间数据缺失, MDS 的插补性能将显著下降^[7]。

近年来, 随机森林 (random forest, RF) 等机器学习模型用于缺失通量数据插补中的研究逐渐兴起。

收稿日期: 2024-07-09 修订日期: 2024-10-09

基金项目: 国家重点研发计划项目 (2020YFA0608100); 国家自然科学基金项目 (32071842)

作者简介: 齐建东, 教授, 主要研究方向为生态系统过程监测与模拟。
Email: qjld@bjfu.edu.cn

MAHABBATI 等^[8]比较机器学习、MDS 等 9 种不同插补方法的性能;XIAO 等^[9]使用了多重插补和决策树等插补方法开展通量数据插补的研究,发现多重插补方法能更准确地反映通量数据的空间分布;VEKURI 等^[10]使用了极端梯度提升(extreme gradient boosting, XGBoost)对二氧化碳通量数据进行研究,发现在欧洲、北美洲北部高纬度地区插补二氧化碳通量缺失数据时,MDS 的插补性能不如 XGBoost。基于数据驱动的机器学习模型在缺失数据插补上的精度和稳定性超越传统插补方法^[11],因此已广泛应用于 NEE 的插补中。然而在插补 30 d 以上的缺失数据时,机器学习模型的插补精度会大幅下降^[7]。

最近,深度学习模型在时间序列数据缺失值插补和预测上取得很大进展^[12-15]。深度学习模型能够拟合时序数据中周期性、趋势性、季节性等复杂的数据模式,从而处理存在不规则缺失值的时序数据,可以更准确地实现 NEE 数据插补任务^[16]。

当前的 NEE 数据插补研究多集中于温带森林生态系统^[17],对插补方法在其他生态系统的效果验证较少。在高纬度地区的生态系统中,植物生长季节短,对于 NEE 数据非常重要的环境因子太阳辐射量在一年中分布非常不均匀,夜间潜在可用的数据缺失较多,这可能会严重影响 NEE 数据的插补精度。尤其是在 NEE 数据缺失长度超过 30 d 的情况下,现有的插补方法插补精度显著降低^[7, 11],这不仅可能导致数据错误,还会影响模型模拟的质量,进而阻碍了气候变化对生态系统碳收支影响的研究。

通量观测数据属于时间序列数据,本研究将时间序列深度学习模型应用于 NEE 数据的插补任务中。探究针对时间序列数据的深度学习模型 DLinear、PatchTST 和 iTransformer 在不同土地利用类型下、不同缺失程度下的 NEE 数据插补效果。进一步提出了 Adapter-Reverse-former (ARformer) 模型,结合了 Attention-MLP 和 Reverseformer,通过 Attention-MLP 捕捉环境因子与 NEE 之间的响应关系,并通过 Reverseformer 捕捉 NEE 数据在时间上的趋势及自相关性。模型综合考虑 NEE 的时序特征和环境因子,能够更准确地模拟 NEE 变化的内在机制。本研究旨在为各种土地利用类型的 NEE 长时间缺失数据插补提供一种更为精确的插补方法。

1 数据源及研究方法

1.1 站点选择

FLUXNET 2015 数据集包含来自 206 个全球分布站点的公开数据(30 min 尺度),包括经过质量控制的生态系统 NEE、H 和 LE 通量以及相关气象和生物变量^[18]。虽然由不同的研究人员安装和维护,但所有站点都应用了统一的通量后处理程序^[18-19]。

本研究使用了每半小时尺度的 FLUXNET 2015 数据产品:NEE_VUT_REF 和 NEE_VUT_REF_QC。质量控制标志(*_QC)用于标识数据集中已经插补过的 NEE 数据的质量。从各站点的数据中进行筛选,截取各站点

连续一年(17 520 条)的原始数据用于插补试验。206 个站点中,81 个站点有超过连续一年的原始数据,其中 16 个站点缺少进行插补试验所需的环境因子。最终,本研究选用符合条件的 65 个站点数据验证 NEE 缺失数据的插补结果。

1.1.1 环境因子选择

植被的生长环境影响植被光合作用和呼吸作用的能力,进而影响 NEE 的量^[20]。因此进行 NEE 插补试验时要选取合适的环境因子。光合有效辐射作为植物光合作用中所需能量的主要来源,直接影响植物的生长、发育和繁殖^[21]。空气温度对植被光合作用和生态系统呼吸都有重要影响^[22],因为温度变化会影响生物体内的生化反应速率。同时,饱和水汽压差可以表示空气干燥程度,该环境因子会显著影响植物气孔开合,从而调控植物的呼吸、光合作用等生理过程。以上 3 项也是 MDS 方法工作所必需的环境因子。本研究选择空气温度(temperature of air, TA)、土壤温度(temperature of soil, TS)、向下短波辐射(shortwave incoming radiation, SW_IN)、净辐射(net radiation, NETRAD)、相对湿度(relative humidity, RH)、土壤含水率(soil water content, SWC)和饱和水汽压差(vapor pressure deficit, VPD)作为环境因子^[22]。

1.1.2 站点特征信息

对每个站点,提取了地理位置、土地利用类型、当地气象、气候分类和仪器设置的特征信息,以提供关于插补性能分析的全面信息。从 FLUXNET 站点元数据中提取了经纬度、海拔、国际地圈生物圈计划(international geosphere-biosphere program, IGBP)等信息^[18]。

所选站点包括 10 种土地利用类型,分别为农田(croplands, CRO)、郁闭灌木林(closed shrublands, CSH)、落叶阔叶林(deciduous broadleaf forests, DBF)、常绿阔叶林(evergreen broadleaf forests, EBF)、常绿针叶林(evergreen needleleaf forests, ENF)、草地(grasslands, GRA)、混交林(mixed forests, MF)、稀疏灌木林(open shrublands, OSH)、稀树草原(savannas, SAV)和有林草地(woody savannas, WSA)。

1.2 人工缺失场景

以连续 1、7、15、30 和 90 d 数据作为单位缺失片段生成 5 类数据缺失场景,各缺失片段均随机生成,每类缺失的数据条数约占数据集总数据条数的 25%,见表 1。

表 1 5 类数据缺失场景

Table 1 Description of five gap scenarios			
缺失片段长度 Missing fragment length/d	单位缺失片段数据条数 Number of missing pieces per unit	缺失片段重复次数 Number of repeats in unit gap	总缺失数据条数 Total number of missing pieces
1	48	90	4 320
7	336	12	4 032
15	720	6	4 320
30	1 440	3	4 320
90	4 320	1	4 320

