arXiv:2611.31064v1 [physics.ao-ph] 23 Nov 2026

人工智能与

动力气候集合中的不确定性比较

Jacob B. Landsberg[1*†]、DevinM.McAfee[1†]、M.A.Fernandez[1†]、ElizabethA. Barnes[1,2]

[1]波士顿大学计算与数据科学学院，美国马萨诸塞州波士顿市联邦大道665号，邮编02215。[2]波士顿大学地球与环境系，美国马萨诸塞州波士顿市联邦大道685号，邮编02215。

*通讯作者。电子邮箱：jlandsbe@bu.edu；共同作者：dmcafee@bu.edu；martmart@bu.edu；eabarnes@bu.edu；

†这些作者对本研究的贡献相同。

摘要

气候模式集合及其不确定性估计对于评估风险和制定气候政策至关重要。然而，在评估新开发的人工智能气候模型时，不确定性的评估往往次于对平均气候状态偏差和趋势的诊断。在这项研究中，我们首次系统比较了最先进的人工智能模型与动力大气气候模式集合的不确定性表征。通过考察2米气温、500百帕位势高度和可降水量，我们发现，与动力学集合相比，人工智能集合在各变量上的内部变率和模型不确定性均有所降低。我们表明，这种减弱的变率会使人工智能集合产生较少的气候极端事件，相较ERA5也呈现欠离散。在均匀变暖情景下，人工智能集合的模型不确定性有所增加。动力学模型也面临自身的挑战：相较ERA5，它们表现出更严重的校准失准（过度离散），且气候偏差大于人工智能模型。随着人工智能气候模型与动力气候模型相互融合并应用于业务气候预测，审慎考量两者在不确定性表征方面的差异，对于切实评估气候风险至关重要。

1 引言

可靠的气候不确定性量化对于评估气候风险对人类健康、能源系统、农业和基础设施造成的影响至关重要。对于月度至十年尺度的气候预测，不确定性主要源于内部变率和模型不确定性 [1]。内部变率指由气候的混沌特性所产生的自然变化。尽管其中包含可预测的成分 [2]，内部变率仍主导着近期气候不确定性。长期气候不确定性主要源于对可能的气候变化情景的不确定性，以及特定情景下受迫响应的模型不确定性 [1]。在较长时间尺度上，内部变率也会影响区域影响，尤其是在高纬度地区 [3]。内部变率可能会减弱或放大受迫响应，对评估近期气候风险至关重要；而由计算限制和物理认知不足造成的模型不确定性，则为长期气候规划提供依据。

以控制物理方程为基础的动力气候模型集合，长期以来一直是模拟气候不确定性的黄金标准 [4, 5]。气候集合由多个采用相同强迫、但受到不同独立扰动的模拟组成。在单模式气候集合中，模拟之间的模式内差异代表内部变率，而集合平均则近似于受迫响应 [1, 6]。

多模式气候集合由多个单模式集合组成，因此可以通过模式间模拟差异来评估模型不确定性 [4, 7]。近几十年来，动力集合一直是气候科学的基石，用于分离气候变率的作用机制，并量化人类对气候系统的影响 [8]。然而，动力气候集合受制于巨大的计算成本，而且随着模式分辨率和复杂度的提高，计算成本还在持续增长 [9]。这限制了动力集合的生成频率，也使得无法生成足够庞大的集合，来充分表征气候极端事件

。 [10]。

近期，一类新型人工智能模型应运而生，通常称为人工智能模拟器，有望以远低于动力学模型的计算成本用于气候预测。人工智能模拟器已达到与最先进的动力天气模型相媲美的预测能力 [11–13] ，近期其应用范围也已扩展至气候预测[14–20]。人工智能气候模拟器采用了多种不同的模型架构，通常旨在模拟气候再分析数据集；不过，开发模拟观测数据的模型的工作也正在兴起 [例如 21]。

