Earthformer:ExploringSpace-TimeTransformersforEarthSystemForecasting

摘要

传统上，地球系统 (例如天气和气候) 预测依赖于具有复杂物理模型的数值模拟，因此计算成本高昂，并且对领域专业知识要求高。随着过去十年时空地球观测数据的爆炸式增长，应用深度学习的数据驱动模型在各种地球系统预测任务中显示出令人印象深刻的潜力。Transformer 作为一种新兴的深度学习架构，尽管在其他领域取得了广泛的成功，但在这一领域的采用却有限。在本文中，我们提出了 Earthformer，一种用于地球系统预测的时空 Transformer。Earthformer 基于一个通用、灵活且高效的时空注意块，称为长方体注意力机制。这个想法是将数据分解为长方体并并行应用长方体级自注意力。这些长方体进一步与一组全局向量相连。我们对新提出的混沌 N-body MNIST 数据集进行了实验，以验证长方体注意力的有效性并找出 Earthformer 的最佳设计。在厄尔尼诺/南方涛动 (ENSO) 预报的真实基准上进行的实验表明，Earthformer 实现了最先进的性能。

关键词： 地球；预测；深度学习；Transformer；Earthformer；长方体注意力机制

1 引言

地球是一个复杂的系统。地球系统的变化影响着我们的日常生活，从温度波动等常规事件到干旱、冰雹和厄尔尼诺/南方涛动 (ENSO) 等极端事件。在所有后果中，地球系统的变化会影响农作物产量、航班延误、引发洪水和森林火灾。准确及时地预测这些变化可以帮助人们采取必要的预防措施来避免危机，或更好地利用风能和太阳能等自然资源。因此，改进地球变化 (例如天气和气候) 的预测模型具有巨大的社会经济影响。尽管它很重要，但近 50 年来，实际的天气和气候预报系统并没有发生根本性的变化 [30]。这些业务模型，包括美国国家海洋和大气管理局 (NOAA) [29] 使用的最先进的高分辨率集合预报 (HREF) 降雨临近预报模型，都依赖于对物理模型进行细致的数值模拟。这种基于模拟的系统不可避免地无法整合来自新兴地球物理观测系统 [12] 的信号，也无法利用 PB 级的地球观测数据 [36]。

作为一种有吸引力的替代方案，深度学习 (DL) 为地球系统预测提供了一种新方法 [30]。基于深度学习的预测模型不是明确地结合物理规则，而是根据地球观测数据进行训练 [31]。通过从大量观察中学习，深度学习模型能够找出系统的内在物理规则并生成优于基于模拟的模型的预测 [9]。这种技术已在多个应用中取得了成功，包括降水临近预报 [6, 29] 和 ENSO 预报 [15]。由于地球系统是混乱的 [20]、高维和时空的，因此设计适当的深度学习架构来对系统进行建模特别具有挑战性。以前的工作依赖于循环神经网络 (RNN) 和卷积神经网络 (CNN)

的结合 [13,31,32,36,38]。这两种架构施加了时间和空间归纳偏差，有助于捕获时空模式。然而，作为一个混沌系统，地球系统的变异性，例如降雨和 ENSO，对系统的初始条件高度敏感，并且可以对内部变化做出突然响应。目前尚不清楚 RNN 和 CNN 模型中的归纳偏差是否仍然适用于此类复杂系统。

另一方面，近年来 Transformer 的广泛应用给深度学习带来了重大突破。该模型最初被提出用于自然语言处理 [7,35]，后来扩展到计算机视觉 [8,21]、多模态文本图像生成 [28]、图学习 [41] 等。Transformer 依赖于注意力机制来捕获数据相关性，并且在建模复杂和远程依赖关系方面功能强大，这两种依赖关系都出现在地球系统中。尽管 Transformer 适合解决这个问题，但它在地球系统预测中的应用有限。单纯地应用 Transformer 架构是不可行的，因为 $O(N^2)$ 注意力机制对于高维地球观测数据来说计算成本太高。如何设计一个适合预测地球系统未来的时空 Transformer 在很大程度上是一个悬而未决的问题。

