利用可解释性指导基于深度学习的统计降尺度，超越标准验证方法

Jose Gonz´ález-Abad[1],**Jorge Ba˜ño-Medina**[1],**Jos´é Manuel Guti´érrez**[1]

[1]坎塔布里亚物理研究所（IFCA）。CSIC - 坎塔布里亚大学，桑坦德，西班牙

arXiv:2302.01771v1 [stat.ML] 3 Feb 2023

要点：

- 深度学习在降尺度方面展现出潜力，但由于缺乏可解释性，其在气候变化领域的应用仍缺乏可信度
- 可解释人工智能（XAI）通过揭示深度降尺度模型的内部行为，促进对这些模型的评估
- 可解释人工智能技术能够检测出标准评估方法无法揭示的结构性问题

通讯作者：Jose Gonz´alez-Abad，gonzabad@ifca.unican.es

摘要

深度学习（DL）已成为一种前景广阔的工具，可根据完美预报（PP）方法，从大尺度大气场将气候预测降尺度至区域—局地尺度。鉴于这些方法的复杂性，对其进行适当评估至关重要，尤其是在气候变化条件下应用时，因为此时外推和泛化能力是关键。在本研究中，我们针对同一项具有挑战性的应用（对 CORDEX 北美域的温度进行降尺度），对从文献中提取的多个深度学习模型进行了比较，并在标准评估方法的基础上，结合可解释人工智能技术加以拓展。我们展示了如何利用这些技术揭示模型的内部行为，从而提供新的评估维度，并有助于模型的诊断与设计。这些结果表明，将可解释人工智能技术纳入统计降尺度评估框架十分有益，尤其是在处理大范围区域和/或气候变化条件下的问题时。

通俗摘要

由于可用计算资源有限，全球环流模式（GCM）通常采用粗分辨率网格模拟气候系统。这限制了全球气候模式产品在区域至局地尺度上的适用性，而不同社会经济部门对此有着迫切需求。统计降尺度旨在通过生成高分辨率气候场来解决这一问题。近年来，机器学习技术（尤其是深度学习模型）在这项任务中展现出良好的前景。这些模型首先利用观测数据集在历史时期进行训练，然后应用于全球气候模式对遥远未来情景的输出，从而生成高分辨率气候变化产品。为评估所生成降尺度气候场的合理性，研究者采用了多种验证框架（例如，评估模型再现当前气候的能力），以检验模型的泛化能力。在此，我们提出一种基于可解释人工智能（XAI）的新型评估方案，用于考察某些深度学习模型是否适用于气候降尺度。我们以北美洲温度降尺度为案例展开分析。与标准验证方法不同，后者的模型间比较结果显示模型之间没有显著差异；我们发现，XAI 方案有助于诊断模型应用于遥远未来情景时可能出现的相关失效。

1 引言

全球气候模式（GCM）模拟历史时期和未来排放情景驱动下气候系统的时空演变（Eyring 等，2016）。然而，计算和物理方面的局限制约了全球气候模式输出的空间分辨率（最先进的模拟为 100–200 km），并导致其无法准确表征局地过程（Maraun & Widmann，2018）。这些局限阻碍了全球气候模式在区域尺度上的应用，而区域尺度的气候信息对于开展影响研究以及制定不同社会经济部门的适应计划需求迫切。

经验统计降尺度（ESD）通过学习一组大尺度大气场（预测因子）与感兴趣的区域/局地变量（预测量）之间的模型/关联来弥补这一差距（Maraun & Widmann，2018）。在完美预报（PP）方法下，统计模型利用同期的预测因子—预测量观测记录（通常为再分析资料和网格化/站点观测），在表征近期气候的 20–30 年时段内进行训练。随后，将全球气候模式模拟中的预测因子（例如历史时期和未来预估）输入所得模型，以生成降尺度信息。

深度卷积神经网络（卷积神经网络，（Goodfellowetal., 2016年；LeCun 等，1995年））近期被引入，成为一种颇具前景的完美预报降尺度技术，因为

它们能够从预测因子场中自动推断编码预测信息的空间特征。不同研究考察了这些方法在当前（Pan 等人，2019；Ba˜no-Medina 等，2020；Sun & Lan，2021）和未来时期气候条件（Ba˜no-Medina 等，2021年、2022年）下用于降尺度的表现，并取得了令人满意的结果。此外，这些进展促成了遵循 CORDEX 协议的大陆范围预估，构成了首个采用 PP 统计降尺度方法开展的大陆范围 CORDEX 项目（Ba˜no-Medina 等，2022年）。然而，卷积神经网络的复杂性和黑箱特性，引发了人们对其广泛应用于生成可用于决策的降尺度结果的担忧。具体而言，现有评估方法无法解决这些问题，因为它们无助于诊断和解释与模型选择（不同的组成部分和设计）相关的潜在问题，例如结构缺陷，或模型外推和泛化的能力。

可解释人工智能（XAI）领域（Zhang et al., 2021; Buhrmester et al., 2021; Das & Rad, 2020）近来兴起，旨在回应日益增长的需求，即解释深度学习模型的运作方式及其结果，使其内部结构能够以人类易于理解的方式呈现。这些技术在降尺度领域的应用尚处于起步阶段，但已取得一些令人鼓舞的成果（Ba˜no-Medina, 2020; Rampal et al., 2022），证明卷积神经网络能够识别相关预测因子和影响空间区域，从而在其隐藏层中自动选择预测因子。

本文采用可解释人工智能技术评估降尺度方法，通过新的诊断指标揭示模型的内部结构及各要素的重要性，从而改进并补充标准验证方法（Maraun et al., 2015）。我们聚焦于一个具有挑战性的问题（对 CORDEX 北美洲区域的温度进行降尺度），并通过比较试验评估文献中提出的多种深度学习方法，以此展示所提出的框架。我们通过应用新的基于可解释人工智能的指标，评估这些方法在当前和未来气候下的表现，超越传统验证指标，并帮助识别标准验证方法无法发现的结构性问题。

本文结构如下：第2节介绍本文使用的数据、相互比较的模型以及可解释人工智能技术。第3节描述实验结果，并重点介绍可解释人工智能技术的应用。最后，第4节总结本文的主要结论。

2 实验框架

2.1 研究区域和数据

我们研究北美洲 CORDEX 区域（12°-70° N 和 60°-165° W）的近地表气温，使用规则网格的 0.5° EWEMBI日数据集（Lange，2019）作为统计降尺度的目标预测变量。图1展示了研究区域内 EWEMBI 温度的第2（P02）和第98（P98）百分位数，以及均值和标准差；还展示了该区域内的 IPCC参考区域（Iturbide 等，2020），本文将据此给出区域结果。该图显示了该区域复杂的气候态：均值、极端值和变率存在显著的纬向梯度（例如，均值从北部接近 -15℃ 到南部 20℃）。这使得在这片横跨整个大陆的区域进行降尺度颇具挑战性，因为必须同时应对北部的极地气候与南部的热带气候，以及两地的高、低变率。

在选择预测因子时，我们遵循以往文献（Ba˜no-Medina 等，2020，2021），采用四个垂直层次（1000、850、700 和 500 百帕）上的五个大尺度变量（位势高度、纬向风和经向风、气温及比湿），时间尺度为每日。ERA-Interim 再分析（Dee 等，2011）（重网格化至