尽管早期结果令人鼓舞 [例如 14, 15], ，但目前仍缺乏对人工智能模拟器在气候不确定性量化方面能力的系统分析。使用确定性损失函数训练的天气模拟器往往存在欠离散现象 [例如 11, 22]„ 但这种欠离散是否会影响气候模拟器的内部变率，以及概率损失函数能否纠正这一行为，仍有待证实。此外，随着气候模拟的应用范围不断扩大，多模式人工智能集合已成为可能，但其不确定性尚未得到全面评估。

人工智能模型比较计划（AIMIP）的第一阶段近期已完成，旨在满足对人工智能气候模拟器进行全面评估的需求[23]。AIMIP协作组遵循统一的协议，汇集了迄今为止规模最大的待评估气候模拟器集合。以往研究 [23, 24] 主要关注AIMIP中强迫变率的偏差。本文则首次对人工智能模型中的不确定性表征进行系统比较。具体而言，我们将AIMIP对内部变率和模型不确定性的表征，与一个由部分 CMIP6 AMIP（大气模式比较计划， [5]）模拟组成的多模型动力学集合进行比较。我们展示了这些表征中的偏差如何影响模拟的历史变率。此外，我们还说明了在理想化气候

变化情景下，AIMIP中的不确定性表征如何变化——这是其未来有望整合进研究和业务流程前的一项重要测试。

2 数据与方法

2.1 数据

AIMIP [23] 模拟遵循为 CMIP7（耦合模式比较计划第 7, [25] 阶段）制定的 AMIP 协议。外部强迫由月度ERA5 [26]数据提供，这些数据已为 AIMIP 预处理[1]。出于对过拟合的担忧，AIMIP模型不包含 CO_2 强迫 [23], ，因此通过统一的 $+2$ K 海表温度和 $+4$ K 海表温度扰动模拟来评估模式对未来气候的泛化能力。

AIMIP 包含八个模式：ArchesWeather 和 ArchesWeatherGen [16, 17], ACE2.1-ERA5 [14], DLESyM [18], NeuralGCM 和 NeuralGCM-HRD [15],cBottle[19], 以及 MD-1p5 [20]。我们排除 NeuralGCM，因为 NeuralGCM-HRD 代表同一模式，但其分辨率从 2.8 度降尺度至其他参与模式共用的 1 度网格（对于采用 HEALPix 网格的模式，则使用相应网格）。除非另有说明，主要分析中不包括全扩散模型（MD-1p5 和 cBottle）。全扩散模型会出现不理想的伪影，例如内部变率偏低（表 2）；此外，与非扩散模型相比，MD-1p5 的时间方差偏高。此外，非扩散型AIMIP模型有五个，与动力学集合的规模相同。

每个人工智能建模中心都针对训练期（1979—2014年，共36年）和测试期（2015—2024年，共10年），生成了一个由五个成员组成的月平均输出集合。不过，我们也纳入了年度结果，这些结果采用滚动的12个月平滑窗口计算，以分析不同

[1]https://zenodo.org/records/17065758

时间尺度（例如，图 1）。集合通过扰动初始条件生成，但对所用的扰动方法没有限制。有关 AIMIP 数据的更多详情、访问说明和示例处理代码，请参阅[2]。

表1 本研究中使用的 AIMIP模型。其中许多模式已根据已发表的设计进行了更新，以适应 AIMIP 协议，例如移除作为输入的二氧化碳，参见 [23]。

模式名称	备注
ArchesWeather [16]	Transformer架构、自回归、均方根误差损失。
ArchesWeatherGen [17]	流匹配模型预测 ArchesWeather 确定性输出。
ACE2.1-ERA5 [14]	球面傅里叶神经算子（SFNO）架构，自回 归，均方误差损失。
DLESyM [18]	U-net 卷积神经网络架构，自回归， 均方根误差损失，HEALPix 网格 [27]。
NeuralGCM-HRD [15]	混合模型（采用编码器-解码器架构的动力核心）， 自回归，连续秩概率得分损失。
cBottle [19]	扩散架构，非自回归，HEALPix 网格。集成—— 通过组合多个训练 检查点的预测生成。
MD1p5 [20]	条件变分自编码器和潜在扩散架构—— 自回归（月度），　　均方误差、KL 正则化和 扩散损失。