在本文中，作者提出了 Earthformer，一种用于地球系统预测的时空 Transformer。为了更好地探索时空注意力的设计，作者提出了 CuboidAttention，它是高效时空注意力的通用构建块。这个想法是将输入张量分解为不重叠的长方体，并并行应用长方体级自注意力。由于我们将 $O(N^2)$ 自注意力限制在局部长方体内，因此整体复杂度大大降低。不同类型的相关性可以通过不同的长方体分解来捕获。通过堆叠具有不同超参数的多个长方体注意力层，我们能够将之前提出的几个 Vision Transformer [4,18,22] 纳入特殊情况，并且还提出了以前未研究过的新注意力模式。这种设计的局限性是缺乏局部长方体相互通信的机制。因此，我们引入了一组关注所有局部长方体的全局向量，从而收集系统的整体状态。通过关注全局向量，局部长方体可以掌握系统的总体动态并相互共享信息。

为了验证长方体注意力的有效性并找出地球系统预测场景下的最佳设计，我们对新提出的 N-body MNIST 数据集进行了广泛的实验：N-body MNIST 中的数字遵循混沌三体运动模式 [33]，这使得数据集不仅比 Moving MNIST 更具挑战性，而且与地球系统预测更相关。综合实验揭示了以下发现：1) 将长方体注意力层与轴向注意力模式堆叠起来既高效又有效，实现了最佳的整体性能，2) 添加全局向量可以在不增加计算成本的情况下提供一致的性能增益，3) 添加层次结构在编码器-解码器架构中可以提高性能。基于这些发现，我们找出了 Earthformer 的最佳设计，并与 ENSO 预报的 ICAR-ENSO 数据集 [15] 的其他基线进行了比较。实验表明，Earthformer 在该项任务上实现了最先进 (SOTA) 的性能。

2 相关工作

2.1 用于地球系统预测的深度学习架构

用于地球系统预测的传统深度学习模型基于 CNN 和 RNN。具有 2D CNN 或 3D CNN 的 U-Net 已用于降水临近预报 [36]、季节性北极海冰预测 [1]ENSO 预报 [15]。施等人 [31] 提出了结合 CNN 和 LSTM 的 ConvLSTM 网络用于降水临近预报。王等人 [38] 提出了 PredRNN，在 ConvLSTM 的基础上添加了时空记忆流结构。为了更好地了解长期的高层关系，王等人 [37] 提出了将 3DCNN 集成到 LSTM 的 E3D-LSTM。为了将 PDE 动力学与未知的补充信息分开，PhyDNet [13] 结合了一个新的循环物理单元来在潜在空间中执行 PDE 约束预测。埃斯佩霍尔特等人 [9] 提出了 MetNet-2，其在降水预测方面优于 HREF。该架构基于 ConvLSTM 和扩张 CNN。最近，有一些工作尝试应用 Transformer 来解决地球系统预测问题。帕塔克等人 [25]

提出了用于全球天气预报的 FourCastNet，它基于自适应傅里叶神经算子 (AFNO) [14]。白等人 [3] 提出了用于降水临近预报的 Rainformer，它基于结合了 CNN 和 SwinTransformer 的架构 [21]。在我们的实验中，我们可以看到 Earthformer 优于 Rainformer。

2.2 用于视频建模的时空 Transformer

受到 ViT [8] 在图像分类方面的成功的启发，采用时空 Transformer 来提高视频理解。为了绕过联合时空注意力带来的巨大内存消耗，一些开创性的工作提出了有效的替代方案，例如分割注意力 [4]、轴向注意力 [4,18]、分解编码器 [2,23] 和可分离注意力 [43]。除了 ViT 的最小适应之外，最近的一些工作在时空变换器的设计之前引入了更多内容，包括轨迹 [26]、多尺度 [11,22] 和多视图 [39]。然而，之前的工作还没有专注于探索用于地球系统预测的时空 Transformer 的设计。

2.3 在视觉 Transformer 中的全局和局部注意力机制

为了使 self-attention 在内存消耗和速度方面更加高效，最近的工作采用了 CNN 的本质，在 Transformer 中执行局部注意力 [16,41]。HaloNets [34] 开发了一个新的自注意力模型系列，由简单的局部自注意力和卷积混合组成，在一系列下游视觉任务上优于 CNN 和普通 ViT。GLiT [5] 引入了局部性模块，并使用神经架构搜索来寻找有效的主干。焦点 Transformer [40] 提出了焦点自注意力，它可以结合细粒度的局部和粗粒度的全局交互。然而，这些架构并不直接适用于时空预测。此外，它们也与我们的设计不同，因为我们保留了 K 个全局向量来总结动态系统的统计数据并连接局部长方体。实验表明，这种全局矢量设计对于成功的时空预测至关重要。