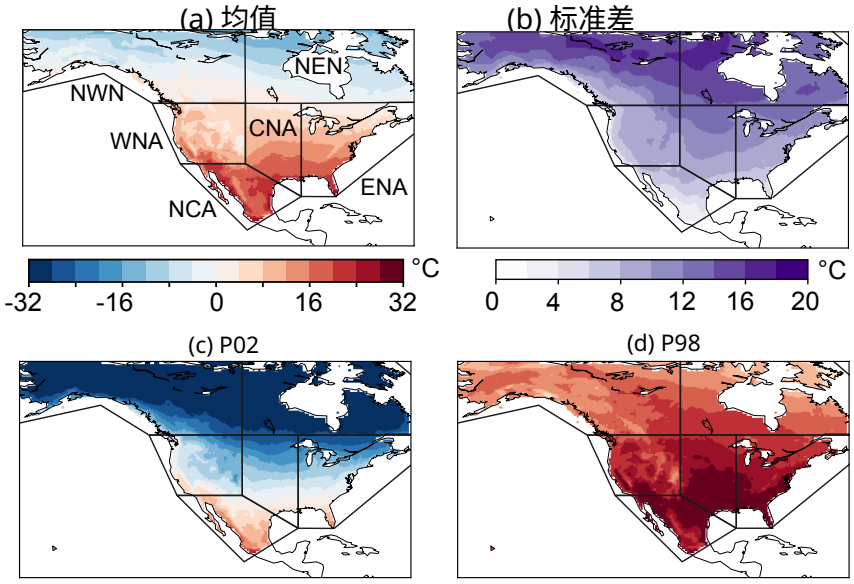

图1. EWEMBI近地表气温在1980—2008年期间的气候态，包括均值、标准差以及第2百分位数（P02）和第98百分位数（P98）。各面板展示了该区域内五个IPCC次大陆参考区域，这些区域在面板(a)中从左到右、从上到下依次标记为：北美西北部（NWN）、北美东北部（NEN）、北美西部（WNA）、北美中部（CNA）、北美东部（ENA）和中美洲北部（NCA）。

水平分辨率为 $2°$，采用双线性插值）与 EWEMBI 预测量一同用于训练，训练依据为共同时间段 1980—2008年内的每日信息。

采用 EC-Earth 模型模拟 $r12i1p1$ (Doblas Reyes 等，2018)，在历史时期（1980—2005年）和未来时期（2006—2100年，基于代表性浓度路径8.5（RCP8.5）情景）条件下测试所得的降尺度模型。已有研究表明，EC-Earth 模型能够恰当地再现关键的大尺度大气环流型，因此已被用于多项降尺度研究中的模式评估（Manzanas 等，2018；Ba˜no-Medina 等，2021）。此外，之所以选择 RCP8.5 情景，是因为在第五阶段耦合模式比较计划（CMIP5）制定的情景中，该情景所描述的气候变化信号最强。与 ERA-Interim 的处理方式相同，我们使用双线性插值将 EC-Earth 预测因子从其原始分辨率（$1.12°$）重网格化至 $2°$。

考虑到 PP假设要求全球气候模式预测因子必须得到逼真的模拟（在本例中，是与 ERA-Interim 再分析进行比较），我们遵循（Ba˜no-Medina 等，2022年）的方法，对全球气候模式预测因子的月平均值和方差进行信号保持调整，以提高它们与对应再分析预测因子场的分布相似性。

最后，由于深度学习算法对特征尺度差异较为敏感，因此需对 ERA-Interim 和 EC-Earth 预测因子进行缩放。缩放方法有多种，Doury 等，2022 年采用每日空间平均值和标准差对预测因子进行缩放。为避免这种空间缩放造成时间信息丢失，他们将每个预测因子的每日空间平均值和标准差时间序列作为额外预测因子加入。基于以往研究（Ba˜no-Medina 等人，2020、2021、2022），我们采用深度学习中更常见的预处理缩放流程，

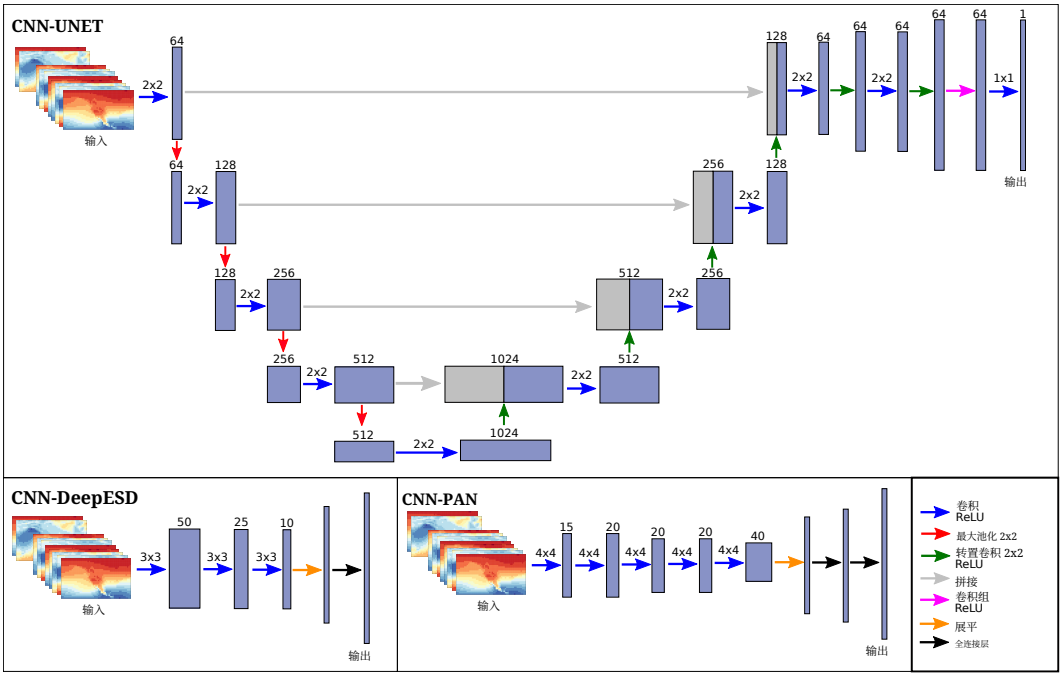

图 2. 本研究中相互比较的不同卷积神经网络模型的拓扑结构概览：CNN-UNET、CNN-DeepESD 和 CNN-PAN。

对 ERA-Interim 和 EC-Earth 预测因子在网格框层级进行缩放，并使用同一组均值和方差参数，因此不会丢失时间分量。这些参数根据 ERA-Interim 数据的训练期计算得出。

2.2 机器学习降尺度模型

本节介绍本研究采用的三种卷积神经网络模型，它们代表了当前最先进的降尺度深度学习方法：CNN-UNET、CNN-DeepESD 和 CNN-PAN，模型复杂度随参数数量逐步增加。CNN-UNET 是一种全卷积模型，由多个卷积层组成，可从预测因子空间生成空间一致的潜在特征图（或滤波器图），并经过优化以解释目标变量。该模型最初为图像识别任务提出，随后作为开箱即用的模型被应用于气候降尺度。CNN-DeepESD 和 CNN-PAN 则将卷积组件（作用于预测因子输入）与后续的稠密组件结合起来，后者作用于卷积生成的特征图，以解释目标变量；其中，CNN-DeepESD 中的稠密组件为线性，而 CNN-PAN 中的稠密组件具有高度非线性。这两种模型专为气候应用（包括统计降尺度）设计。图 2 概述了这三种模型，并展示了它们不同的拓扑结构（全卷积结构，以及卷积加稠密结构）。

所有卷积神经网络均采用相同的流程进行训练，分别使用 ERA-Interim 和 EWEMBI 作为预测因子场和预测目标场。标准化后的变量作为通道堆叠，并输入 CNN 模型的输入层。使用 ADAM优化器（Kingma & Ba，2014），以 1e-4 的学习率和 100 的批次大小，最小化均方误差（MSE）。我们遵循以往文献（Ba˜no-Medina 等，2020）的做法，将观测数据划分为训练集和测试集。为防止