为提高数据利用率,消除季节性对插补性能的影响,保证插补结果的准确性和稳定性,各类场景随机重复

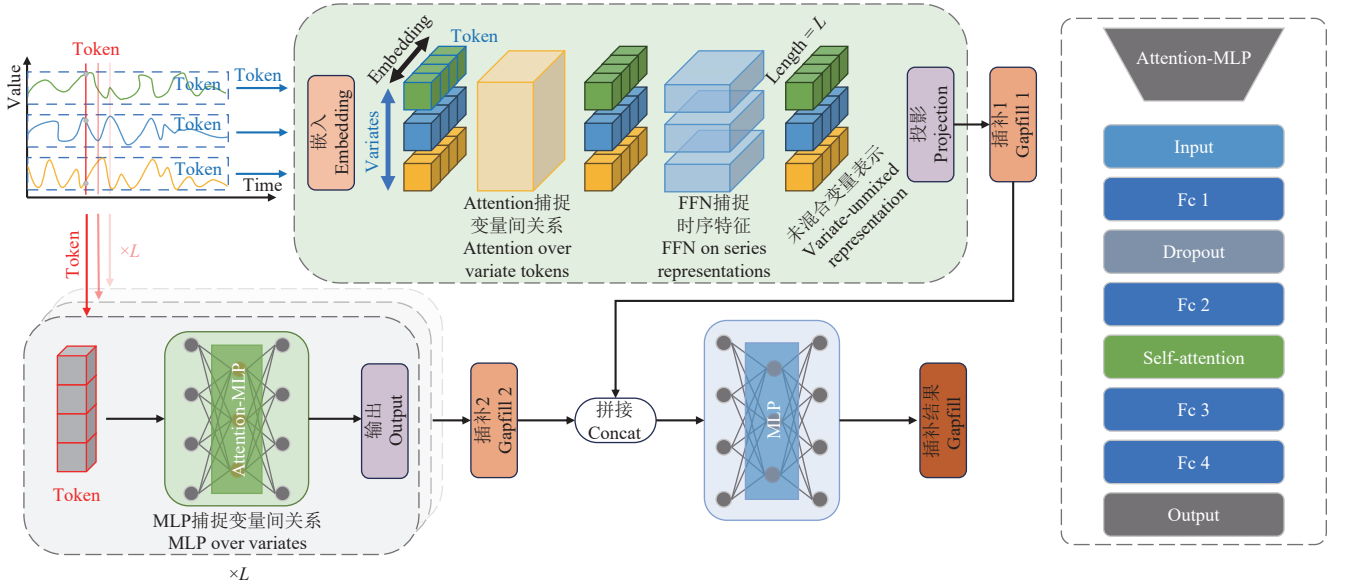
10 次，再跟原始数据集相叠加，生成缺失数据集。

1.3 插补方法

在研究多变量时间序列插补任务中，传统的插补方法无法捕捉 NEE 数据的自相关性，难以处理长时间缺失；而现有的时序深度学习模型虽然在捕捉时间序列的自相关性方面表现出色，但在处理多变量间复杂关系时存在一定局限性，尤其是在不同时间戳上变量之间的映射关系难以充分捕捉。而 NEE 数据与环境因子之间存在显著

的非线性关系，且在每个时间戳上这种关系都需要精确捕捉。因此，传统的方法在处理 NEE 数据时，往往难以取得理想的效果。为了在捕捉时间序列数据的自相关性的同时更好地捕捉 NEE 数据与环境因子之间的复杂非线性关系，同时提高模型的计算效率和泛化能力，本文提出了 Adapter-Reverseformer(ARformer) 模型。

Adapter-Reverseformer 模型分为 Attention-MLP 和 Reverseformer 模型两个主要模块，模型结构见图 1。



注：图中的 L 表示时间序列数据的长度。
Note: The L means the length of the time series data.

图 1 Adapter-Reverseformer 模型结构图
Fig.1 Structure of Adapter-Reverseformer model

1.3.1 Reverseformer 模型

Reverseformer 模型针对 Transformer 处理长回望窗口时间序列效果不佳的问题，重新设计和优化了 Transformer 模型的层归一化 (layer normalization)、前馈网络 (feed-forward network) 和自注意力 (self-attention) 3 个模块的功能，以便更好地处理多变量时间序列数据。

层归一化最初是为了提高深层网络的收敛性和训练稳定性而提出的，有助于减小内部协变量转移，使得模型在训练过程中能够更快地收敛，同时提高模型的泛化能力。Reverseformer 将层归一化应用于单一变量时间序列，能够更好地处理多变量时间序列插补中的非平稳问题，提高了插补的稳定性和准确性。层归一化的这种应用方式使得模型能够更好地适应不同变量之间的差异，并避免了传统方法中将层归一化应用于同一时间戳的多变量导致引入噪声和使时间序列过于平滑等问题。

层归一化的公式如下：

$$\text{LayerNorm}(x) = \frac{x - E(x)}{\sqrt{\text{Var}(x)}} \quad (1)$$

式中 x 为输入向量； $E(x)$ 为输入向量的均值； $\text{Var}(x)$ 为输入向量的方差。

前馈网络作为 Transformer 模型编码的基本构建块，在传统 Transformer 中，被应用于同一时间戳的多变量制成的 token，这些变量可能会错位并且过于局部化，导致前馈网络无法学习到足够的信息来进行插补。Reverseformer 将前馈网络应用于单一变量时间序列制成的 token，根据 HORNICK 的通用近似定理 (universal approximation theorem)，前馈网络可以捕捉到时间序列中复杂的时序特征。通过这些基本构建块的堆叠，使用密集的非线性连接对观测到的时间序列进行编码，并解码这些特征以实现序列的插补。前馈网络这一应用方式更好地利用了线性层的优势，可以使用更长的回望窗口，从更长的时间序列中捕获时序特征，并提高了在其他变量上的泛化能力。其输出式如下：

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (2)$$

式中 W_1 为第一层的权重矩阵； b_1 为第一层的偏置向量； W_2 为第二层的权重矩阵； b_2 为第二层的偏置向量。

自注意力在传统 Transformer 中通常用于对时间序列的时序特征进行建模，通过计算序列中所有位置之间的注意力权重得分，捕捉时间依赖关系。Reverseformer 将自注意力机制应用于每个单一变量时间序列制成的 token，以捕捉不同变量之间的相关性。具体来说，对于

每个变量 token, 计算其与其他所有变量 token 之间的注意力权重得分, 并根据这些得分加权求和, 生成一个新的表征, 因此可以更好地捕捉变量之间的相互关系。

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (3)$$

式中 $\mathbf{Q}$ 为查询矩阵, 表示输入序列中每个位置的查询向量; $\mathbf{K}$ 为键矩阵, 表示输入序列中每个位置的键向量; $\mathbf{V}$ 为值矩阵, 表示输入序列中每个位置的值向量; d_k 为键向量的维度, 用于缩放点积; Softmax 表示归一化。

在 Reverseformer 中, 自注意力机制还结合了多头注意力 (multi-head attention) 机制。多头注意力通过并行计算多个独立的自注意力, 并将它们的结果进行拼接, 进一步增强了模型的表示能力。这种方法使得模型能够捕捉到更加多样和细致的模式, 提高了模型在复杂的多变量时间序列数据上的表现。