3 模型

与之前的工作 [3,31,36] 类似，我们将地球系统预测制定为时空序列预测问题。地球观测数据，例如来自 NEXRAD [17] 的雷达回波图和来自 CIMP6 [10] 的气候数据，被表示为时空序列 $[X_i]_{i=1}^{T}$, $X_i \in \mathbb{R}^{H\times W\times C_{in}}$。基于这些观察，模型预测未来 K 步 $[Y_{T+i}]_{i=1}^{K}$, $Y_{T+i} \in \mathbb{R}^{H\times W\times C_{out}}$。这里，H、W 表示空间分辨率，$C_{in}$、$C_{out}$ 分别表示输入序列和目标序列在每个时空坐标处可用的测量数量。如图1所示，我们提出的 Earthformer 是一个基于 CuboidAttention 的分层 Transformer 编码器解码器。输入观测值被编码为隐藏状态的层次结构，然后解码为预测目标。接下来，我们将介绍 Earthformer 中采用的长方体注意力机制和分层编码器-解码器架构的详细设计。

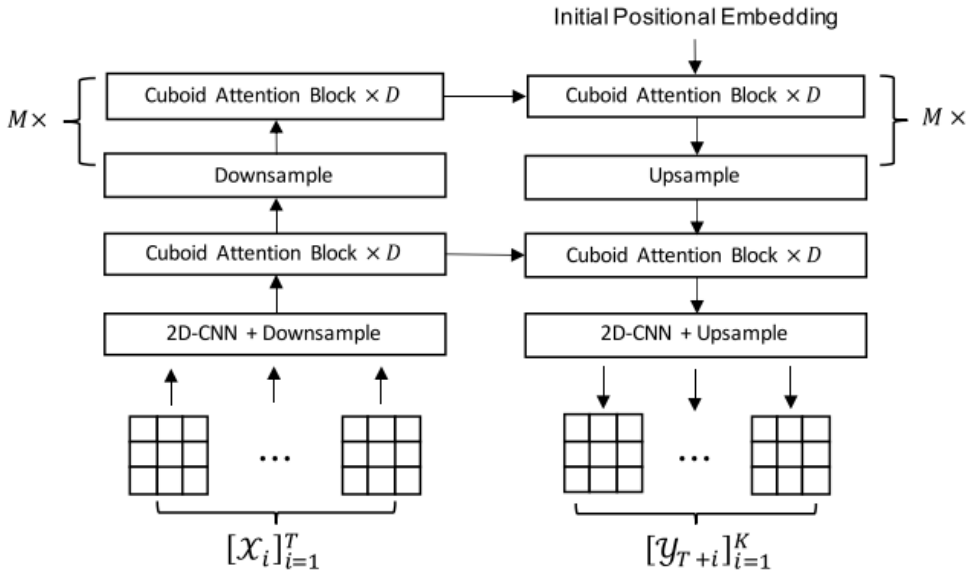


图 1. Earthformer 模型结构图

3.1 长方体注意力机制

与图像和文本相比，地球系统中的时空数据通常具有更高的维度。因此，将 Transformer 应用于此任务具有挑战性。例如，对于形状为 (T,H,W) 的 3D 张量，普通自注意力的复杂度为 $O(T^2H^2W^2)$，并且在计算上可能不可行。先前的文献提出了各种结构感知的时空注意力机制来降低复杂性 [4, 18, 22]。

这些时空注意力机制具有堆叠多个基本注意力层的共同设计，这些基本注意力层专注于不同类型的数据相关性 (例如时间相关性和空间相关性)。基于这一观察，我们提出了通用的长方体注意力层，它涉及三个步骤：“分解”、“参与”和“合并”。

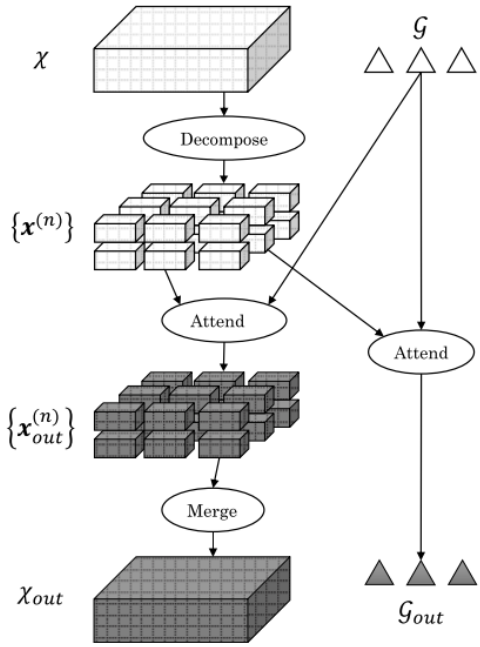


图 2. 带有全局向量的长方体注意力机制结构图

(1) 分解

我们首先将输入时空张量 $X \in \mathbb{R}^{T\times H\times W\times C}$ 分解为长方体序列 $\{x^{(n)}\}$。