模型过拟合，我们采用早停策略，并从训练数据中随机划分出 10% 用于验证。

2.2.1 CNN-UNET

UNET 模型是一种广受欢迎的现成卷积神经网络配置，已广泛应用于图像识别问题（Ronneberger 等人，2015）。在气候降尺度领域，UNET 已用于欧洲的多项研究（Doury 等，2022年；Quesada-Chac´on 等，2022年）。该模型由两个不同的模块组成：编码器和解码器（见图 2）。编码器通过卷积和最大池化压缩输入的维度，而解码器则通过转置卷积将其扩展至特定维度。网络中有一条路径连接编码器和解码器的不同层级，使它们保持连通（更多细节见 Ronneberger 等人，2015）。

我们构建了一个 UNET 模型，其编码器由五个模块组成，每个模块均包含卷积层、批归一化层和最大池化层——最后一个模块除外，该模块不含最大池化层；这些模块分别使用 64、128、256、512 和 1024 个卷积核。解码器则由依次排列的四个模块构成，每个模块包含转置卷积和标准卷积，卷积核数量分别为 512、256、128 和 64（详见图 2）。所有这些层均使用修正线性单元（ReLU）激活函数。解码器之后，依次应用标准卷积层和转置卷积层，以获得所需的输出维度。由于不含全连接层，该模型为全卷积模型。

2.2.2 CNN-DeepESD

CNN-DeepESD 模型由 Ba˜no-Medina 等人（2020）提出，用于对欧洲区域的温度和降水进行降尺度，并在多项研究中针对欧洲的气候变化条件进行了测试（Ba˜no-Medina 等人，2021年、2022年）。CNN-DeepESD 由三个卷积层组成，分别使用 50、25 和 10 个卷积核，隐藏层采用 ReLU 激活函数（详见图 2）。最后一个卷积层的输出会被展平并传入一个采用线性激活函数的全连接层，该层的 10870 个神经元对应预报变量区域中的网格点。请注意，这是结合卷积组件和稠密组件的最简单模型（本例中为单个线性层）。

2.2.3 CNN-PAN

Pan 等人（2019）提出了一种结合卷积层和全连接层的 CNN-PAN 模型，用于对代表北美洲气候态的 14 个不同网格点的日降水进行降尺度。在本研究中，我们采用作者提出的默认网络架构，并根据我们的任务进行了调整。由此得到的模型与 CNN-DeepESD 类似，但更深（见图 2），由五个卷积层组成（分别使用 15、20、20、20 和 40 个卷积核，并采用 ReLU 激活函数），后接两个全连接层（分别包含 5435 和 10870 个隐藏神经元，后者对应预报变量网格点）。按照 Pan 等人（2019）的建议，我们尝试了最大池化、随机失活和批归一化层，但均未取得理想效果。

2.3 可解释人工智能（可解释人工智能）

人们设计了不同的可解释人工智能技术，用于解释深度学习模型的运行机制，使其内部结构能够以人类易于理解的方式呈现。尤其是，显著性图通过量化输入空间特征对特定预测的影响，实现了可解释性。在降尺度领域，显著性图有助于识别相关的预测变量及其影响空间区域，从而便于对降尺度结果进行诊断和解释。这发挥着与.....类似的作用

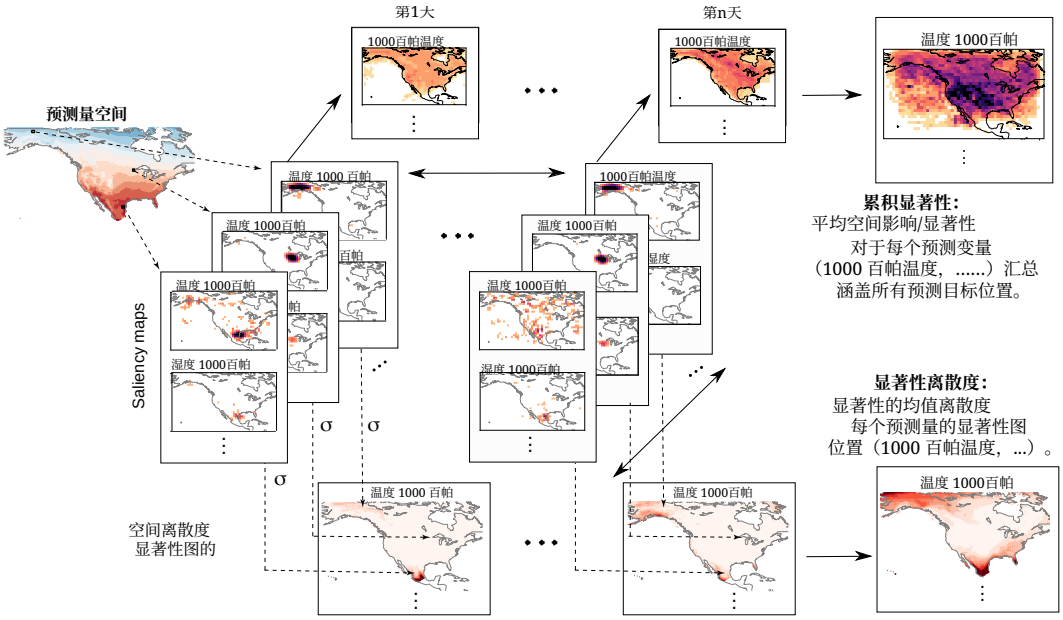

图 3. 本文所用基于可解释人工智能的显著性指标示意图：累积显著性图（ASM）和显著性离散图（SDM）。针对每个特定输出位置（一个预测量网格点）——图中以北部、中部和南部三个不同位置为例——以及每次观测（日）——图中按列展示了两天的数据——计算显著性图。请注意，ASM 场和 SDM 场的维度不同，因为前者位于输入（预测因子）空间，后者则位于输出（预测量）空间。

a.....标准机器学习技术中使用的特征重要性算法
p先前的统计降尺度研究（He 等人，2016年；Oh 等人，2022年）。

近年来，研究者在降尺度领域探索了多种基于显著性图的可解释人工智能技术（Gonz´alez-Abad, Ba˜no-Medina, & Guti´errez, 2022），发现各类技术得出的结果相似，尤其是那些依赖梯度计算的技术。基于此，我们采用一种称为积分梯度（IG）的基于梯度的技术（Sundararajan 等人，2017年），因为该技术能够克服标准梯度方法固有的问题，例如梯度饱和（Glorot & Bengio，2010年），并已应用于其他气候 CNN 应用（Kondylatos 等人，2022年）。

显著性图通过在卷积神经网络各层中反向传播梯度，从输出层回传至输入特征来计算。如图 3 所示，针对输出层中的每个神经元（对应某个预测量位置/网格点），以及每次观测（本例中为每天）分别计算显著性图。结果是在预测因子空间（网格框经度/纬度—变量）上定义的三维显著性图，其中每个值表示特定特征对模型预测的贡献。正/负显著性值表示有助于提高/降低目标位置预期温度的输入。由于我们只关注各输入特征的相关性，而不考虑其正负号，因此取绝对值。遵循近期研究（Toms 等人，2021年；Mamalakis 等人，2022年），我们按逐日方式对显著性图进行归一化（针对每个预测量位置分别除以最大值），从而着重识别信息丰富的输入变量和空间区域，并便于对不同模型进行比较。此外，为减少结果中的噪声并避免梯度碎裂问题（Mamalakis 等人，2022年），我们遵循 Toms 等人（2021年）的做法，移除低于 0.1 的相关性值。

近期研究中，作者会针对特定时期计算并研究显著性图，以揭示模型预测特定（极端）事件时的行为（Rampal 等，2022）。然而，与标准机器学习算法中使用的特征重要性技术类似，本研究关注的是模型在整个（训练/测试/降尺度）气候时期内的总体行为。因此，我们提出以下显著性幅度：

