

《Climax: 一个天气和气候预测的基础模型》复现

叶毅强

摘要

大多数最先进的天气和气候建模方法都是基于大气的物理信息数值模型。这些方法旨在模拟多个变量之间的非线性动力学和复杂的相互作用，这很难近似。此外，许多这样的数值模型计算量大，尤其是在以细粒度空间和时间分辨率对大气现象进行建模时。最近基于机器学习的数据驱动方法旨在通过使用深度神经网络学习数据驱动的函数映射来直接解决下游预测或投影任务。然而，这些网络是针对特定时空任务的策划和同质气候数据集进行训练的，因此缺乏数值模型的通用性。本文复现和改进了 Climax，这是一种灵活且可推广的天气和气候科学深度学习模型，可以使用跨越不同变量、时空覆盖和物理基础的异构数据集进行训练。Climax 使用新颖的编码和聚合块扩展了 Transformer 架构，允许在保持通用效用的同时有效利用可用计算。Climax 使用来自 CMIP6 的气候数据集的自我监督学习目标进行预训练。然后可以对预训练的 Climax 进行微调以解决广泛的气候和天气任务，包括那些涉及预训练期间看不见的大气变量和时空尺度的任务。与现有的数据驱动基线相比，本文在原论文的基础上，主要增加了自编码模块和 case-base 检索模块，表明即使以较低分辨率和计算预算进行预训练，Climax 中的这种通用性在天气预报和气候预测基准上产生了卓越的性能。

关键词：天气和气候预测；基础模型；

1 引言

天气和气候建模是科学和社会的一个无所不在的挑战。随着对极端天气事件和气候变化的日益关注，对长期政策制定和适应努力的减灾和气候预测的改进天气预报的需求越来越大 [5]。目前，天气和气候全球建模的数值方法是通过各种一般环流模型 (GCM) 参数化的 [11]。GCMC 代表了将能量和物质流动随时间积分的大气、陆地和海洋联系起来的微分方程系统，以获得相关大气变量的预测 [1, 11]。虽然在实践中非常有用，但 GCM 也面临着许多挑战，例如准确地表示精细分辨率的物理过程和初始条件，以及大规模数据同化和计算模拟中的技术挑战。这些因素限制了它们在许多情况下的使用，特别是在以非常短的时间尺度 (例如，几个小时) 快速模拟大气变量或在长时间尺度上准确 (例如，超过 5-7 天) [17]。

相比之下，数据驱动的大气变量预测方法稳步上升，特别是对于气象应用。这里的关键思想是训练深度神经网络使用几十年的历史全局数据集预测目标大气变量，例如 ERA-5 再分析数据集 [8]。与 GCMC 不同，这些网络没有明确基于物理学，并且缺乏对地球系统科学的通用效用，因为它们针对特定的预测建模任务进行了训练。然而，随着计算和数据集的增长，越来越多的证据表明，这些模型在许多情况下可以达到与最先进的数值模型竞争的准确

性，例如降水临近预报和温度、风湿度等变量的中等范围预测。虽然这些趋势令人鼓舞，但关于这种数据驱动方法对不同现实场景的通用性仍然存在担忧，例如极端天气事件的预测和长期气候预测，特别是在有限的时空监督和计算预算下。

论文受上述成功的启发，这项工作研究了这个问题：我们如何设计和训练天气和气候的基础模型？该模型可以有效地适应有关地球大气的通用任务。论文提出了 ClimaX，这是天气和气候的基础模型。为了预训练任何基础模型，关键是使用无监督目标在大型数据集上训练深度架构。例如，语言和视觉的许多基础模型使用生成建模在互联网规模的数据集上训练大型转换器。虽然在概念上很简单，但这种缩放方式充满了来自天气和气候领域的挑战。我们希望通过 climax 天气和气候领域的第一个基础模型，挖掘并有效利用其学习到的知识，进一步增强低分辨率天气数据的预测能力。

2 相关工作

2.1 数值方法

数值方法是一种根据大气的数学模型，包含一连串的偏微分、物理及化学公式，利用当前的气压、温度、风速、湿度作为输入数据而作出天气预报的方式。难点在于以精细分辨率准确地表示物理过程和初始条件，以及大规模数据同化和计算模拟中的技术挑战。当今使用的当前天气和气候模型广泛依赖于数值方法和计算模拟来预测和理解地球天气和气候系统。这些任务包括各种数值天气预报 (NWP) 系统，该系统使用计算机模拟对天气条件进行短期预测，以及气候模型，这些模型使用类似的技术来模拟和预测地球气候的长期变化。最值得注意的是，天气和气候模型的核心是同一组原始方程。