$$\{x^{(n)}\} = Decompose(X, cuboid\_size, strategy, shift), \tag{1}$$

其中 cuboid_size=(b_T,b_H,b_W) 为局部长方体的大小，$strategy \in \{\text{“}local\text{”},\text{“}dilated\text{”}\}$ 控制是采用局部分解策略还是采用扩张分解策略 [4]，shift=(s_T，s_H，s_W) 是窗口移位偏移量 [21]。图3提供了三个示例，展示了如何根据 Decompose(·) 的不同超参数来分解输入张量。$\{x^{(n)}\}$ 中共有 $\lceil\frac{T}{b_T}\rceil\lceil\frac{H}{b_H}\rceil\lceil\frac{W}{b_W}\rceil$ 长方体。为了简化符号，我们假设 T、H、W 可被 b_T、b_H、b_W 整除。在实现中，如果输入张量不可整除，我们会对其进行填充。

假设 x(n) 是 $\{x(n)\}$ 中的第 (n_T, n_H, n_W) 个长方体。$x^{(n)}$ 的第 (i, j, k) 元素可以通过公式映射到 X 的第 (i', j', k') 元素。如果策略是 “本地的” 就依据公式 2，如果策略是 “扩张的” 就依据公式 3。

$$\begin{aligned} i' &\leftrightarrow s_T + b_T(n_T - 1) + i \mod T, \\ j' &\leftrightarrow s_H + b_H(n_H - 1) + j \mod H, \\ k' &\leftrightarrow s_W + b_W(n_W - 1) + k \mod W \end{aligned} \tag{2}$$

$$\begin{aligned} i' &\leftrightarrow s_T + b_T(i - 1) + n_T \mod T, \\ j' &\leftrightarrow s_H + b_H(j - 1) + n_H \mod H, \\ k' &\leftrightarrow s_W + b_W(k - 1) + n_W \mod W \end{aligned} \tag{3}$$

由于映射是双向的，因此可以通过逆运算将元素从 X 映射到 $\{x^{(n)}\}$。

(2) 参与

将输入张量分解为一系列不重叠的长方体 $\{x^{(n)}\}$ 后，我们在每个长方体中并行应用自注意力。

$$x_{out}^{(n)} = Attention_\theta(x^{(n)}, x^{(n)}, x^{(n)}), 1 \le n \le N. \tag{4}$$

$Attention_\theta(Q, K, V) = Softmax(((W_Q Q)(W_K K)^T)\sqrt{C})(W_V V)$ 的查询矩阵、键矩阵和值矩阵 Q、K 和 V 都是 $\{x^{(n)}\}$ 的扁平化版本，我们将生成的矩阵分解回 3D 张量。W_Q、W_K 和 W_V 是线性投影权重，一起缩写为 θ。自注意力参数 θ 在所有长方体之间共享。“参与” 步骤的计算复杂度为 $O(\lceil\frac{T}{b_T}\rceil\lceil\frac{H}{b_H}\rceil\lceil\frac{W}{b_W}\rceil(b_T b_H b_W)^2) \approx O(THW \cdot b_T b_H b_W)$，其与长方体大小线性缩放。由于长方体的大小可以比输入张量的大小小得多，因此该层比完全注意更有效。