多头注意力的公式如下:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h)\mathbf{W}^O \quad (4)$$

$$\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) \quad (5)$$

式中 $\mathbf{W}_i^Q$ 为第 i 个注意力头的查询权重矩阵; $\mathbf{W}_i^K$ 为第 i 个注意力头的键权重矩阵; $\mathbf{W}_i^V$ 为第 i 个注意力头的值权重矩阵; $\mathbf{W}^O$ 为多头注意力输出的权重矩阵。

1.3.2 Attention-MLP 模块

基于 Transformer 架构的模型在模拟时间序列的自相关性方面表现出色, 但其在多变量方面主要捕捉的是不同变量序列之间的关系。然而, 在本研究中, NEE 数据和环境因子之间的关系在每一个时间戳上都有较为明显的映射关系, 因此, 单纯依赖不同变量序列间关系的模拟方法存在一定的局限性。为了解决这一问题, 设计了特征融合模块 Attention-MLP, 以拟合 NEE 与环境因子的即时响应关系。

Attention-MLP 是一种浅层全连接神经网络, 包含 4 个全连接层和一个自注意力层, 所需的训练数据和计算资源较少。通过 4 层使用 ReLU 激活函数的全连接层, 模型能够捕捉 NEE 与环境因子之间的非线性关系, 自注意力层则用于在较低计算密集度下增强对变量间上下文信息的捕捉。此外, 模型中嵌入了 dropout 层以避免过拟合, 通过随机丢弃部分神经元, 防止模型在训练过程中对训练数据过拟合, 提升了模型的泛化能力。结合上述设计, Attention-MLP 能够在较低计算成本下, 有效捕捉和融合 NEE 与环境因子之间复杂的非线性关系, 提升插补性能和模型的稳定性。

Attention-MLP 的自注意力机制相比于 Reverseformer 更简单, 计算量较低, 通过一个线性变换和 tanh 激活函数计算评分, 再通过另一个线性变换计算注意力权重, 并将这些权重直接用于输入向量的加权求和。这种设计在保证计算效率的同时, 仍能捕捉 NEE 与环境因子之间的非线性关系。其公式如下:

$$\text{scores} = \tanh(\mathbf{W}(\mathbf{x})) \quad (6)$$

$$\text{attention_weights} = \text{Softmax}(\mathbf{V}(\text{scores}), \text{dim} = 1) \quad (7)$$

$$\text{context_vector} = \text{attention_weights} \odot \mathbf{x} \quad (8)$$

式中 $\mathbf{W}$ 为一个线性变换矩阵; $\mathbf{V}$ 为另一个线性变换矩阵, 将评分映射到单个标量; scores 表示输入特征的重要性评分; attention_weights 表示输入特征的注意力权重; context_vector 表示输入特征的加权求和结果; $\odot$ 表示逐元素乘法。

最后, 将两者输出的 NEE 插补结果通过 1 个包含 3 个使用 ReLU 激活函数的全连接层和 1 个自注意力层的 MLP, 得到 NEE 最终的插补结果。

1.4 基准插补方法

为验证 ARformer 模型在 NEE 数据集上的插补效果, 选用了 5 种插补模型作为基准插补方法, 包括 NEE 插补领域最通用的 MDS 方法、目前 NEE 插补中研究较多的机器学习方法 RF、以及在时间序列数据预测和插补领域最新、效果最佳的研究成果 DLinear、PatchTST 和 iTransformer 模型^[13-15]。

1.4.1 边际分布采样法

边际分布采样法 (MDS) 作为 FLUXNET 处理缺失值的标准方法, 是 MDV 法和 LUT 法的综合使用。

1.4.2 随机森林模型

随机森林 (RF)^[23] 计算成本低、结果可解释性强, 通过组合多个决策树并采用投票或取平均值的方式, 能够有效防止过拟合。

1.4.3 DLinear 模型

DLinear 模型^[13] 将时间序列数据分解为趋势和季节成分, 并通过单层线性网络处理这 2 个序列, 提升了时序数据的预测性能。

1.4.4 PatchTST 模型

PatchTST 模型^[14] 通过 patch 分段和通道独立两个设计优化多变量时间序列预测。它通过划分子序列并共享权重, 让模型能够更好地捕获数据的局部模式和特征。

1.4.5 iTransformer 模型

iTransformer 模型^[15] 通过分离变量, 将不同变量建模为独立 token, 并分别处理其相关性和时序特征, 提高了时序数据的插补性能。

1.5 评价指标

使用均方根误差 (root mean square error, RMSE)、平均绝对误差 (mean absolute error, MAE)、决定系数 (coefficient of determination, R^2) 和偏差 (bias) 对每个站点不同插补方法的 NEE 插补值和实测值进行统计比较, 以评估插补结果的准确性和性能。

bias 表示插补结果与实测值之间的系统性偏差, 是对插补方法是否存在系统性误差的一种量化, 计算方式见式 (9)。通过计算 bias, 可以评估插补方法是否存在高估或低估现象。

$$\text{bias} = \frac{(\sum \text{fill.} - \sum \text{meas.})}{n} \quad (9)$$

式中 fill. 为人工缺失的通量插补值, $\mu\text{mol}/(\text{m}^2 \cdot \text{s})$; meas.

为人工缺失的通量实测值, $\mu\text{mol}/(\text{m}^2\cdot\text{s})$; n 为 30 min 通量数据的总缺失条数。

1.6 试验环境及参数配置

本研究的试验环境为: E5-2 678 处理器、32GB 内存、RTX2080super 显卡、Ubuntu20.04 系统、Python3.8 环境、PyTorch1.7.1 框架。为保证不同模型对插补效果的公平性, 本文试验中的 4 种时间序列深度学习模型采用相同的超参数, 随机森林模型采用网格搜索得出的最优超参数, 各模型的超参数见表 2。

表 2 6 种模型主要超参数设置
Table 2 Primary hyper-parameter setting of six models

模型 Model	超参数 Hyperparameter	设定值 Setting value
DLinear PatchTST iTransformer Reverseformer	top_k	5
	e_layers, d_layers	2, 1
	d_model, d_ff	128, 128
	mask_rate	0.25
	activation	GELU
	solver	Adam
	batch_size	16
	learning_rate	0.001
	epoch	10
	missing_len	48, 336, 720
Attention-MLP	seq_len	384, 672, 1 440
	activation	ReLU
	batch_size	256
随机森林 Random forest	learning_rate	0.001
	epoch	300
	n_estimators	193
	max_depth	35

在随机森林模型的训练过程中, 以环境因子作为特征输入, 以 NEE 数据为目标变量, 构建随机森林回归模型以完成数据插补。在 4 种时间序列深度学习模型的训练过程中, 模型的输入是形状为 (B, T, N) 的张量, 其中 B 为 batch size, T 为时间步数, N 为特征数。模型在运算时, 每次会处理 B 个长度为 T 的时间序列, 因此对于不同的缺失长度 missing_len, 需设置不同的参数 T , 即序列长度 seq_len。在数据缺失长度为 1、7 和 15 d 时, 缺失长度为 48、336 和 720, 序列长度分别设置为 384、672 和 1 440。在数据缺失长度为 30 和 90 d 时, 如直接进行单步插补, 序列长度过长, 运算成本过高, 因此使用 1 440 的序列长度进行循环插补。