对于气候建模，地球系统模型 (ESM) [9] 或“耦合模型”将控制大气、冰冻圈、陆地和海洋过程的模拟结合在一起被认为是最先进的。这些模拟主要基于一般循环模型 (GCMs)，可以追溯到 Phillips [12] 和 Lorenz [10] 在旋转球上求解 Navier-Stokes 方程来模拟流体循环。这些模型通常用于执行各种因素敏感性研究，以检查温室气体浓度等某些强迫因素的变化如何影响全球或地区气候，并有助于气候预测以帮助理解未来的条件。

数值天气预报 (NWP) 模型共享 GCM 的许多组成部分，特别是大气成分。然而，结合数据同化，涉及结合大气和海洋的观测和各种测量以及这些数值模型对于准确预测和模拟很重要。天气和气候模型之间的另一个显着区别是基础方程的解的框架：天气预报——初始值问题和气候——边界值问题。这些解决方案方法的不同难度导致了这样一个事实，即气候模型往往是全球性的，通常具有较粗略的时空分辨率，而天气模型的范围从全球到地方和区域，具有非常高的时空分辨率。

2.2 基础模型

Bommasani, Hudson 等人 [2] 将“基础模型”一词赋予了通过自我监督在广泛数据上训练可扩展深度学习模型的新兴范式，然后可以将其适应（通常通过微调）广泛的下游任务。当前值得注意的例子包括 BERT [6]、GPT [3] 和 PaLM [4]，在语言中，CLIP [13]、Florence [16]、BEiT [15]，用于视觉语言。在网络数据抓取的外部应用程序中，这种范式也开始在蛋白质设计等各个科学领域取得成功 [14]。这些模型的关键意义已被确定为在大规模迁移学习原则的

支持下，在模型能力方面的出现，以及在不同任务、领域和模式的方法方面的同质化。虽然基础模型本身应该被认为是不完整的，但它可以提供一个共同的基础，从中可以派生出各种特定任务的模型。目前，天气和气候科学与机器学习领域交叉的研究主要集中在为每一项感兴趣的任務设计单独的模型上，尽管这些任务可能有相当多样化的大规模数据，并具有共享的基础物理和地质。最近的一些工作提出了卫星图像和遥感的预训练技术，但到目前为止，它们还没有应用于天气和气候的多感受数据和变量。

3 本文方法

3.1 本文方法概述

本文在 Vision Transformers(ViT) [7] 的基础上构建了 ClimaX 架构，并复现了两个主要的架构变化，即通道标记化和变量聚合，以进一步提高灵活性和通用性。进一步微调了基于 ERA5 数据集的全球预测任务。ClimaX 预训练模型框架如1所示：