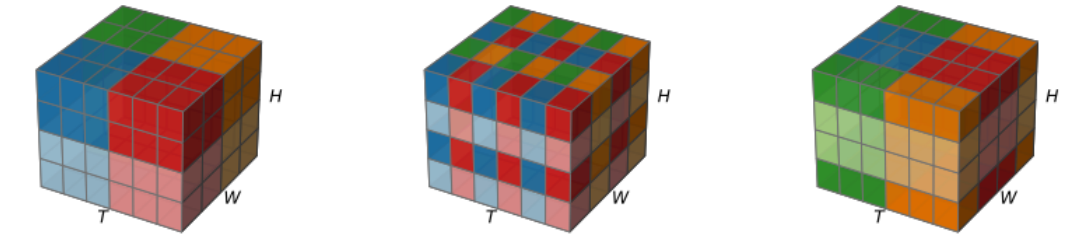
(3) 合并

合并 (·) 是分解 (·) 的逆操作。将注意力步骤 $\{x_{out}^{(n)}\}$ 后获得的长方体序列合并回原始输入形状，以产生长方体注意力的最终输出，如公式 5 所示，其映射遵循与公式 2 和公式 3 相同的映射关系。

$$X_{out} = Merge(\{x_{out}^{n}\}_n, cuboid\_size, strategy, shift). \tag{5}$$

我们结合了公式 1，4，5 中描述的 “分解”、“参与” 和 “合并” 步骤，具体如公式 6。

$$X_{out} = CubAttn_\theta(X, cuboid\_size, strategy, shift). \tag{6}$$



(a) strategy="local" shift=(0, 0, 0)　(b) strategy="dilated" shift=(0, 0, 0)　(c) strategy="local" shift = (0, 1, 1)

图 3. 输入形状为 (T,H,W) = (6, 4, 4) 且长方体大小 (bT, bH, bW) = (3, 2, 2) 时长方体分解策略图示

3.2 全局向量

先前公式的一个限制是长方体彼此之间不通信。这是不希望的，因为每个长方体无法理解系统的全局动态。因此，受 BERT [7,42] 中采用的 [CLS] 标记的启发，我们建议引入 P 个全局向量 $\mathcal{G} \in \mathbb{R}^{P \times C}$ 的集合来帮助长方体分散和收集关键的全局信息。当每个长方体执行自注意力时，元素不仅会关注同一长方体内的其他元素，还会关注全局向量 $\mathcal{G}$。我们修改公式 4 至公式 7 实现本地与全球信息交换。我们还使用公式 8 通过聚合输入张量 X 的所有元素的信息来更新全局向量 $\mathcal{G}$。

$$X_{out}^{(n)} = Attention_\theta(x^{(n)}, Cat(x^{(n)}, \mathcal{G}), Cat(x^{(n)}, \mathcal{G})), 1 \leq n \leq N. \tag{7}$$

$$\mathcal{G}_{out} = Attention\varphi(\mathcal{G}, Cat(\mathcal{G}, X), Cat(\mathcal{G}, X)). \tag{8}$$

在这里，Cat(·) 展平并连接其输入张量。通过结合公式 1,7,8,5，我们用全局向量简化了长方体注意力层的整体计算，如公式 9 所示。

$$\begin{aligned} X_{out} &= CubAttn_\theta(X, \mathcal{G}, cuboid\_size, strategy, shift), \\ \mathcal{G}_{out} &= Attn_\phi^{global}(\mathcal{G}, X). \end{aligned} \tag{9}$$

全局向量引起的额外复杂度大约为 $O(THW \cdot P + P^2)$。鉴于 P 通常很小 (在我们的实验中，P 最多为 8)，全局结构引起的计算开销可以忽略不计。长方体注意力层的架构如图2所示。

3.3 分层编码器-解码器结构

Earthformer 采用如图 2 所示的分层编码器-解码器架构。分层架构逐渐将输入序列编码为多个表示级别，并通过从粗到细的过程生成预测。每个层次堆叠 D 个长方体注意力块。编码器中的长方体注意块使用 "Divided Space-Time" 模式，解码器中的每个长方体块采用 "Axial" 模式。为了降低长方体注意力层输入的空间分辨率，我们引入了一对初始下采样和上采样模块，它们由堆叠的 2D-CNN 和最近邻插值 (NNI) 层组成。与其他采用 Transformer 进行视频预测的论文不同 [18,27]，我们以非自回归的方式生成预测，而不是逐个自回归的方式生成预测。这意味着我们的解码器直接根据初始学习的位置嵌入生成预测。

4 复现细节

4.1 与已有开源代码对比

为了更好地复现论文结果，我参考并使用了论文作者提供的开源代码https://github.com/amazon-science/earth-forecasting-transformer。在复现过程中，我仔细研究了代码的实现细节，并对其数据预处理、模型结构和训练过程等核心模块进行了深入理解和适配。