- **累积显著性图 (ASM)**：通过累积与所有预测目标位置对应的显著性图，获得每个特定网格框中每个预测因子值的总体空间影响/显著性（例如 1000 百帕温度，如图 3 顶部所示）。随后，在特定时间段内对其进行汇总。该指标反映预测因子空间中不同元素对所有预测目标位置进行预测的总体重要性。也就是说，给定预测因子网格框的 ASM 表明该输入对预测所有目标位置的总体重要性。数值较高表示局部影响强和/或影响范围大。

- **显著性离散图 (SDM)**：显著性离散图通过在（预测目标）地图的每个位置表示相应显著性图的离散度来计算；该值衡量给定目标位置的显著性图的空间稀疏性，因此离散度越大，预测因子的影响越分散（例如 1000 百帕温度，如图 3 底部所示）。与前一种情况一样，显著性离散图会在特定时间段内进行汇总。具体而言，显著性图的离散度可以通过多种方式计算（熵、标准差等）；在本研究中，我们使用以到目标位置的地理距离为权重的加权距离。然后将这些加权值相加，得到每个位置的一个实数值。为考虑纬度轴上面积的差异，我们采用哈弗辛公式（Robusto，1957）计算距离数组。

3 结果

本节展示本研究在不同条件下对比的三个 CNN 模型所得结果。我们首先介绍理想条件下（即使用再分析预测因子）进行交叉验证所得的结果；随后，测试这些模型外推未来气候条件的表现（使用全球气候模式预测因子），并表明在某些区域，卷积神经网络模型之间会得出不同结果。最后，我们介绍本研究提出的新型可解释人工智能指标，这些指标有助于部分解释上述结果。三个 CNN 模型均使用 ERA-Interim 预测因子和 EWEMBI 预测量进行训练，具体说明见第 2.1 节。

3.1 标准评估：交叉验证与外推

图4展示了在ERA-Interim测试期计算的四项标准验证评分（P02、均值和P98的偏差，以及均方根误差）的交叉验证结果，用于评估模型的整体性能。三种卷积神经网络模型按复杂度（根据所涉及的参数数量）递增排列在各行中，从全卷积的CNN-UNET到卷积与全连接的CNN-DeepESD和CNN-PAN模型。对于每项验证指标，图中均展示了空间结果以及绝对值的空间均值。图中显示，各模型的偏差具有相似的空间结构（CNN-DeepESD和CNN-PAN尤为如此），且极端值（P02和P98）的偏差较大。就均方根误差而言，CNN-UNET的表现最佳，尤其是在中东部和南部地区。然而，没有任何一个模型能在所有评分上都胜过其他模型，并在所有地区表现最佳；总体而言，没有理由舍弃其中任何一个模型。

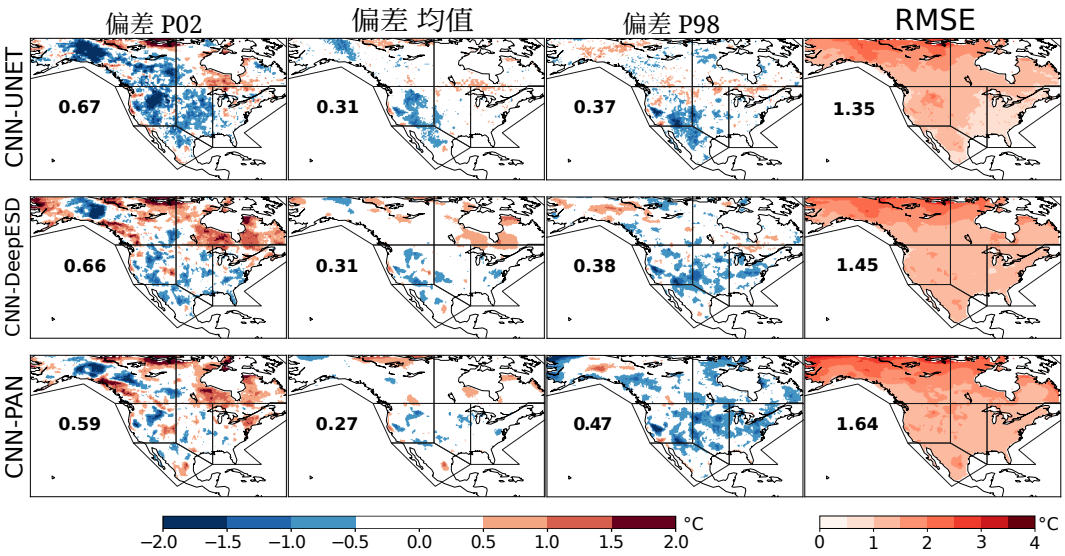

图4. 不同 CNN 模型在测试集上的验证指标。对于每个模型（按行排列），图中展示了 P02、均值和 P98 的偏差，以及均方根误差（按列排列）。每个面板中的数字表示绝对值的空间均值。图中还标出了政府间气候变化专门委员会区域。

为了检验卷积神经网络模型在未来气候变化条件下的外推能力（即将其应用于全球气候模式预测中的预测因子；见第2.1节），我们沿用先前研究的方法，将全球气候模式的“原始”预测作为伪现实（Vrac等，2007；Ba˜no-Medina等，2022年）。我们将未来情景划分为三个时期（2006—2040年、2041—2070年和2071—2100年），并计算全球气候模式和卷积神经网络模型在未来情景与历史时期（1980—2005年）之间的变化量。虽然降尺度后的预测可能不同于全球气候模式信号（例如，在局地尺度上增加附加价值），但原始和降尺度后的气候变化信号在次大陆尺度上总体应保持一致（例如，在本文使用的政府间气候变化专门委员会区域尺度上）。图5比较了不同政府间气候变化专门委员会区域中各模型的变化量（气候变化信号）。变化量按不同未来时段（行）计算，涵盖P02、均值和P98指数（列）。我们还展示了最后一个时段（2071—2100年）夏季（八月）和冬季（十二月）条件下的变化量。

在近期未来时期（2006—2040年），三个 CNN 模型给出的信号似乎总体上与全球气候模式的信号一致，但暖极端（P98）除外：在南部地区（NCA），CNN-DeepESD 和 CNN-PAN 的结果与全球气候模式及 CNN-UNET 略有差异。在中期和远期未来时期（2041—2070年和 2071—2100年），CNN-DeepESD 和 CNN-PAN 在暖极端（P98）和均值方面的差异逐渐增大，并在中部和南部地区（WNA、CNA、ENA 和 NCA）产生不合理的结果，部分情况下差异超过 2 度。不过，在不同未来时期，所有 CNN 模型对冷极端（P02）的结果均与全球气候模式一致。图5最后两列展示了冷季（十二月）和暖季（八月）的结果，进一步凸显了 CNN-DeepESD 和 CNN-PAN 方法在冷暖季节表现上的差异。总体而言，CNN-UNET 模型得出了合理的结果，在所有未来时期均呈现出与全球气候模式相似的区域格局。

CNN-UNET（与全球气候模式相比，结果看似合理）与 CNN-DeepESD 和 CNN-PAN 方法之间的差异（这些方法的气候.....