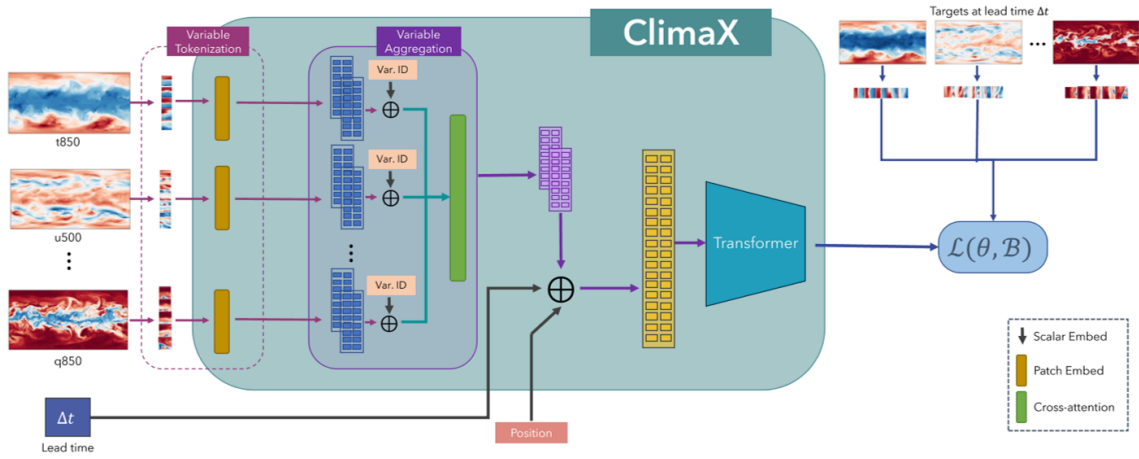


图 1. 模型框架示意图

任务：预测未来任意时间的任意一组输入变量，可以理解为图片到图片的转换问题。

输入：数据集 CMIP6, $V * H * W$, 其中 V 是输入变量的个数, H 和 w 取决于分辨率

模型架构：变量使用变量分离标记化进行编码，然后使用变量聚合进行聚合。连同位置嵌入和引导时间嵌入，这些嵌入被馈送到 ViT 主干。

3.2 通道标记化

给定形状的输入 $V * H * W$, ViT 分词器将输入转化为为 $(H/p) * (W/P) = h * w$ 的块，每个块的大小为 $V * p^2$, 其中 p 是块的大小。这种标记化方案适用于图像数据，因为 V 始终是 RGB 通道，这对于所有数据集都是相同的。然而，气候和天气数据并非如此，因为不同数据集之间的物理变量数量可能不同。例如，在 CMIP6 项目中，每个数据集都包含不同气候模型的模拟数据，因此具有不同的物理变量。因此，我们提出了变量标记化，这是一种新的标记化方案，可以分别标记输入中的每个变量。具体来说，每个通道的输入变量都是形状的空间映射 $H * W$ 被标记为 $h * w$ 补丁，这导致 $V * h * w$ 的 patch 总数。最后，每个大小的输入

$\text{patch}p^2$ 线性嵌入到维度为 D 的向量, 其中 D 是所选择的 embedding 大小。因此, 变量标记化模块的输出具有的维度为 $V * h * w * D$ 。

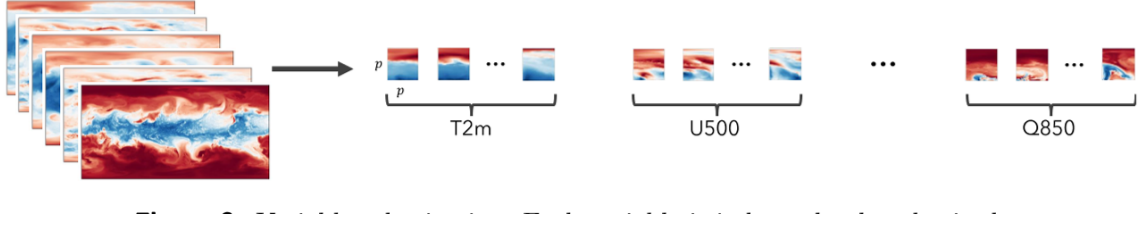


图 2. 通道标记化示意图

3.3 变量聚合

对于每个空间位置 $h * w$ 映射, 我们执行一个交叉注意力操作, 其中查询是一个可学习的向量, 键和值是该位置的变量 V 的嵌入向量。交叉注意力模块为每个空间位置输出一个向量, 从而将序列长度减少到 $h * w$, 显著降低了计算成本。此外, 该序列现在包含具有通用语义的统一标记, 为注意力层创建了一个更简单的任务。

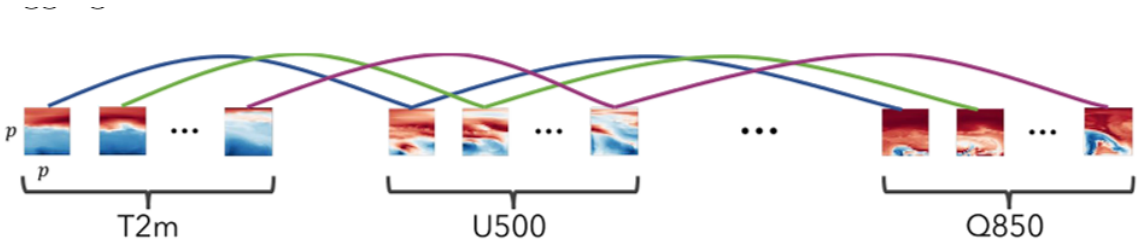


图 3. 模型框架示意图

4 复现细节

4.1 与已有开源代码对比

由于本次实验复现的论文已开源, 我们结合论文的基本框架和计算公式, 参考开源代码进行复现。使用 AutoEncoder 对预训练模型编码的特征进行压缩并存储在 KB 内, 使用 Case-Base 的思想, 融合过去的经验指导预测结果。如图5所示:

4.2 实验环境搭建

实验环境构建主要是 ERA5 数据集的下载和开源代码的部署, 包括模型权重的下载和数据集的预处理, 具体可见文件 readme.md。

在全球气候预测的任务中, 输入数据 t , 预测 $t + \Delta t$ 的气候数据, 其中本次实验设置输入变量包括 7 个垂直高度的 6 个大气变量、3 个地表变量和 3 个恒定场, 共 48 个输入变量, 输出变量分别是 {500hPa 的位势 (Z500)、850hPa 的温度 (T850)、距离地面 2 米的温度 (T2m) 和距离地面 10 米的纬向风速 (U10)}。

4.3 创新点

受 Case-Base 的思想启发，即利用过去的经验指导当前的预测，本次复现首先建立一个知识库 KB ，其中知识库 KB 内存储 $(Data_t, CData_t)$ 的二元组， t 表示观测时刻。

自编码器是一种高效的数据压缩方式，其模型如4所示。输入数据经过 AE Encoder 下采样和 AE Decoder 的上采样还原输入数据本身，经过 AE Encoder 编码的低维嵌入压缩输入数据的信息，作为 $Data_t$ 的低维表示 $CData_t$ ，从而降低检索过去经验的计算开销，同时压缩后的信息提高鲁棒性，减少噪声对模型检索结果的影响。

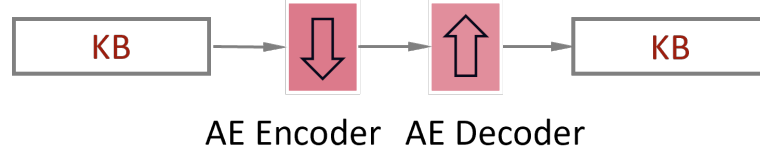


图 4. AE 示意

出于帮助模型充分利用过去经验和已经学习到的知识的目标，我们基于 Case-base 的思想输入经过 Climax 的 encoder 编码 $Data_t$ ，在经过余弦相似度比较知识库存储的 $CData_{t'}$ 后，选取 top-k 个相似度最高的原型，并读取 $CData_{t'+\Delta t}$ 的真实结果，最后 $CData_{t'+\Delta t}$ 和 $CData_t$ 经过两层线性层连接的解码层后输出预测结果。

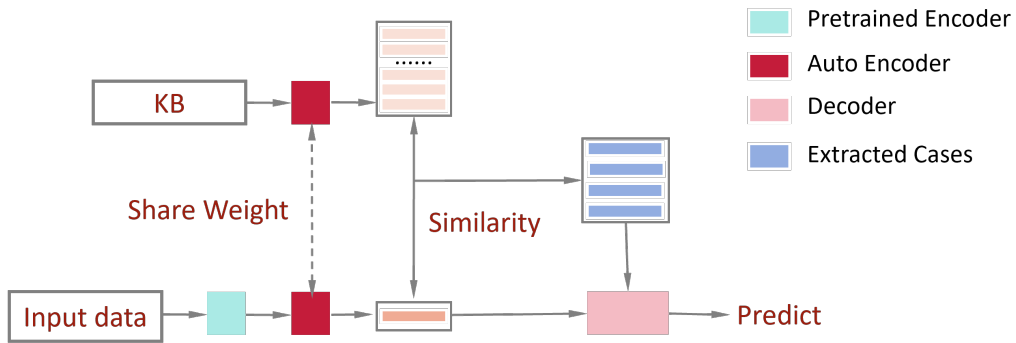


图 5. 改进方法示意

5 实验结果分析

实验结果如1所示，可以发现，经过基于 Case 比较的改进后，各预测变量的准确率均有小幅度的提升，但是对模型预测的效果的提升有限。通过实验过程中发现，这可能是与时间序列的不平稳性有关，同时说明了利用过去的经验可以提高模型预测的准确性。