在复现过程中，我对 Earthformer 模型的超参数学习率及批量大小进行了调整，并通过实验探索添加了额外的正则化技术 Dropout 以提升模型的泛化能力。在原始代码的基础上，我还优化了训练日志记录模块，使得实验结果更加直观且便于分析。

在许多实际应用中，时空数据往往呈现出不同尺度上的变化。传统的时序建模方法大部分仅仅依赖于时域或空间域的特征，无法充分捕捉数据在不同尺度上的局部细节。时空小波变化能够在多尺度上同时分析数据的时域和空域特征，有助于揭示数据中的突变、周期性模式或局部趋势。在此动机下，我尝试在模型中引入时空小波变化，在实际实验中，通过引入时空小波变换，模型能够更好地捕捉到数据中的高频和低频特征，提升了对时空数据的建模能力。在 N-body MNIST 预测任务中，模型的预测精度得到一定的提升。

4.2 实验环境搭建

本实验在服务器环境中完成，服务器运行 Ubuntu 20.04.6 LTS 操作系统，通过 Xshell 终端进行 SSH 连接和交互。硬件配置包括 Intel Core i7-10700K CPU、NVIDIA A40 GPU (24GB 显存)、64GB 内存以及 2TB SSD。软件环境使用 Python 3.9.7 作为主要编程语言，基于 PyTorch 1.13.0 + cu117 深度学习框架进行模型复现。相关依赖库包括 Numpy、Pandas、Matplotlib、Scikit-learn、Torchvision 和 tqdm 等，依赖管理工具为 Conda 4.11.0。复现论文代码时，从 Earthformer 的开源代码库克隆代码并根据提供的 requirements.txt 文件安装相关依赖，同时对数据加载过程进行了优化，提升了训练效率。模型训练与测试在 GPU 上进行，并启用了混合精度加速。整个环境搭建严格按照开源代码要求完成，确保实验的可复现性和结果的可靠性。

4.3 界面分析与使用说明

本次实验复现采用的代码基于 Earthformer 的开源实现，整体代码结构清晰，主要包含模型定义、数据处理、训练与测试等功能模块。核心文件包括 models 文件夹 (用于模型架构与子模块实现)、data 文件夹 (提供数据处理和加载功能)、scripts 文件夹 (包含训练、验证和测试脚本)、config 文件夹 (存放配置文件) 以及 utils 文件夹 (提供工具函数)。

代码运行基于命令行交互，用户通过安装依赖、运行数据预处理脚本、指定训练参数、运行训练与评估脚本即可完成实验。在实验过程中，系统提供清晰的日志输出，方便用户实时跟踪训练进度、验证误差等信息，并且支持模型检查点的自动保存。用户界面简洁易用，能够灵活调整模型参数并进行结果可视化，方便用户分析与优化模型表现。

5 实验结果分析

在本次论文复现中，我们主要针对 N-body MNIST 和 Enso 数据集进行了实验，旨在评估模型在不同数据集上的表现，并分析其预测能力和泛化性能。

5.1 N-body MNIST

地球是一个复杂的系统，其中有大量的变量相互作用。与地球系统相比，合成的 Moving MNIST 数据集的动力学 (其中数字以恒定速度独立移动) 过于简化。因此，在 Moving MNIST 上取得良好性能并不意味着该模型能够对地球系统中复杂的相互作用进行建模。另一方面，现实世界的地球观测数据虽然经历了快速发展，但仍然存在噪音，可能无法为模型开发提供有用的见解。因此，我们将 Moving MNIST 扩展到 N-body MNIST，其中 N 个数字在 64×64 帧内按照 N 体运动模式移动。每个数字都有其质量并受到其他数字的重力。我们在实验中选择 N=3，以便数字遵循混沌三体运动 [33]。N-body MNIST 中的高度非线性交互使其比原始的 Moving MNIST 更具挑战性。我们生成 20,000 个序列用于训练，1,000 个用于验证，1,000 个用于测试。表1将 Earthformer 的性能与 N-body MNIST 数据集的基线进行了比较。

表 1. N-body MNIST 数据集上的性能比较

Model	#Param. (M)	GFLOPS	MSE ↓	MAE ↓	SSIM ↑
UNet [36]	16.2	0.8	37.83	93.11	0.8143
ConvLSTM [31]	13.8	28.8	31.66	71.55	0.8874
PredRNN [38]	22.8	230.8	21.54	53.85	0.9244
PhyDNet [13]	3.3	14.9	27.99	77.86	0.8103
E3D-LSTM [37]	12.3	299.4	22.63	61.48	0.9064
Rainformer [3]	18.8	1.1	38.77	95.78	0.8011
Earthformer w/o global	6.2	33.4	<u>15.46</u>	<u>41.06</u>	<u>0.9489</u>
Earthformer	7.4	33.7	**14.27**	**38.76**	**0.9518**