我们将人工智能模型与一组按照 CMIP6 [5] 的 AMIP协议运行的仅大气动力学模型进行比较。我们筛选出在1979—2014年期间，2米气温（TAS）、500百帕位势高度（Z500）和比湿这三个变量各至少包含五个成员的模式。虽然AIMIP模型的测试集延伸至2024年，但AMIP试验的数据期仅一致延伸至2014年，这限制了我们的比较时段。本文使用的动力学模型包括：CESM2（10个成员）、GISS-E2-1-G（20个成员）、IPSL-CM6A-LR（21个成员）、MIROC6（10个成员）和NorCPM1（10个成员）。我们采用保守重网格化，将这些动力学模型重网格化至与AIMIP模型相同的一度网格。下文将这一动力学集合称为DENS。

[2]https://github.com/ai2cm/AIMIP

在本研究中，我们评估TAS、Z500和低层自由对流层可降水量（PRW）。PRW通过对500百帕至850百帕之间的比湿进行积分计算；延伸至该气柱范围内的高海拔区域不纳入计算。我们选择考察整个大气柱中的这一部分，原因是：（1）并非所有模式都提供其他气压层的数据；（2）低层自由对流层中的水汽对深对流的发生具有重要的物理意义，尤其是在热带地区 [28]。DLESyM不包含比湿，因此不参与PRW分析。对于TAS，我们屏蔽海洋区域，因为TAS与模式所受强迫的指定海表温度紧密耦合。我们将比较重点放在1979—2014年，因为这是AIMIP与DENS的共同时间段。我们还分析了AIMIP测试时段（2015—2024年）以及统一的 $+2$ 开尔文SST模拟，以考察AIMIP的样本外行为。

2.2 方法

对于 AIMIP 和 DENS 中的每个模式，都会在每个网格点减去该模式的时间平均值，以滤除平均态偏差的影响。我们在每个月时间步计算方差，既计算每个模式内部成员之间的方差（大气内部变率），也计算各模式均值之间的方差（模型不确定性）。这种表述是气候集合中的标准做法，并假设内部变率和模型不确定性彼此独立，且可视为可加的方差来源 [例如， 7]。多模式集合的剩余方差则由大气受迫响应的时间方差（大气受迫变率）解释。将每个框架的方差分解为大气内部变率和受迫变率的贡献，同时考虑模型不确定性，可以通过比较各方差分量的差异，解释两个框架之间变率的差异。重要的是，尽管这种方差分解方法很常见 [1, 7], 但它会高估模型不确定性和大气受迫

在样本量较小时的变率 [29]。模式均值和多模式均值都会保留抽样噪声，从而放大模式间差异（模型不确定性）和时间方差（大气受迫变率）。能够校正这一问题的随机效应模型会给出较小的这两种方差估计值，但不会改变 AIMIP 与 DENS 之间的比较，相关讨论见补充材料第S9节。

在全文中，我们通常将大气内部变率简称为内部变率，将大气受迫变率简称为强迫变率。不要将这些概念与耦合气候集合中的内部成分和受迫成分混淆。具体而言，在耦合集合中，受迫响应衡量的是气候系统对辐射强迫的响应。本文所说的受迫响应，是指大气对海表温度强迫的响应。该受迫响应在月时间步上的变率主要由季节性导致；而在年际至年代际时间尺度上，大气受迫响应则受到海表温度内部变率（如厄尔尼诺—南方涛动）以及气候变化导致的局地海表温度趋势的影响。

我们通过对时间进行自助抽样、采用留一模型法，以及对 DENS 模式进行成员子抽样，评估内部变率和模型不确定性估计的置信度。在每次时间自助抽样中，通过从 1979—2014 年期间有放回地随机抽样，计算内部变率和模型不确定性。在计算模型不确定性时，每次时间自助抽样中，我们都会从每个 DENS模式（最多有 21 个成员）中随机抽取 5 个成员（重复 20 次，且不放回），以匹配 AIMIP模型的集合规模。这样做是必要的，以确保 DENS 的模型不确定性不会因受迫响应的不确定性较低而被人为压低 [例如 6]。我们对时间自助抽样维度（以及 DENS 模型不确定性中的成员维度）取平均值，迭代次数则根据自助抽样过程的收敛情况确定（见补充图 S3）。由于我们主要关注比较完整的 DENS