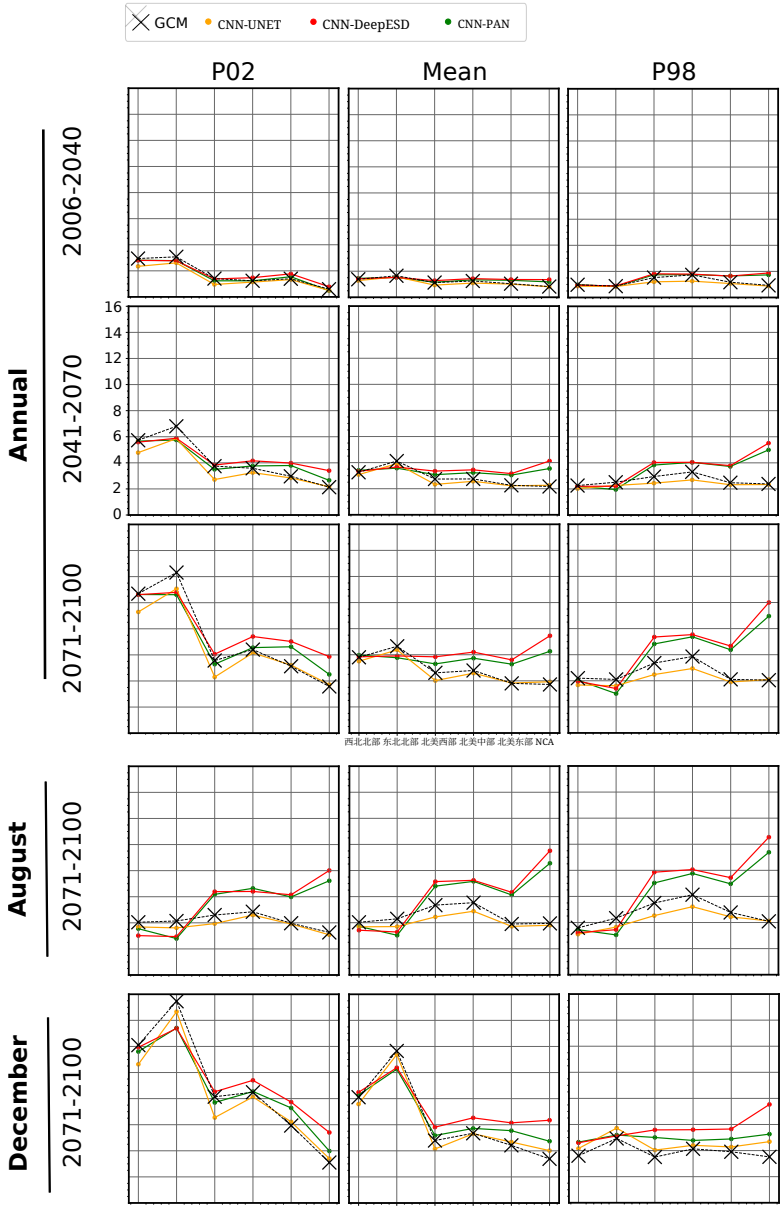

图5。 不同未来时期相对于历史时期（1980—2005年）的全球气候模式和卷积神经网络模型变化量，按不同政府间气候变化专门委员会区域（NWN、NEN、WNA、北美中部、北美东部和NCA；见图1）汇总，并按从左上到右下的顺序排列。前三行显示2006—2040年、2041—2070年和2071—2100年的年度信号。最后两行显示后一时期八月和十二月的季节信号。对于上述各时期，变化量分别根据P02、均值和P98计算（各列）。

在中部和南部地区的气候变化信号）并未明显体现在图4所示的标准评估指标结果中，所有模型在测试集上的表现都相对较好。这些结果凸显了潜在的模型缺陷，而由于底层模型的复杂性和黑箱特性，这些缺陷难以诊断。下文将看到，第2.3节介绍的可解释人工智能（XAI）技术能够解释这些差异。

3.2 运用可解释人工智能技术提升可解释性

第2.3节介绍的可解释人工智能诊断方法提供了相关信息，有助于理解第3.1节评估的卷积神经网络方法的表现。具体而言，累积显著性图（ASM）揭示了模型所使用的相关预测因子及其影响区域。图6展示了测试期内八月和十二月的累积显著性图汇总结果。对于每个CNN模型（按列排列），该图展示了1000百帕层五个大尺度预测因子对应的累积显著性图（按行排列）；其他气压层（850、700和500百帕）的结果显著性较低，为简洁起见未予展示。

对于 CNN-UNET 和 CNN-DeepESD 模型，温度是累积相关性最高的预测因子，不过在十二月的 NWN区域，比湿也会影响预测。CNN-UNET 显示的数值较低，表明与 CNN-DeepESD 相比，该模型针对不同目标位置的预测因子影响区域更为局部化；而 CNN-DeepESD 则在太平洋（八月）上也呈现出一些较弱的影响区域。

对于 CNN-PAN 模型，气温也是相关性最高的变量，但显著性值也分布于其他所有预测因子。我们可以看到，显著性模式更加嘈杂，整个北美洲的累计显著性更高，其中还包括一些大面积海洋区域。这意味着该模型会将相关性赋予目标位置局部影响区域之外的预测因子区域。这可能是该模型中稠密隐藏层的作用结果；这些层作用于卷积，并捕捉到比其他模型更广泛的影响区域。该模型在测试中取得的结果略差，这可能表明最终层中增加的复杂性导致了轻微过拟合。

请注意，ASM 总体刻画了整个区域内预测因子的重要性，并可根据（空间上）累积的显著性强度，比较不同模型中预测因子影响区域的空间范围。显著性离散图（SDM）提供了补充信息，用于判断这些预测因子对每个局地目标预测变量的影响是局限于局部还是扩展至更大空间范围。由此可以识别目标预测变量所在区域中预测因子的相关空间结构可能不符合物理规律的区域，从而指出模型中可能存在的问题。

图7展示了1000百帕温度的 SDM 结果（根据先前结果，这是影响力最大的变量）。图中按对比模型排列（列），并分别展示八月和十二月（行）。对于某个特定网格点，该指标的值较低，表明相应目标预测变量的显著性图基于集中在该区域的预测因子；值较高则表示空间范围更广（请注意，数值非常高可能意味着存在不符合物理规律的情况）。对于 CNN-UNET，我们可以观察到该指标在整个北美地区（八月和十二月均如此）都接近于零，这证实了该模型具有很强的局地性（如图6所示）。对于 CNN-DeepESD，我们观察到北部和南部地区的值较高，尤其是八月的南部地区。这可能表明模型对该区域的预测依赖于不符合物理规律的广泛影响区域，也可能解释气候变化的巨大偏差