5.2 ICAR-ENSO

厄尔尼诺/南方涛动 (ENSO) 与区域极端气候和生态系统影响有着广泛的关联。ENSO 海面温度 (ST) 异常预报的提前期长达一年 (12 个步长)。Nino3.4 指数是太平洋某一区域 (170°-120°W，5°S-5°N) 的区域平均海温距平，是此次气候事件的重要指标。预测质量通过三个月移动平均 Nino3.4 指数 $C^{Nino3.4} = \frac{\sum_N (X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum_N (X-\bar{X})^2 \sum_N (Y-\bar{Y})^2}} \in \mathbb{R}^K$ 进行评估，在大小为 N 的整个测试集上计算，其中 $Y \in \mathbb{R}^{N\times K}$ 是 K 步 Nino3.4 的 ground-truth,$X \in \mathbb{R}^{N\times K}$ 是 Nino3.4 对应的预测。

ICAR-ENSO 由气候和应用研究所 (ICAR) 提供的历史气候观测和模拟数据组成。在 12 步 SST 距平观测的背景下，我们预测了 14 步 SST 距平 (计算三个月移动平均值时比一年多 2 步)。表2将 Earthformer 的性能与 ICAR-ENSO 数据集的基线进行了比较。我们报告了

K=12 个预测步骤的平均相关技能 $C-Nino3.4-M=\frac{1}{K}\sum_K C_k^{Nino3.4}$，以及加权平均相关技能 $C-Nino3.4-WM=\frac{1}{K}\sum_k a_k\cdot C_k^{Nino3.4}$，以及时空海表温度异常预测与相应的地面实况之间的 MSE。我们可以发现 Earthformer 在所有方面都始终优于基线。同时在实验中可以发现，评估指标和使用全局向量进一步提高了模型性能。

表 2. ICAR-ENSO 数据集上的性能比较

Model	#Param. (M)	GFLOPS	C-Nino3.4-M ↑	C-Nino3.4-WM ↑	MSE (10^{-4}) ↓
Persistence	-	-	0.3214	0.449	4.552
UNet [36]	12.3	0.6	0.6938	2.085	2.859
ConvLSTM [31]	14.2	11.4	0.6945	2.102	2.649
PredRNN [38]	23.2	85.5	0.6434	1.906	3.053
PhyDNet [13]	3.5	5.4	0.6656	1.958	2.712
E3D-LSTM [37]	11.8	99.2	0.7047	2.118	3.091
Rainformer [3]	18.6	1.1	0.7112	2.147	3.039
Earthformer w/o global	6.6	23.2	0.7221	2.209	2.544
Earthformer	7.6	23.6	**0.7317**	**2.246**	**2.542**

6 总结与展望

在本文中，我们提出了 Earthformer，一种用于地球系统预测的时空 Transformer。Earthformer 基于称为 CuboidAttention 的通用且高效的构建块。它在我们新提出的 N-body MNIST 和 ICAR-ENSO 上实现了 SOTA。

我们的工作有一定的局限性。第一个是 Earthformer 是一个确定性模型，不会模拟不确定性。这可能会导致预测所有可能未来的平均值，从而导致模型生成低感知质量的模糊预测，并且缺乏有价值的小规模细节。事实上，业界缺乏衡量地球系统预测模型中不确定性成分的适当指标。将 Earthformer 扩展到概率预测模型可能是一个令人兴奋的未来方向。第二个是该模型纯粹是数据驱动的，没有利用地球系统的物理知识。最近关于添加物理约束 [19, 24] 以及整合数据驱动模型和基于物理的模型 [29] 的预测的研究表明，这是一个积极且有前途的研究方向。我们计划未来研究如何将物理知识融入到 Earthformer 中。

参考文献

[1] Tom R. Andersson, J. Scott Hosking, María Pérez-Ortiz, Brooks Paige, Andrew Elliott, Chris Russell, Stephen Law, Daniel C. Jones, Jeremy Wilkinson, Tony Phillips, and et al. Seasonal arctic sea ice forecasting with probabilistic deep learning. *Nature communications*, 12(1):1–12, 2021.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Luvcić, and Cordelia Schmid. Vivit: A video vision transformer. In *International Conference on Computer Vision (ICCV)*, 2021.