和 AIMIP 框架，而非单个模式，因此我们还采用留一模型法，为框架平均结果提供不确定性估计（例如，表 2）。

除了比较各框架的方差外，我们还评估其极端值和边际校准。为分析极端值，我们汇集每个模式的所有时次和成员，计算空间分位数——这些分位数由成员均值经验累积分布函数（CDF）定义。对分位数进行自助法处理的方式与模型不确定性的自助法相同，但不采用留一模型法，因为这对分位数分布的影响很小。为评估各框架的边际校准，我们利用 AIMIP 和 DENS 的多模式分布，计算 ERA5再分析的概率积分变换（PIT）。

对于网格点 i，我们汇集各时次、成员（对于 DENS，采用前五个成员）和模式，以构建多模式经验累积分布函数 F_i。PIT 值计算为 $F_i(\mathrm{ERA5}_{i,\,t})$在网格点 i处随时间变化的每个 ERA5 样本的值。我们注意到，PIT 值通常使用时间条件预报分布 $F_{i,\,t}(\mathrm{ERA5}_{i,\,t})$计算，用于检验天气模型的概率校准 [例如， 30]。在此，我们关注气候模型的边际校准，因此仔细考察气候累积分布函数 F_i。对于经过边际校准的集合，PIT 值应近似服从标准均匀分布。通过将 PIT 值的方差与标准均匀分布的方差进行比较，可以评估离散度；标准均匀分布的方差为 $1/12$。欠离散集合的 PIT 方差为 $> 1/12$，而过度离散集合的 PIT 方差为 $< 1/12$ [31]。我们在图 3 中将离散度量化为 $\log(12 \times \mathrm{Var(PIT)})$，即 PIT 方差与标准均匀分布方差之比的对数。该离散度指标在过度离散时为负，在欠离散时为正。

通过将 PIT 值划分为十分位区间，绘制 PIT 直方图，以考察各网格点的全球校准。对每个

网格点的 PIT 直方图进行纬度加权平均，得到图中报告的全球PIT直方图4。为衡量边际校准的偏差，我们报告 D 指标（参见 [32] 和第S6节）。

3 结果

表2报告了不同框架下TAS、PRW和Z500的月平均内部变率和模式不确定性，以及由自助法和留一模型迭代范围所定义的置信区间。我们发现，与非扩散型AIMIP模型相比，DENS在各变量中模拟出更强的内部变率，并表现出更大的模式不确定性（表2），但这种内部变率差异在PRW中未达到显著水平。此外，尽管并非总是显著，DENS在年时间尺度上也表现出类似的更大内部变率和模式不确定性。在年度平滑模拟中也观察到了这一现象（图1）。此外，对于两种扩散模型，这种差异进一步扩大；与所有非扩散模型相比，它们在各变量中的内部变率均显著较弱（表2，另见表S2）。

表2 DENS 和 AIMIP 模型的内部变率与模型不确定性。由于 AIMIP 扩散模型仅有两个，因此我们不报告扩散框架的模型不确定性。粗体数值表示其置信区间与训练期内 AIMIP 非扩散模式的置信区间不重叠。

	TAS (K^2)	PRW (mm^2)	Z500 (m^2)
Internal variability			
DENS	**2.84 [2.63, 3.09]**	2.01 [1.85, 2.16]	1294 [1226, 1347]
AIMIP non-diffusion	2.37 [2.21, 2.56]	1.96 [1.87, 2.04]	1167 [1125, 1227]
AIMIP non-diffusion (test)	2.29 [2.04, 2.59]	2.04 [1.87, 2.17]	1169 [1100, 1243]
AIMIP non-diffusion (+2K)	2.60 [2.16, 2.95]	**2.82 [2.37, 3.08]**	1274 [1140, 1366]
AIMIP diffusion	**0.906 [0.650, 1.18]**	**0.467 [0.227, 0.733]**	**439 [368, 510]**
Model uncertainty			
DENS	**2.18 [1.76, 2.53]**	**1.40 [1.17, 1.59]**	**428 [390, 468]**
AIMIP non-diffusion	0.731 [0.640, 0.822]	0.538 [0.487, 0.591]	296 [277, 312]
AIMIP non-diffusion (test)	0.688 [0.590, 0.791]	0.545 [0.468, 0.613]	298 [269, 326]
AIMIP non-diffusion (+2K)	1.42 [0.780, 1.70]	**0.916 [0.670, 1.12]**	**373 [319, 415]**