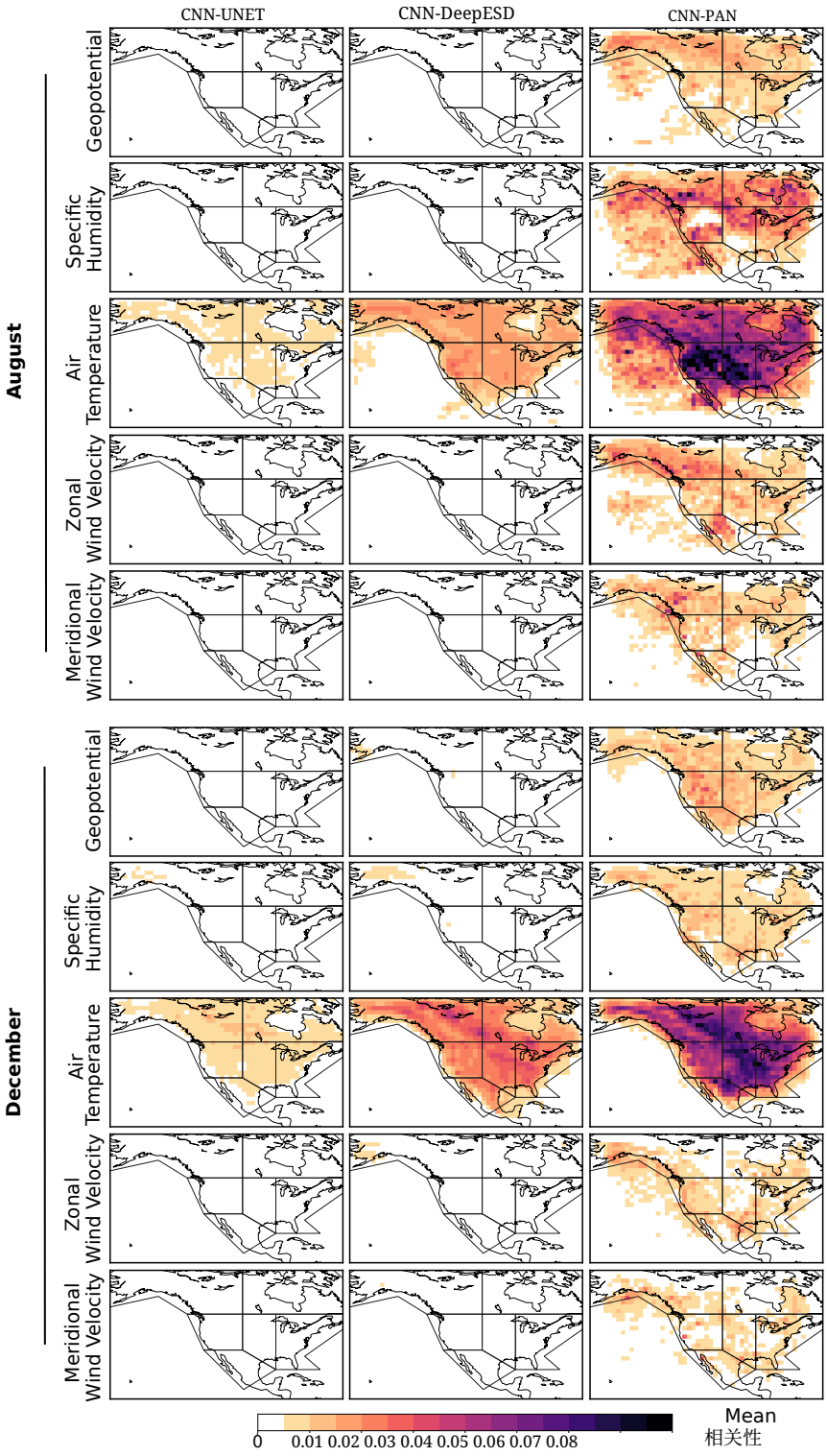

图6。八月和十二月各CNN模型（按列排列）针对五个大尺度预测因子的累积显著性图（ASM）（单个1000百帕气压层上的位势、比湿、温度以及纬向和经向风速）。

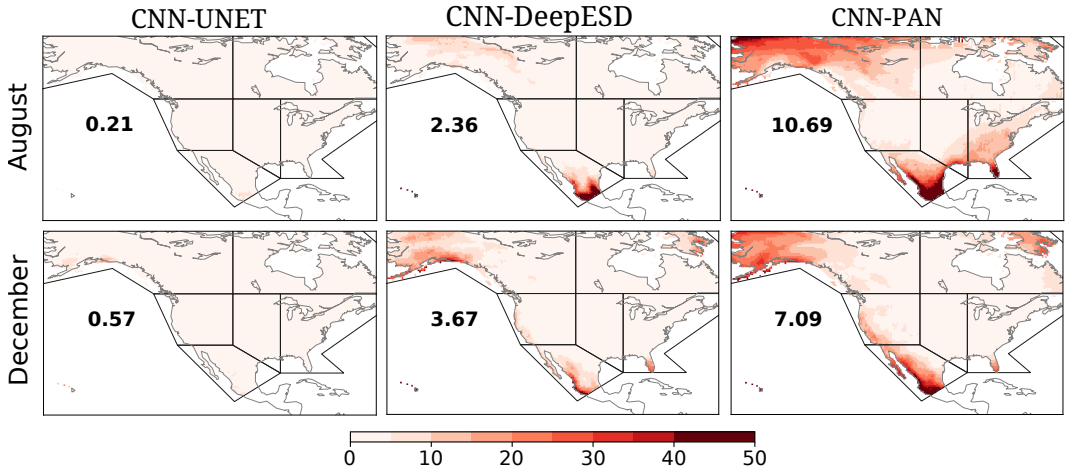

图7。 八月和十二月（按行排列）各模型（按列排列）在1000百帕处的温度显著性离散图（SDM）。

该区域暖季的信号（见图5）。对于 CNN-PAN 模型，这一现象更加显著，且空间范围更广。

为更好地理解这一潜在问题，图8按列展示了不同卷积神经网络模型的显著性图，按行展示了三个具有代表性的预测点位置，分别位于不同纬度（北部、中部和南部）。为简化起见，我们绘制了1000百帕处的气温和比湿。前者是影响最大的变量，后者则用于说明其他预测因子的表现。对于北部点（温度和湿度分别对应第一行和第四行），可以看到，所有模型都学到了相似的局地模式。其中，CNN-UNET的局地影响范围更窄。此外，对于CNN-DeepESD和CNN-PAN，可以看到局地范围之外也存在相关网格点，CNN-PAN尤其如此。图7所示的SDM指标反映了这种非局地性：这两个模型在NWN区域的取值较高，尤其是CNN-PAN。对于中部点（第二行和第五行），所有模型都显示出局地影响范围，其中CNN-UNET的模式比CNN-DeepESD和CNN-PAN更窄。最后，南部点在整个区域呈现出杂乱的显著性值，这表明CNN-DeepESD和CNN-PAN模型存在某种结构性问题，CNN-PAN尤为明显。这解释了为何这些模型在该区域缺乏外推能力，并对气候变化信号给出不合理的结果；另一方面，CNN-UNET在这三个点上保持了相同的局地模式，表明学习过程中不存在冲突。

4 结论

近期，已有多项研究表明，深度学习方法（尤其是卷积神经网络，CNN）在降尺度应用中前景广阔。这些方法已采用标准验证方法进行评估，但由于缺乏可解释性，它们在气候变化应用中的可信度仍然不足。本研究对北美洲地区的多个CNN模型进行比较，并引入多种可解释人工智能技术，分析模型的内部行为，为这些方法的评估提供依据，从而揭示结构缺陷可能导致的问题。首先，我们使用再分析预测因子，并通过交叉验证对CNN-UNET、CNN-DeepESD和CNN-PAN进行评估。基于多项广泛用于评估的指标