本试验中数据集包含 65 个站点的全年 NEE 数据。为了评估不同缺失时长下的插补效果, 每个站点针对 1、7、15、30 和 90 d 这 5 种缺失时长分别设置了 10 种随机缺失场景。因此, 每种缺失时长对应 650 个缺失样本集, 从而保证了在不同缺失时长下进行公平的对比分析。

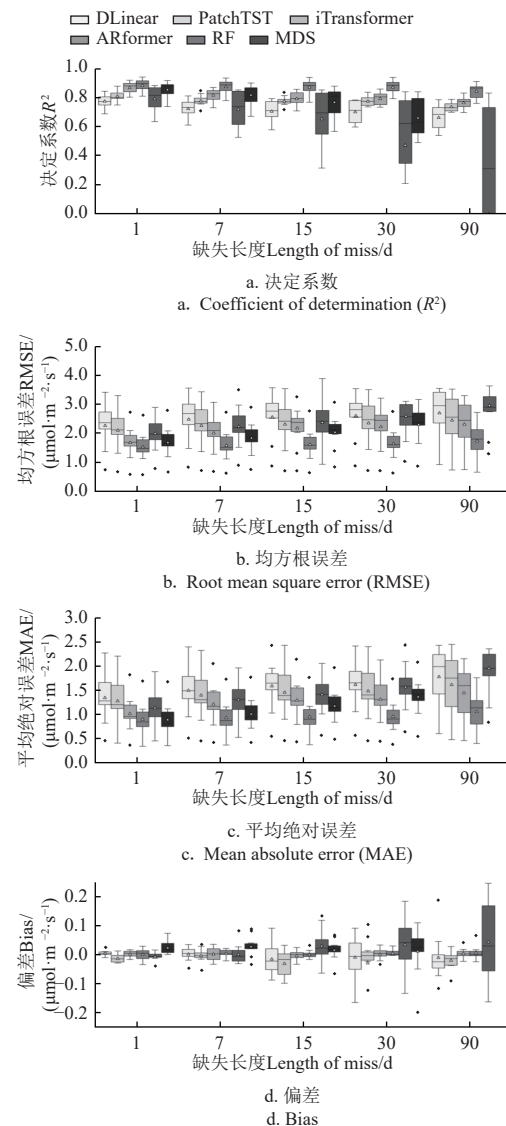
数据集被随机分成 80% 的训练集和 20% 的测试集。不同缺失长度的数据缺失场景模型分别进行训练。最终, 利用训练完成的模型对人工 NEE 缺失数据集进行插补。

2 结果与分析

2.1 各方法在不同缺失长度下的插补效果

当 NEE 数据缺失长度小于 7 d 时, 各模型的插补精度和稳定性最佳; 随着缺失长度增加, 6 种模型的插补精度和稳定性均有所下降 (图 2)。缺失长度不超过

15 d 时, ARformer 表现最佳, 插补精度较高, MDS、RF 和 iTransformer 表现良好, 而 DLinear 稍逊一筹。图 3 展示了缺失 15 d 时 4 种深度学习模型的插补数据与实际数据的对比, 显示 ARformer 最能准确捕捉时间序列的变化趋势。当缺失长度达到 30 d, RF 和 MDS 的精度下降, 而深度学习模型保持了较高的插补精度, 尤其是 ARformer 表现优异, 模型的 R^2 在 0.762~0.913 之间, RMSE、MAE 和 bis 在 0.668~2.724、0.410~1.751 和 -0.024~0.067 $\mu\text{mol}/(\text{m}^2\cdot\text{s})$ 之间。缺失长度为 90 d 时, MDS 无法应对如此长的缺失, RF 精度也显著降低, 而 ARformer 的插补效果依然明显优于其他模型。



注: 箱体上的横线, 从上到下依次为第 3 分位数、中位数以及第 1 分位数, $\triangle$ 表示插补结果的均值, $\bullet$ 表示插补结果中的异常值。下同。
Note: Bars show the third quartile, median and the first quartile as three bars on the boxes in descending order, while the $\triangle$ indicate the mean and the $\bullet$ indicate the outlier. Same behind.

图 2 6 种 NEE 插补模型在不同缺失长度下的插补效果对比
Fig.2 Comparison of interpolation results of 6 NEE interpolation models with different length of miss

综合考虑 4 个指标, 4 种时序深度学习模型在不同缺失情况下插补性能表现稳定, 并且对长时间连续性缺

失的 NEE 数据的插补效果较好, 而 RF 和 MDS 适合较短缺失的插补任务, 但在缺失长度达到 30 d 或以上时表现显著下降。总体来看, iTransformer 和 ARformer 在各类缺失长度情况下表现稳定, 尤其是 ARformer 插补效果最优, 全面优于其他方法。