[3] Cong Bai, Feng Sun, Jinglin Zhang, Yi Song, and Shengyong Chen. Rainformer: Features extraction balanced network for radar-based precipitation nowcasting. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding. *arXiv preprint arXiv:2102.05095*, 2(3):4, 2021.

[5] Boyu Chen, Peixia Li, Chuming Li, Baopu Li, Lei Bai, Chen Lin, Ming Sun, Junjie Yan, and Wanli Ouyang. Glit: Neural architecture search for global and local image transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–21, 2021.

[6] Christian Schroeder de Witt, Catherine Tong, Valentina Zantedeschi, Daniele De Martini, Freddie Kalaitzis, Matthew Chantry, Duncan Watson-Parris, and Piotr Bilinski. Rainbench: towards global precipitation forecasting from satellite imagery. In *AAAI*, 2021.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Lasse Espeholt, Shreya Agrawal, Casper Sønderby, Manoj Kumar, Jonathan Heek, Carla Bromberg, Cenk Gazen, Jason Hickey, Aaron Bell, and Nal Kalchbrenner. Skillful twelve hour precipitation forecasts using large context neural networks. In *arXiv preprint arXiv:2111.07470*, 2021.

[10] Veronika Eyring, Sandrine Bony, Gerald A Meehl, Catherine A Senior, Bjorn Stevens, Ronald J Stouffer, and Karl E Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016.

[11] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021.

[12] Steven J Goodman, Timothy J Schmit, Jaime Daniels, and Robert J Redmon. *The GOES-R series: a new generation of geostationary environmental satellites*. Elsevier, 2019.

[13] Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11474–11484, 2020.

[14] John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive fourier neural operators: Efficient token mixers for transformers. In *arXiv preprint arXiv:2111.13587*, 2021.

[15] Yoo-Geun Ham, Jeong-Hwan Kim, and Jing-Jia Luo. Deep learning for multi-year enso forecasts. *Nature*, 573(7775):568–572, 2019.

[16] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. In *NeurIPS*, 2021.

[17] William H Heiss, David L McGrew, and Dale Sirmans. Nexrad: next generation weather radar (wsr-88d). *Microwave Journal*, 33(1):79–89, 1990.

[18] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019.

[19] Aditi Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, and Michael W Mahoney. Characterizing possible failure modes in physics-informed neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 26548–26560, 2021.

[20] Christophe Letellier. *Chaos in nature*, volume 94. World Scientific, 2019.

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[22] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021.

[23] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. *arXiv preprint arXiv:2102.00719*, 2021.

[24] Geoffrey Négiar, Michael W Mahoney, and Aditi S Krishnapriyan. Learning differentiable solvers for systems with hard constraints. *arXiv preprint arXiv:2207.08675*, 2022.

[25] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, and et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

[26] Mandela Patrick, Dylan Campbell, Yuki M. Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F. Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[27] Ruslan Rakhimov, Denis Volkhonskiy, Alexey Artemov, Denis Zorin, and Evgeny Burnaev. Latent video transformer. *arXiv preprint arXiv:2006.10704*, 2020.

[28] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831, 2021.

[29] Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, and et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021.

[30] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and et al. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019.

[31] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NeurIPS*, volume 28, 2015.

[32] Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai kin Wong, and Wang chun Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. In *NeurIPS*, volume 30, 2017.

[33] Mauri Valtonen and Hannu Karttunen. *The three-body problem*. Cambridge University Press, 2006.

[34] Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. In *CVPR*, 2021.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, volume 30, 2017.

[36] Mark Veillette, Siddharth Samsi, and Chris Mattioli. Sevir: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. In *Advances in Neural Information Processing Systems*, volume 33, pages 22009–22019, 2020.

[37] Yunbo Wang, Lu Jiang, Ming-Hsuan Yang, Li-Jia Li, Mingsheng Long, and Li Fei-Fei. Eidetic 3d lstm: A model for video prediction and beyond. In *International Conference on Learning Representations*, 2018.

[38] Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, Philip Yu, and Mingsheng Long. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[39] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *CVPR*, 2022.

[40] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. In *NeurIPS*, 2021.

[41] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[42] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and et al. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297, 2020.

[43] Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. Vidtr: Video transformer without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13577–13587, 2021.