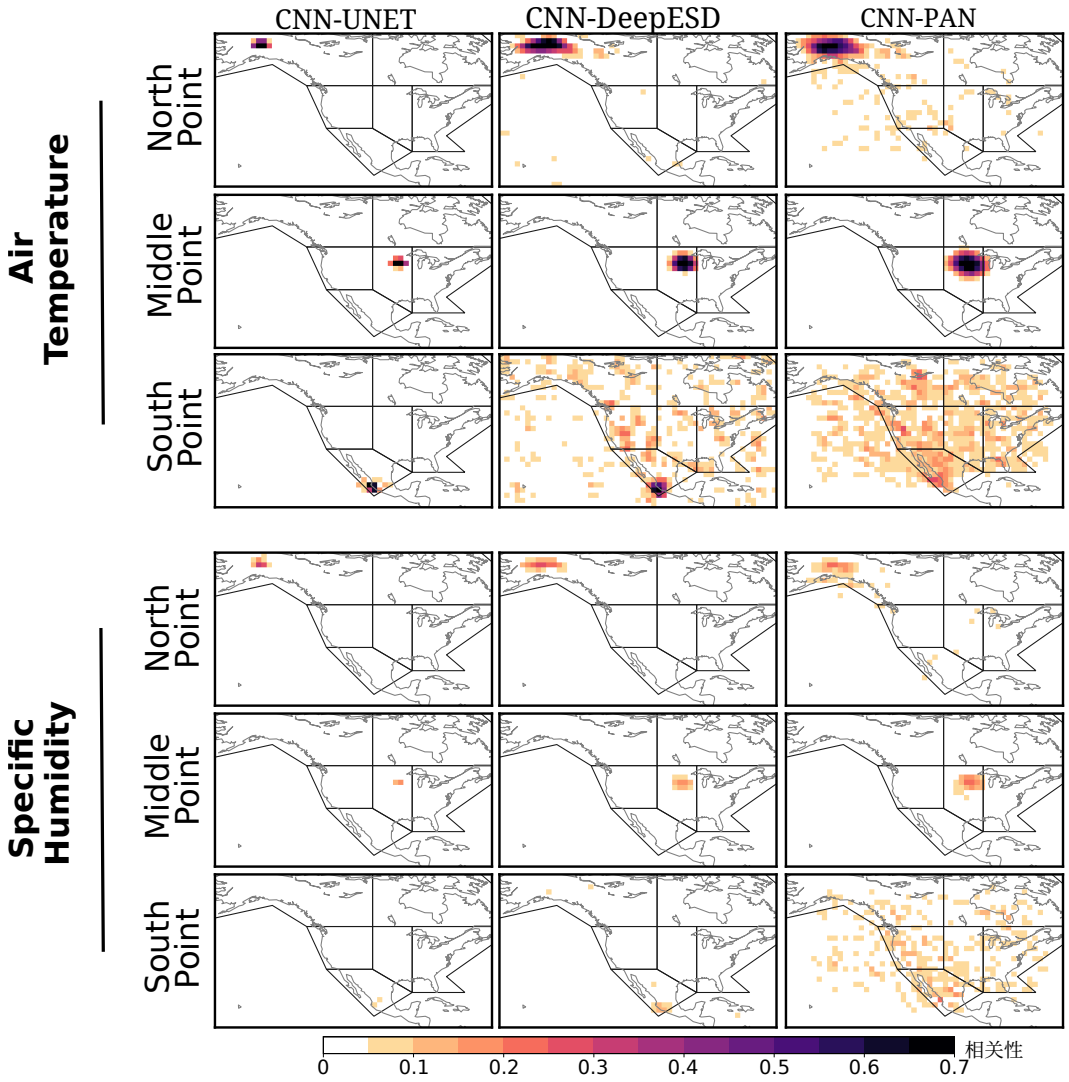

图8. 测试期八月（上图）和十二月（下图）的 1000 百帕气温和比湿显著性图。

在这些模型的性能方面，我们没有观察到明显差异。将这些模型置于气候模型空间（全球气候模式预测因子）中进行评估时，我们发现部分卷积神经网络模型（尤其是 CNN-DeepESD 和 CNN-PAN）的投影存在差异：与作为伪现实的 GCM结果相比，它们预测南部地区（尤其是 NCA）的温度更高。

为深入了解这些差异，我们分析了卷积神经网络模型在区间内的行为，并提出和应用了几种基于可解释人工智能技术的指标。具体而言，我们使用显著性图评估不同预测因子对模型结果的影响。我们发现，对于这三个模型，1000百帕处的气温是最重要的变量，这与先前的研究一致（Maraun & Widmann，2018年）。然而，CNN-DeepESD 和 CNN-PAN 模型未能正确学习南部地区（尤其是 NCA）预测因子与预测量之间的局地关系，导致未来情景下的降尺度气温出现偏差。相比之下，CNN-UNET 能够不受区域影响地学习正确的局地关系，表明其对局地性概念的学习更加稳健。这些差异源于卷积神经网络不同的拓扑结构。CNN-UNET 是一种全卷积模型，内含局地性的先验假设，这有助于在预测因子空间中学习高度局地化的降尺度模式。然而，CNN-DeepESD 和 CNN-PAN 的拓扑结构分别在网络末端包含一个和两个全连接层。这有利于学习更广泛的可能函数和影响区域，但也会引入噪声和虚假关系，这一点在本研究考察的可解释人工智能指标中有所体现。对于 CNN-PAN，这一问题尤为突出，因为两个全连接层使其更难学习局地性概念。

CNN-DeepESD 和 CNN-PAN 学到的虚假非局地关系，在八月等较暖月份造成的负面影响比在十二月等较冷月份更为严重。这与未来时期观测到的偏差相符：高百分位数和较暖月份的偏差更为显著。北美洲极端的纬度梯度加剧了 CNN-DeepESD 和 CNN-PAN 在南部地区的非局地性问题。变率较低（如 NCA）且属于热带气候的南部地区，其预测误差对总体误差的贡献较小。因此，在这些地区通过全连接层学习虚假关系不会受到较大的误差惩罚；由此，这些低变率地区在学习过程中得到的优化也较少。

本研究结果表明，可解释人工智能技术的应用能够扩展标准评估技术工具箱，用于评估基于深度学习的降尺度方法。此外，这些技术有助于我们理解模型的内部行为，并识别偏差和模型缺陷。当研究范围较大时，气候变率可能会导致所学模型产生偏差，因此这些技术尤为重要。我们鼓励研究人员将此类技术纳入评估框架，特别是在研究大范围区域或气候变化条件下；在这些情形中，充分理解卷积神经网络模型的行为有助于避免结果中的偏差和伪影。

5 数据和代码可用性声明

重现本研究实验所需的所有数据均可公开获取。ERA-Interim 数据可从欧洲中期天气预报中心网站（https://www .ecmwf.int）下载，EC-Earth 数据可从地球系统网格联合会（ESGF）门户网站（https://esgf-data.dkrz.de）下载。EWEMBI 数据集可从 ISIMIP 获取（https:// www.isimip.org）。

复现这些实验所需的代码已公开发布在 GitHub（https://github.com/jgonzalezab/XAI-metrics-North-America）。如需下载和处理

数据并训练模型，我们使用*climate*4R框架（Iturbide 等，2019）。具体而言，训练卷积神经网络模型时，我们使用*downscaleR.keras* （https://github.com/SantanderMetGroup/downscaleR.keras），它是*climate*4R的一部分，使用 TensorFlow（Abadi 等，2015）和 Keras（Chollet 等，2015）。用于计算可解释人工智能指标的积分梯度技术通过 Python 开源库*innvestigate*实现（Alber 等，2019）。所有这些实验均在西班牙国家研究委员会（CSIC）位于坎塔布里亚物理研究所 IFCA（CSIC - 坎塔布里亚大学）的深度学习计算基础设施上执行。我们在配备图形处理器（GPU）的节点上训练模型，具体使用 NVIDIA Tesla V100 图形处理器。在这些使用单个 GPU 的节点上，CNN-DeepESD 的训练每个周期约需 2 秒，CNN-PAN 需 3 秒，CNN-UNET 需 50 秒。为便于复现这些实验，并简化其在高性能计算集群上的运行，我们提供了按（Gonz´alez-Abad, L´opez Garc´ıa, & Kozlov, 2022）所述工作流程执行所需的脚本。GitHub 代码库中还提供了 Dockerfile。

致谢

我们感谢由 MCIN/AEI 资助的 ATLAS 项目（PID2019-111481RB-I00）提供的部分经费支持，资助编号为 10.13039/501100011033。J. Gonz´alez-Abad 感谢西班牙国家研究局通过卓越中心 Mar´ıa de Maeztu 单位提供的资助支持，项目编号为 MDM-2017-0765。此外，J. Ba˜no-Medina 感谢坎塔布里亚大学以及坎塔布里亚政府大学、平等、文化与体育部通过“用于探测宇宙本质的仪器与数据科学”项目提供的支持。