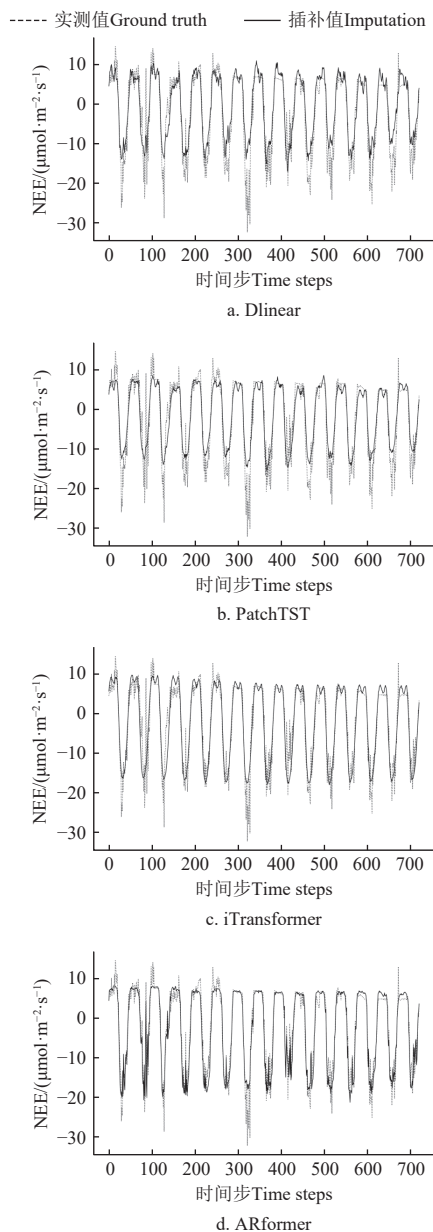


图3 DK-Sor 站缺失 15 d 的净生态系统交换量插补结果可视化

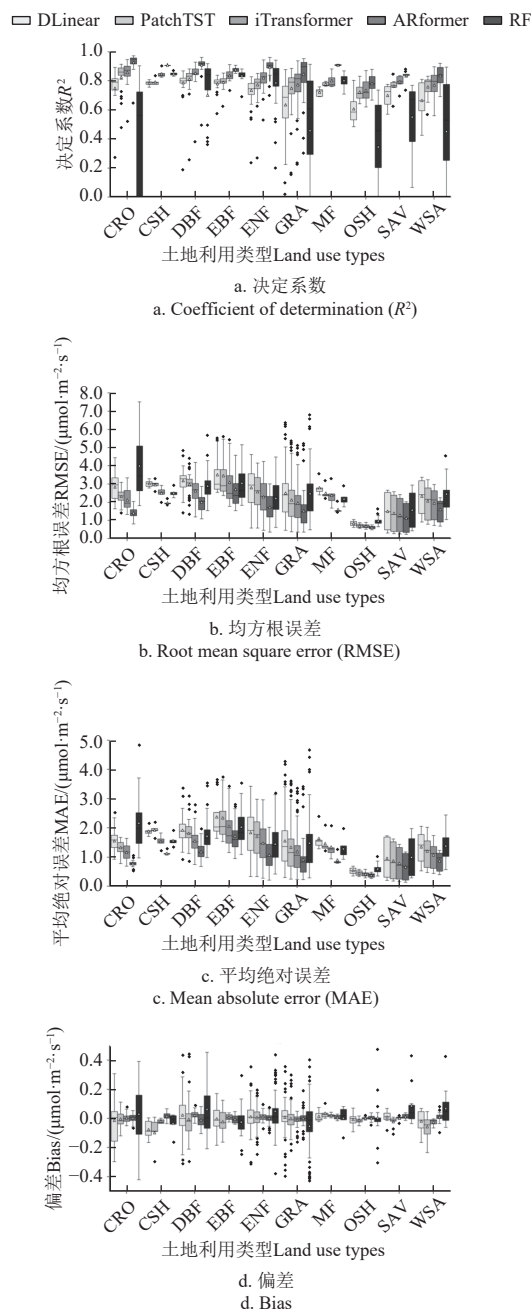
Fig.3 Visualization of gapfill-15 d results of net ecosystem exchange (NEE) on the DK-Sor station

2.2 各方法在不同土地利用类型下的插补效果

在试验所涉及的 10 种土地利用类型中, 时序深度学习模型 DLinear 和 PatchTST 在 CRO、GRA、OSH、SAV 和 WSA 的插补精度优于机器学习模型 RF。RF 在这 5 种土地利用类型的平均 RMSE 为 4.014、2.492、0.988、1.591 和 2.468 $\mu\text{mol}/(\text{m}^2\cdot\text{s})$, 平均 R^2 为 0.461、0.349、0.555 和 0.454。精度和拟合程度低于所有时序深度学习模型。各模型在不同土地利用类型各站点的平均

RMSE 见图 4。

而在 CSH、DBF、EBF、ENF、MF 这 5 种土地利用类型中, RF 的平均 RMSE 为 2.516、2.915、3.068、2.236 和 2.191 $\mu\text{mol}/(\text{m}^2\cdot\text{s})$, 插补精度优于 DLinear 和 PatchTST。



注: CRO、CSH、DBF、EBF、ENF、GRA、MF、OSH、SAV、WSA 分别代表农田、郁闭灌木林、落叶阔叶林、常绿阔叶林、常绿针叶林、草地、混交林、稀疏灌木林、稀树草原、有林草地。

Note: CRO, CSH, DBF, EBF, ENF, GRA, MF, OSH, SAV, WSA represent croplands, closed shrublands, deciduous broadleaf forests, evergreen broadleaf forests, evergreen needleleaf forests, grasslands, mixed forests, open shrublands, savannas, woody savannas.

图4 5种 NEE 插补模型在不同土地利用类型的插补效果对比
Fig.4 Comparison of interpolation results of 5 NEE interpolation models with different land cover types

总体而言, 通过捕捉环境因子与 NEE 关系来进行插补的 RF 模型, 在 CSH、DBF、EBF、ENF 和 MF 这 5 种土地利用类型的插补精度优于时序深度学习模型 DLinear

和 PatchTST。在 CRO、GRA、OSH、SAV 和 WSA 这 5 种土地利用类型的结果则相反。同时，在 GRA、OSH、SAV 和 WSA 这 4 种土地利用类型中，使用多层感知机融合环境因子特征的 ARformer 模型的插补效果相比 PatchTST 提升较小。

在所有的 10 种土地利用类型中，iTransformer 和 ARformer 模型的 RMSE、MAE、bias 都较低， R^2 较高，插补性能优秀，ARformer 模型插补性能最好，全面优于其他插补方法。

2.3 各方法运算复杂度与效率分析

观察表 3 可知，DLinear 由于计算复杂度最低，推理时间最快，适合在计算资源有限的场景中使用，但其精度不如其他模型。PatchTST 虽然在插补性能上表现较好，但其计算复杂度和推理时间较高，适用于计算资源充足

且对精度要求较高的场景。iTransformer 计算复杂度较低，并且有较高的推理速度，但其插补精度略逊于 ARformer。ARformer 在保证插补精度最高的前提下，兼顾了计算复杂度和推理效率，表现出更好的均衡性，是一个兼顾性能与效率的选择。

上述效率的对比集中在推理时间，但是深度学习模型需要在训练数据上进行多次迭代优化，整个训练过程往往耗费大量时间和计算资源。相比之下，RF 和 MDS 方法在时间消耗上具有明显的优势。由于 RF 和 MDS 方法无法计算推理时间和参数量，RF 平均只需约 70 s 即可完成对一年数据的训练和插补，而 MDS 方法无需训练，通过一次运算就能完成插补，整个过程只需约 15 s。在计算资源有限或需要快速处理大规模数据时，这些传统方法仍具有实用性。

表 3 4 种时间序列深度学习模型运算复杂度

序列长度 Sequence length	Dlinear			PatchTST			iTransformer			ARformer		
	运算量	参数量	推理时间	运算量	参数量	推理时间	运算量	参数量	推理时间	运算量	参数量	推理时间
	FLOPs/M	Params/M	Time/ms	FLOPs/M	Params/M	Time/ms	FLOPs/M	Params/M	Time/ms	FLOPs/M	Params/M	Time/ms
384	4.72	0.30	0.6	189.91	2.56	7.6	7.64	0.30	2.5	10.69	0.33	3.9
672	14.46	0.90	0.6	382.76	7.43	10.1	9.56	0.37	2.6	30.91	0.40	6.9
1 440	66.38	4.15	0.9	1 108.29	33.38	38.9	14.69	0.57	2.8	60.42	0.60	7.5

注：表中的推理时间为推理一个形状为 (1, T, 8) 的张量所需的时间。FLOPs 为浮点运算量。
Note: The time in the table is the time required to infer a tensor of shape (1, T, 8). FLOPs is floating point operations.