在测试期内，AIMIP的内部变率和模式不确定性基本保持不变。另一方面，在理想化的 +2 K情景下，模式不确定性明显增加。这或许是意料之中的，因为AIMIP模型对样本外强迫产生了一系列不符合物理规律的响应，例如全球范围内TAS敏感性减弱，以及无法捕捉陆地—海洋增暖对比。例如，ACE2.1-ERA5和cBottle表现出显著的样本外伪影，模拟出全球大部分陆地近地表降温响应 [23]。AIMIP的内部变率也有所增加，但仅PRW的增加达到显著水平。表2 中TAS和Z500内部变率的适度增加，是由单个模式ACE2.1-ERA5的不符合物理规律的响应所导致的，该模式的成员在不同变量之间出现分歧（补充表S2）。然而，PRW内部变率的增加具有显著性，并且在所有非扩散型AIMIP模型中均有观察到，而不仅限于ACE2.1-ERA5（见补充表S2）。在 +2K情景下，PRW模式不确定性的很大一部分增加同样归因于ACE2.1-ERA5。

在图1中，我们展示了 AIMIP 和 DENS 在月和年时间尺度上的强迫变率，以及时间平均的内部变率和模型不确定性。在月时间尺度上，各变量的方差主要由强迫变率主导（图 1a），这主要是季节性造成的；经过年度平滑后，强迫变率的贡献显著降低（图 1b）。对于 AIMIP，在月时间尺度上，模型不确定性对方差的贡献远小于内部变率。然而，对于 DENS，月度 TAS 和 PRW 的模型不确定性与内部变率的量级相当。模型不确定性的这一差异，可能是因为 AIMIP 模型均经过训练，以模拟 ERA5 的季节循环。因此，AIMIP 对其季节循环的模型不确定性相对较小，导致其月度模型不确定性低于 DENS。

在年时间尺度上，季节性被消除，因此总方差降低（图 1b）。对于 TAS 和 Z500，AIMIP 和 DENS 中剩余方差的最大来源都是内部变率。然而，年度 PRW 方差仍主要由强迫变率贡献。这可能是因为全球水文循环主要受湿润热带，尤其是热带太平洋支配 [例如， 33]；热带太平洋变率受厄尔尼诺—南方涛动调制，其峰值强度出现在年际时间尺度 [34]。总体而言，在图 1所示的两个时间尺度上，DENS 的强迫变率都强于 AIMIP。然而，与 TAS 和 Z500 不同，在年尺度上，DENS 和 AIMIP 的 PRW 强迫变率几乎相同。我们推测这也与厄尔尼诺—南方涛动有关。虽然 AIMIP 模型总体上对规定的全球海表温度敏感度较低，但其 PRW 响应更强，且主要局限于热带太平洋（见补充图 S4 和S7）。尽管在年时间尺度上，DENS 的模型不确定性仍相对较大，但与月度模拟相比，两种框架之间的不确定性差异较小。换言之，AIMIP 对季节循环的不确定性较低，而其对较长期强迫变率的不确定性则与动力学集合更为接近。

现在，我们说明两个框架中变率分量的月度差异如何体现在极端事件表征中。在图2 中，我们展示了DENS与AIMIP之间第1、第5、第10、第90、第95和第99百分位数（跨模式平均）的差异，即月度TAS的差异（PRW和Z500见补充图S4 和S5）。我们发现，DENS产生的极端值高于AIMIP，并且两个框架之间的差异在最极端的分位数处最大。例如，对于TAS，DENS中的TAS第99百分位数比AIMIP高出超过1开尔文，而第95百分位数则高出略低于1开尔文（图2b）。