Metric	climax	Ours
test/acc	0.738	0.767
test/acc 10m u component of wind 7 days	0.492	0.556
test/acc 2m temperature 7 days	0.890	0.896
test/acc geopotential 500 7 days	0.765	0.796
test/acc temperature 850 7 days	0.806	0.821
test/w mse 0.129 0.115		
test/w mse geopotential 500 7 days	0.434	0.385
test/w mse temperature 850 7 days	0.013	0.012
test/w rmse	0.034	0.830
test/w mse 10m u component of wind 7 days	0.034	0.832
test/w mse 2m temperature 7 days	163.016	151.985
test/w rmse 10m u component of wind 7 days	3.651	3.437
test/w rmse 2m temperature 7 days	2.380	2.292
test/w rmse geopotential 500 7 days	643.139	599.429
test/w rmse temperature 850 7 days	2.896	2.783

表 1. 实验结果

6 总结与展望

本次前沿工作复现 ClimaX，一个天气气候预测的基础模型。通过本次实验，对天气气候预测的原理和相关工作有了更深的认识，同时，基础模型的潜力在自然语言领域外，在各种科学自然领域都有着巨大的价值，希望后续工作能将数据驱动的和基于物理建模的模型相互融合，相互映证。基于 Case 比较的改进对模型预测的效果的提升有限，实验过程中发现，这可能是与时间序列的不平稳性有关，如果仅利用后二十年的数据微调，与使用全体数据微调相比，预测性能有着一定提升，下一步将针对数据集的时空属性和现有数值物理建模的方法进行探索，希望能得到对自然科学有价值的发现。

参考文献

- [1] Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, page 47–55, Sep 2015.
- [2] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- [3] T.B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Askell Amanda, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Henighan Tom, Rewon Child, A. Ramesh, DanielM. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, EricJ.

- Sigler, Mateusz Litwin, Scott Gray, Chess Benjamin, Jack Clark, Christopher Berner, McCandlish Sam, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv: Computation and Language*, *arXiv: Computation and Language*, May 2020.
- [4] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, HyungWon Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, and Liam Fe. Palm: Scaling language modeling with pathways.
 - [5] GrupoIntergubernamentaldeExpertossobreelCambio Climático. Climate change 2021 : the physical science basis. Jan 2021.
 - [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, Jan 2019.
 - [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv: Computer Vision and Pattern Recognition*, *arXiv: Computer Vision and Pattern Recognition*, Oct 2020.
 - [8] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.
 - [9] James W. Hurrell, M. M. Holland, P. R. Gent, S. Ghan, Jennifer E. Kay, P. J. Kushner, J.-F. Lamarque, W. G. Large, D. Lawrence, K. Lindsay, W. H. Lipscomb, M. C. Long, N. Mahowald, D. R. Marsh, R. B. Neale, P. Rasch, S. Vavrus, M. Vertenstein, D. Bader, W. D. Collins, J. J. Hack, J. Kiehl, and S. Marshall. The community earth system model: A framework for collaborative research. *Bulletin of the American Meteorological Society*, 94(9):1339–1360, Sep 2013.
 - [10] EdwardN. Lorenz. The nature and theory of the general circulation of the atmosphere. *World Meteorological Organization eBooks*, *World Meteorological Organization eBooks*.
 - [11] Peter Lynch. The origins of computer weather prediction and climate modeling. *Journal of Computational Physics*, 227(7):3431–3444, Mar 2008.

- [12] NORMAN A. PHILLIPS. *THE GENERAL CIRCULATION OF THE ATMOSPHERE: A NUMERICAL EXPERIMENT*, page 18–25.
- [13] Alec Radford, JongWook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Askell Amanda, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *Cornell University - arXiv, Cornell University - arXiv*, Feb 2021.
- [14] Robert Verkuil, Ori Kabeli, Yilun Du, Basile I. M. Wicky, Lukas F. Milles, Justas Dau-paras, David Baker, Sergey Ovchinnikov, Tom Sercu, and Alexander Rives. Language models generalize beyond natural proteins. Dec 2022.
- [15] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks.
- [16] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision.
- [17] Fuqing Zhang, Y. Qiang Sun, Linus Magnusson, Roberto Buizza, Shian-Jiann Lin, Jan-Huey Chen, and Kerry Emanuel. What is the predictability limit of midlatitude weather? *Journal of the Atmospheric Sciences*, 76(4):1077–1091, Apr 2019.