参考文献

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., . . . Zheng, X. (2015). *TensorFlow*：异构系统上的大规模机器学习。检索自 https://www.tensorflow.org/（软件可从 tensorflow.org 获取）Alber, M., Lapuschkin, S., Seegerer, P., H¨agele, M., Sch¨utt, K. T., Montavon, G., . . . Kindermans, P.-J. (2019). innvestigate 神经网络！《机器学习研究期刊》,*20*(93), 1–8. Ba˜no-Medina, J. (2020). 理解气候统计降尺度模型中的深度学习决策。载于第十届国际气候信息学会议论文集 (第79–85页)。Ba˜no-Medina, J., Manzanas, R., Cimadevilla, E., Fern´andez, J., Gonz´alez-Abad, J., Cofi˜no, A. S., & Guti´errez, J. M. (2022). 使用深度学习对多模式气候预测集合进行降尺度（deepesd）：对 CORDEX EUR-44 的贡献。《地球科学模型开发讨论》, 1–14. Ba˜no-Medina, J., Manzanas, R., & Guti´errez, J. M. (2020). 用于气候统计降尺度的深度学习神经网络模型配置与比较。《地球科学模型开发》, *13*(4), 2109–2124. Ba˜no-Medina, J., Manzanas, R., & Guti´errez, J. M. (2021). 深度卷积神经网络是否适用于大陆尺度气候变化预测降尺度。《气候动力学》, *57*(11), 2941–2951. Buhrmester, V., M¨unch, D., & Arens, M. (2021). 计算机视觉中黑箱深度神经网络解释方法分析：综述。《机器学习与知识提取》, *3*(4), 966–989. Chollet, F., et al. (2015). *Keras*。https://keras.io。Das, A., & Rad, P. (2020). 可解释人工智能（XAI）领域的机遇与挑战：综述。*arXiv* 预印本 *arXiv:2006.11371*。Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., . . . 等 (2011). ERA-Interim 再分析：资料同化系统的配置与性能

。Quarterly Journal of the Royal Meteorological Society, 137 (656), 553–597. Doblas Reyes, F., Acosta Navarro, J. C., Acosta Cobos, M. C., Bellprat, O., Bil- bao, R., Castrillo Melguizo, M., . . . 等 (2018). 使用 EC-Earth 开展气候预测研究。欧洲中期天气预报中心通讯(154), 35–40. Doury, A., Somot, S., Gadat, S., Ribes, A., & Corre, L. (2022). 基于深度学习的区域气候模型模拟器：一种新型混合降尺度方法的概念与初步评估。《气候动力学》, 1–29. Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). 第六阶段耦合模式比较计划（CMIP6）实验设计与组织概述。《地球科学模型开发》, 9 (5), 1937–1958. Glorot, X., & Bengio, Y. (2010). 理解深度前馈神经网络训练的困难。载于第十三届国际人工智能与统计学会议论文集 (第249–256页)。Gonz´alez-Abad, J., Ba˜no-Medina, J., & Guti´errez, J. M. (2022). 测试深度气候统计降尺度的可解释性技术。载于 ICLR 2022 地球与空间科学人工智能研讨会。Gonz´alez-Abad, J., L´opez Garc´ıa, ´A., & Kozlov, V. Y. (2022). 用于在高性能计算集群上分布式训练深度学习算法的基于容器的工作流。《集群计算》, 1–20. Goodfellow, I., Bengio, Y., & Courville, A. (2016). 深度学习。麻省理工学院出版社 (2016). He, X., Chaney, N. W., Schleiss, M., & Sheffield, J. (2016). 使用可调整的随机森林对降水进行空间降尺度。《水资源研究》, 52 (10), 8217–8237. Iturbide, M., Bedia, J., Herrera, S., Ba˜no-Medina, J., Fern´andez, J., Fr´ıas, M. D., . . . 等 (2019). 基于 R 的 climate4r 开放框架：用于可复现的气候数据访问与后处理。《环境建模与软件》, 111 , 42–54. Iturbide, M., Guti´errez, J. M., Alves, L. M., Bedia, J., Cerezo-Mota, R., Ci madev- illa, E., . . . 等 (2020). 更新政府间气候变化专门委员会气候参考区域，以便对气候模型数据进行次大陆尺度分析：定义与汇总数据集。《地球系统科学数据》, 12 (4), 2959–2970. Kingma, D. P., & Ba, J. (2014). Adam：一种随机优化方法。arXiv 预印本 arXiv:1412.6980。Kondylatos, S., Prapas, I., Ronco, M., Papo utsis, I., Camps-Valls, G., Piles, M., . . . Carvalhais, N. (2022). 基于深度学习的野火危险预测与理解。《地球物理研究快报》, e2022GL099368. Lange, S. (2019). EartH2Observe、WFDEI 和 ERA-Interim 数据：经合并并进行偏差订正后用于 ISIMIP（EWEMBI）。GFZ数据服务。LeCun, Y., Bengio, Y., et al. (1995). 用于图像、语音和时间序列的卷积网络。《脑理论与神经网络手册》, 3361 (10), 1995. Mamalakis, A., Barnes, E. A., & Ebert-Uphoff, I. (2022). 研究可解释人工智能方法在地球科学卷积神经网络应用中的忠实度。《地球系统人工智能》, 1 (4), e220012. Manzanas, R., Guti´errez, J., Fern´andez, J., Van Meijgaard, E., Calmanti, S., Mag- ari˜no, M., . . . Herrera, S. (2018). 欧洲季节气温预报的动力与统计降尺度：对用户应用的附加价值。《气候服务》, 9 , 44–56. Maraun, D., & Widmann, M. (2018). 气候研究中的统计降尺度与偏差订正。剑桥大学出版社, (2018). Maraun, D., Widmann, M., Guti´errez, J. M., Kotlarski, S., Chandler, R. E., Hertig,

E., . . . Wilcke, R. A.（2015）。Value：用于验证气候变化研究中降尺度方法的框架。地球的未来，*3*(1)，1–14。Oh, M., Lee, J., Kim, J.-Y., & Kim, H.-G.（2022）。利用多分辨率地形数据对风能资源图进行基于机器学习的统计降尺度。风能。Pan, B., Hsu, K., AghaKouchak, A., & Sorooshian, S.（2019）。使用卷积神经网络改进降水估算。《水资源研究》，*55*(3)，2301–2321。Quesada-Chac´on, D., Barfus, K., & Bernhofer, C.（2022）。通过深度学习实现可重复的高分辨率统计降尺度。地球科学模型发展，*15*(19)，7353–7370。Rampal, N., Gibson, P. B., Sood, A., Stuart, S., Fauchereau, N. C., Brandolino, C., . . . Meyers, T.（2022）。利用可解释的深度学习进行高分辨率降尺度：新西兰的极端降雨。天气与气候极端事件，100525。Robusto, C. C.（1957）。余弦-半正矢公式。《美国数学月刊》，*64*(1)，38–40。Ronneberger, O., Fischer, P., & Brox, T.（2015）。U-net：用于生物医学图像分割的卷积网络。载于国际医学图像计算与计算机辅助干预会议 （第234–241页）。Sun, L., & Lan, Y.（2021）。利用深度学习神经网络模型对中国日温度和降水进行统计降尺度：局地化及与其他方法的比较。国际气候学杂志，*41*(2)，1128–1147。Sundararajan, M., Taly, A., & Yan, Q.（2017）。深度网络的公理归因。载于国际机器学习会议 （第3319–3328页）。Toms, B. A., Barnes, E. A., & Hurrell, J. W.（2021）。利用可解释神经网络评估地球系统模型的年代际可预测性。《地球物理研究快报》，*48*(12)，e2021GL093842。Vrac, M., Stein, M., Hayhoe, K., & Liang, X.-Z.（2007）。一种在未来气候变化背景下验证统计降尺度方法的通用方法。《地球物理研究快报》，*34*(18)。Zhang, Y., Tiˇno, P., Leonardis, A., & Tang, K.（2021）。神经网络可解释性综述。*IEEE*计算智能新兴主题汇刊。