2.4 ARformer 模型的泛化应用效果

本研究使用 FLUXNET 中各站点数据训练得到 ARformer 模型，为验证 ARformer 模型在其他生态系统监测网络数据插补中的应用效果，本研究针对北京园林绿化生态系统监测网络中的十三陵蟒山森林生态监测站 2023 年全年 NEE 数据进行插补，表 4 展示了 ARformer 模型在长时间缺失情景下对碳通量的插补效果。图 5 展示了缺失 15 d 时 ARformer 的插补数据与实际数据的对比。

表 4 十三陵站 NEE 插补效果

Table 4 Interpolation results of NEE in Shisanling station				
缺失长度 Gap length/d	R^2	RMSE/ ($\mu\text{mol}\cdot\text{m}^{-2}\cdot\text{s}^{-1}$)	MAE/ ($\mu\text{mol}\cdot\text{m}^{-2}\cdot\text{s}^{-1}$)	Bias/ ($\mu\text{mol}\cdot\text{m}^{-2}\cdot\text{s}^{-1}$)
1	0.827	2.657	1.559	0.009
7	0.821	2.659	1.566	-0.054
15	0.818	2.678	1.591	0.048
30	0.817	2.689	1.615	0.046
90	0.802	2.744	1.652	0.039

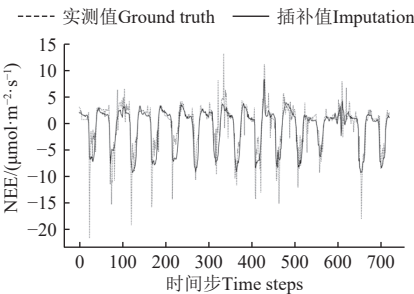


图 5 十三陵站 NEE 缺失 15 d 的插补结果可视化
Fig.5 Visualization of gapfill-15 d NEE results in Shisanling station

插补结果显示，即使在数据连续缺失较长时间的情

况下，ARformer 依旧能捕获时序特征和环境因子的关系，从而实现较高的插补精度。

3 讨 论

3.1 连续缺失程度对 NEE 数据插补效果的影响

对比不同缺失情况下 NEE 插补结果表明，连续缺失天数显著影响插补效果，这与 BENAVIDES 等^[24]的研究一致。在短期缺失情况下，各种插补算法的表现相对接近，均能有效地插补 NEE 数据。然而，随着连续缺失天数的增加，各种插补方法的 RMSE 和 MAE 均呈上升趋势，性能普遍下降。其中，深度学习模型的上升幅度较小，而 RF 和 MDS 的上升幅度较大。这表明各插补方法在超过一定缺失长度后误差会显著上升。

MDS 作为 FLUXNET 处理缺失值的标准方法，在进行较长时间缺失的插补时，性能不如 RF，这与 LUCAS-MOFFAT 等的研究结论一致^[4, 7]。MDS 在缺失不超 15 d 时效果较佳，但超过 15 d 后精度下降。RF 在处理连续、离散和分类变量方面表现出色，适合用气象因子辅助进行 NEE 数据插补^[25]。当 NEE 数据的缺失时间在 30 d 以内时，RF 的插补精度和结果的可靠性都较高，这一结果与 ZHU 等的研究保持一致^[7]。

时序深度学习模型 DLinear、PatchTST 和 iTransformer 随 NEE 数据缺失长度的增加，插补精度的降低幅度较小。深度学习模型可以捕捉时间序列中的长期依赖关系且能够提取多尺度特征，在面对不同长度的缺失数据时，能够保持较高的插补精度。在缺失长度不超过 7 d 时，DLinear、PatchTST 的插补精度低于 RF，因为仅有线性

层的 DLinear 模型整体结构以及 PatchTST 模型用于捕捉环境因子与 NEE 关系的 FFN 层结构较为简单, 导致这 2 个模型并不能很好地对 NEE 与环境因子的相关性进行拟合。当缺失长度超过 15 d 时, DLinear、PatchTST 的插补精度反超 RF 模型。而 iTransformer 模型使用前馈网络捕捉时序自相关性、注意力机制捕捉变量间的关系, 使其在应对各种长度的缺失时都有较高的插补精度。

ARformer 模型的插补精度最高, 而且随着 NEE 数据缺失长度的增加, 插补精度的降低程度最小。当 NEE 数据缺失长度超过 15 d 时, ARformer 模型的插补效果明显优于其他插补方法。即使 NEE 缺失长度达到 90 d, ARformer 模型的插补结果依旧可靠, 在一定程度上提高了 NEE 数据长期缺失的插补精度。因其使用前馈神经网络捕捉时序自相关性、注意力机制捕捉变量间的关系, 在此基础上设计了基于多层感知机的多特征融合模块 Attention-MLP, 进一步拟合 NEE 与环境因子的关系, 与 Reverseformer 的结果进行特征融合。该模型结合了 NEE 数据的环境因子与时序特征, 能够在长期数据缺失的情况下拟合 NEE 的数据走向, 不仅利用了 NEE 数据自身的时序特征, 还精确捕捉了环境因子与 NEE 之间的响应关系, 从而更精确地拟合 NEE 数据的内在趋势, 提高数据插补的精度。

对于连续缺失时间不超过 15 d 时, MDS 和 RF 插补精度与深度学习模型相差不大, 但其时间开销远低于深度学习模型, 在计算资源有限的场景中建议选用。当连续缺失长度较长 (≥ 30 d) 时, 建议使用以 ARformer 模型为代表的深度学习模型进行插补。

3.2 土地利用类型对 NEE 数据插补效果的影响

通过对比不同土地利用类型下 NEE 的插补结果, 研究发现各插补方法在不同土地利用类型下的表现各有优劣, 这一结果很可能与植被变化有关。不同土地利用类型的 NEE 观测数据之间存在差异, 这些差异会导致数据分布特征不同, 从而显著影响插补方法的性能, 这与 RIVERA-MUÑOZ 等^[26]的研究结果相似。

对比不同土地利用类型的环境因子与 NEE 的相关系数发现, 在 CSH、DBF、EBF、ENF 和 MF 这 5 种土地利用类型中, 对 NEE 影响最重要的向下短波辐射、空气温度和饱和水汽压差 3 个环境因子与 NEE 的相关性均高于其他 5 种土地利用类型。这导致了捕捉环境因子与 NEE 关系来进行插补的 RF 模型表现不佳以及 ARformer 模型的插补效果相比其他模型提升较小。SAV 和 GRA 的通量信号较弱, 难以收集通量数据, 导致数据大量缺失, 而且弱信号会产生噪声和干扰, 使得收集到的数据不准确, 造成时间序列数据不均衡, 不利于模型插补, 这与 ZHU 等^[7, 27]的研究结论一致。OSH 和 WSA 插补性能相对较弱是因为其地上覆盖植被相对较少, 地上植被对 NEE 的贡献比重较小^[28], NEE 观测数据容易受到更多环境干扰和噪声的影响, 包括大气变化和周围环境的干扰, 导致数据质量下降, 增加了模型拟合的难度。CRO 中的 NEE 观测数据受到人类活动 (如灌溉和施肥) 的高

度影响^[29], 可能导致土地利用类型变化、化学物质影响环境因子、数据噪声和干扰以及数据不均衡等问题, 这些不利因素会对 NEE 观测数据产生负面影响, 使得插补模型的性能下降。

ARformer 模型通过结合 Attention-MLP 特征融合模块, 显著提升了插补精度。与 DLinear 模型仅依赖线性层进行插补不同, ARformer 能够更好地捕捉 NEE 与环境因子之间的复杂非线性关系, 并在不同时间尺度上整合这些信息, 使其在面对多种土地利用类型时, 表现出较高的泛化能力。相比之下, PatchTST 模型虽然采用了通道独立设计, 但其 FFN 层在捕捉环境因子与 NEE 间复杂关系的能力有所欠缺, 尤其在环境因子较强的土地利用类型 (如 DBF 和 ENF) 中, PatchTST 的插补精度较 ARformer 有所下降。同时, ARformer 的 Reverseformer 模块通过建模时间序列的自相关性, 进一步增强了其在处理长时间缺失数据上的表现, 即使在环境因子较弱的土地利用类型下, ARformer 的插补效果始终优于其他模型。

4 结 论

本研究设计了基于多层感知机的多特征融合模块 Attention-MLP, 有效利用环境因子与 NEE 之间的响应关系, 并通过 Reverseformer 模型拟合了 NEE 在时间上的自相关性, 最后使用 MLP 将 NEE 的时序特征和与环境因子之间的响应关系相结合, 提出了 ARformer 模型。通过随机剔除连续 1、7、15、30 和 90 d 的 5 种缺失场景, 评估了 ARformer 与 5 种基准插补算法在不同数据缺失长度和各种土地利用类型下的插补效果。主要结论如下:

1) 时间序列深度学习模型均能对不同土地利用类型的长时间连续性 NEE 缺失数据进行有效插补。其中, ARformer 模型在不同土地利用类型的 NEE 缺失数据插补中表现出色, 其 R^2 在 0.762~0.913 之间, 均方根误差、平均绝对误差和偏差分别在 0.668~2.724、0.410~1.751 和 -0.024~0.067 $\mu\text{mol}/(\text{m}^2\cdot\text{s})$ 之间。

2) 尤其在数据缺失超过 30 d 的情况下, ARformer 模型的插补精度明显优于其他插补方法, 表现出最佳的综合插补性能, 可应用于 NEE 长期缺失数据插补, 为地球碳循环的后续研究提供关键数据。

[参 考 文 献]

- [1] BALDOCCHI D. How eddy covariance flux measurements have contributed to our understanding of Global Change Biology[J]. *Global Change Biology*, 2020, 26(1): 242-260.
- [2] 时元智, 时红, 崔远来, 等. 基于涡度相关法的长江中下游稻田水分利用效率变化规律[J]. *农业工程学报*, 2021, 37(4): 130-139.
- SHI Yuanzhi, SHI Hong, CUI Yuanlai, et al. Variation of water use efficiency in a paddy field in the middle and lower reaches of the Yangtze River using eddy covariance method[J]. *Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE)*, 2021, 37(4): 130-139.

- (in Chinese with English abstract)
- [3] CHU H, LUO X, OUYANG Z, et al. Representativeness of Eddy-Covariance flux footprints for areas surrounding AmeriFlux sites[J]. *Agricultural and Forest Meteorology*, 2021, 301: 108350.
 - [4] LUCAS-MOFFAT A, SCHRADER F, HERBST M, et al. Multiple gap-filling for eddy covariance datasets[J]. *Agricultural and Forest Meteorology*, 2022, 325: 109114.
 - [5] TENG D, HE X, WANG J, et al. Uncertainty in gap filling and estimating the annual sum of carbon dioxide exchange for the desert Tugai forest, Ebinur Lake Basin, Northwest China[J]. *PeerJ*, 2020, 8: e8530.
 - [6] ZHU S, MCCALMONT J, CARDENAS M, et al. Gap-filling carbon dioxide, water, energy, and methane fluxes in challenging ecosystems: Comparing between methods, drivers, and gap-lengths[J]. *Agricultural and Forest Meteorology*, 2023, 332: 109365.
 - [7] ZHU S, CLEMENT R, MCCALMONT J, et al. Stable gap-filling for longer eddy covariance data gaps: A globally validated machine-learning approach for carbon dioxide, water, and energy fluxes[J]. *Agricultural and Forest Meteorology*, 2022, 314: 108777.
 - [8] MAHABBATI A, BERINGER J, LEOPOLD M, et al. A comparison of gap-filling algorithms for eddy covariance fluxes and their drivers[J]. *Geoscientific Instrumentation, Methods and Data Systems*, 2021, 10(1): 123-140.
 - [9] XIAO Q, GENG G, CHENG J, et al. Evaluation of gap-filling approaches in satellite-based daily PM_{2.5} prediction models[J]. *Atmospheric Environment*, 2021, 244: 117921.
 - [10] VEKURI H, TUOVINEN J, KULMALA L, et al. A widely-used eddy covariance gap-filling method creates systematic bias in carbon balance estimates[J]. *Scientific Reports*, 2023, 13(1): 1720.
 - [11] KANG M, ICHII K, KIM J, et al. New gap-filling strategies for long-period flux data gaps using a data-driven approach[J]. *Atmosphere*, 2019, 10(10): 568.
 - [12] WU H, HU T, LIU Y, et al. TimesNet: Temporal 2D-variation modeling for general time series analysis[EB/OL]. 2022[2023-04-12]. <https://arxiv.org/abs/2210.02186v2>.
 - [13] ZENG A, CHEN M, ZHANG L, et al. Are transformers effective for time series forecasting?[EB/OL]. 2022[2022-08-17]. <https://arxiv.org/abs/2205.13504>.
 - [14] NIE Y, NGUYEN N, SINTHONG P, et al. A time series is worth 64 words: Long-term forecasting with transformers [EB/OL]. 2022[2023-03-05]. <https://arxiv.org/abs/2211.14730>.
 - [15] LIU Y, HU T, ZHANG H, et al. Itransformer: Inverted transformers are effective for time series forecasting[EB/OL]. 2023[2024-03-14]. <https://arxiv.org/abs/2310.06625>.
 - [16] CHEN Z, XU H, JIANG P, et al. A transfer learning-based LSTM strategy for imputing large-scale consecutive missing data and its application in a water quality prediction system[J]. *Journal of Hydrology*, 2021, 602: 126573.
 - [17] SOLOWAY A, AMIRO B, DUNN A, et al. Carbon neutral or a sink? Uncertainty caused by gap-filling long-term flux measurements for an old-growth boreal black spruce forest[J]. *Agricultural and Forest Meteorology*, 2017, 233: 110-121.
 - [18] PASTORELLO G, TROTTA C, CANFORA E, et al. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data[J]. *Scientific Data*, 2020, 7(1): 225.
 - [19] PASTORELLO G, PAPALE D, CHU H, et al. A new data set to keep a sharper eye on land-air exchanges[J]. *Eos, Transactions American Geophysical Union*, 2017, 98(8): 28.
 - [20] 韩信, 张宝忠, 车政, 等. 荒漠草原生长季不同时间尺度水热通量变化特征及其影响因子[J]. *农业工程学报*, 2024, 40(15): 65-75.
 - HAN Xin, ZHANG Baozhong, CHE Zheng, et al. Characteristics and impact factors of water and heat flux changes at different time scales during the growing season of desert grasslands[J]. *Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE)*, 2024, 40(15): 65-75. (in Chinese with English abstract)
 - [21] 彭丽, 赵仲辉, 项文化, 等. 辐射变化对中亚热带杉木人工林净 CO₂ 交换的影响[J]. *应用生态学报*, 2022, 33(1): 17-24.
 - PENG Li, ZHAO Zhonghui, XIANG Wenhua, et al. Effects of radiation changes on net ecosystem exchange of carbon dioxide in a middle subtropical Chinese fir plantation[J]. *Chinese Journal of Applied Ecology*, 2022, 33(1): 17-24. (in Chinese with English abstract)
 - [22] XU L, BALDOCCHI D. Seasonal variation in carbon dioxide exchange over a Mediterranean annual grassland in California[J]. *Agricultural and Forest Meteorology*, 2004, 123(1/2): 79-96.
 - [23] CHANG X, XING Y, GONG W, et al. Evaluating gross primary productivity over 9 ChinaFlux sites based on random forest regression models, remote sensing, and eddy covariance data[J]. *Science of the Total Environment*, 2023, 875: 162601.
 - [24] BENAVIDES F, SANTACRUZ M, ROMERO-LEITON J, et al. Assessing methods for multiple imputation of systematic missing data in marine fisheries time series with a new validation algorithm[J]. *Aquaculture and Fisheries*, 2023, 8(5): 587-599.
 - [25] 王少影, 张宇, 孟宪红, 等. 机器学习算法对涡动相关缺失通量数据的插补研究[J]. *高原气象*, 2020, 39(6): 1348-1360.
 - WANG Shaoying, ZHANG Yu, MENG Xianhong, et al. Fill the gaps of eddy covariance fluxes using machine learning algorithms[J]. *Plateau Meteorology*, 2020, 39(6): 1348-1360. (in Chinese with English abstract)
 - [26] RIVERA-MUÑOZ L, GIRALDO-FORERO A, MARTINEZ-VARGAS J. Deep matrix factorization models for estimation of missing data in a low-cost sensor network to measure air

- quality[J]. *Ecological Informatics*, 2022, 71: 101775.
- [27] 于洪亮, 赵越, 李淑贞, 等. 呼伦贝尔草甸草原不同显热通量观测方法比较[J]. *农业工程学报*, 2023, 39(21): 88-97.
- YU Hongliang, ZHAO Yue, LI Shuzhen, et al. Comparison of observation methods of sensible heat flux in Hulunbuir meadow steppe[J]. *Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE)*, 2023, 39(21): 88-97. (in Chinese with English abstract)
- [28] 金艳强, 李敬, 张一平, 等. 元江干热河谷稀树灌草丛植被碳储量及净初级生产力[J]. *生态学报*, 2017, 37(17): 5584-5590.
- JIN Yanqiang, LI Jing, ZHANG Yiping, et al. Carbon storage and net primary productivity of a savanna ecosystem in a dry-hot valley in Yuanjiang, Yunnan[J]. *Acta Ecologica Sinica*, 2017, 37(17): 5584-5590. (in Chinese with English abstract)
- [29] ZHANG W, LUO G, YUAN X, et al. New data - driven method for estimation of net ecosystem carbon exchange at meteorological stations effectively increases the global carbon flux data[J]. *Methods in Ecology and Evolution*, 2023, 14(9): 2449-2463.

Constructing ARformer: A deep learning model for the long-term gap-filling of carbon flux data

QI Jiandong^{1,2}, WU Peng¹, ZHA Tianshan³

(1. School of Information Science and Technology, Beijing Forestry University, Beijing 100083, China; 2. Engineering Research Center for Forestry-oriented Intelligent Information Processing of National Forestry and Grassland Administration, Beijing 100083, China; 3. School of Soil and Water Conservation, Beijing Forestry University, Beijing 100083, China)

Abstract: Net ecosystem exchange (NEE) measurement is very critical to understanding carbon flux in ecosystems. But some gaps are still common in data collection, due to the harsh weather or sensor malfunction. Traditional interpolation can struggle with the long-term missing data, leading to inaccuracies in carbon flux analysis. Long-term missing data can also occur in NEE measurements. In this study, the Adapter-Reverseformer (ARformer) model was proposed to enhance the accuracy of NEE gap-filling, especially for extended periods of data loss. A multi-layer perceptron (MLP) was integrated with the Reverseformer module. The longer data gaps were effectively utilized to leverage both environmental data and the temporal patterns in the NEE. A novel system was established to focus on the non-linear relationships between NEE and environmental factors. The time-dependent trend was characterized by NEE data. The obtained model was tested using the FLUXNET 2015 dataset. Half-hourly carbon flux data was collected from 65 sites across 10 types of land use. Five artificial gap scenarios were generated by randomly removing data for continuous periods of 1, 7, 15, 30, and 90 days. The performance of ARformer was compared with marginal distribution sampling (MDS), random forest (RF), and three advanced deep learning models: DLinear, PatchTST, and iTransformer. The results demonstrated that the ARformer model consistently outperformed the baseline, especially when dealing with long-term missing data. Specifically, the performance of RF decreased significantly, when the missing data spanned 90 days, and MDS failed to reasonably estimate the model. In contrast, the ARformer model maintained a high accuracy, with R^2 values ranging from 0.762 to 0.913. The root mean square error (RMSE) ranged between 0.668 and 2.724 $\mu\text{mol}/(\text{m}^2\cdot\text{s})$, the mean absolute error (MAE) ranged from 0.410 to 1.751 $\mu\text{mol}/(\text{m}^2\cdot\text{s})$, and bias values remained between -0.024 and 0.067 $\mu\text{mol}/(\text{m}^2\cdot\text{s})$. The ARformer model demonstrated superior performance across different land-use types, including closed shrublands, deciduous broadleaf forests, evergreen broadleaf forests, evergreen needle-leaf forests, and mixed forests. As such, the ARformer model was used to more effectively capture these vegetation types with the complex relationships between NEE and environmental drivers. Furthermore, it was observed that the time-series deep learning models in general provided the better interpolation for the long-term missing NEE data, with ARformer leading in accuracy. In conclusion, deep learning models, particularly the ARformer model, were highly effective in filling the gaps in NEE data for the various ecosystems. The ARformer model was recommended when the data gaps were extended beyond 30 days. The accuracy of interpolation was also attributed to the temporal dependencies and the relationship between NEE and environmental factors. More reliable NEE data was then obtained to clarify the carbon flux dynamics across different ecosystems. Thus, the ARformer model was represented for the long-term data gaps in NEE measurements.

Keywords: carbon flux; Adapter-Reverseformer; deep learning; long-term gap-filling