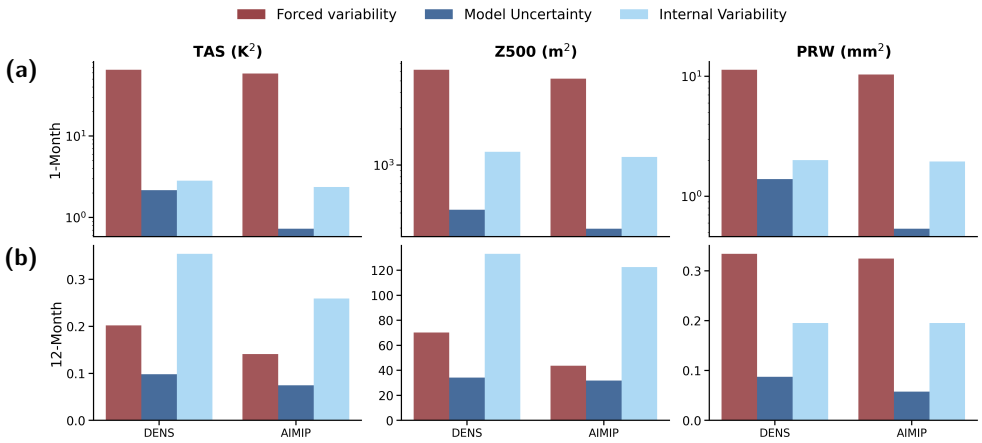

图1 AIMIP 和 DENS 在训练期内的内部变率、模型不确定性以及多模式平均（强迫变率）的时间变化。图中展示了（a）月度和（b）年度（经过 12 个月平滑处理）的结果；月度结果的 y 轴采用对数刻度，以便显示内部变率和模型不确定性的差异。

在月时间尺度上，这一差异主要受AIMIP较弱的季节性影响，因为对于两个框架而言，月度强迫变率均显著强于内部变率（图1）。例如，夏普利值分析表明，月度TAS $q = 0.99$ 差异中有86%归因于强迫变率的差异（见补充章节S10）。如图2a所示，在全球陆地的大部分地区，DENS的近地表温度极端值强于AIMIP。这一TAS差异在北半球温带地区最为明显，而在南半球相对较小，这与北半球更强的季节性气温变率一致 [35]。同样，AIMIP的大多数地区的Z500极端值较弱。然而，最大的Z500差异出现在温带地区，那里季节性变率最强（补充图S5）。

月度可降水量极端值的差异在空间上更加不均匀。在温带地区的大部分区域，AIMIP的离散程度较低，并且全球平均分位数较弱。然而，AIMIP主要在热带太平洋中部和东部产生更强的可降水量极端值（补充图S4），该区域的强迫变率由厄尔尼诺—南方涛动主导。例如，在NINO3.4区域 [36]， $q = 0.99$ 高出2.58毫米

AIMIP中的数值更高——纬度加权全球平均差异为 -0.37 毫米——其中2.13毫米归因于强迫变率的差异（补充图S14）。

在年时间尺度上，我们观察到类似的差异：随着分位数越来越极端，AIMIP 产生的极端值比 DENS 越来越弱（见补充图S6、S7和S8）。由于年方差相较于月度模拟有所减弱，年度分位数差异也相对较小。与季节性占主导地位的月时间尺度相比，内部变率对 TAS 极端值差异的贡献更大。例如，全球平均 $q = 0.99$ 年度 TAS 在 AIMIP 中低 0.27 K，其中 53% 的差异归因于内部变率（见补充图S16）。然而，对于 Z500 和 PRW，不同框架之间单模式强迫变率的差异仍起主导作用（见补充章节 S10）。

显然，AIMIP 多模式集合的离散程度低于 DENS。然而，尽管这一比较具有参考价值，但若不与观测或再分析数据关联，其局限性很大。因此，我们使用 PIT 分析，评估这两个框架相对于 ERA5 的边际校准。图3展示了月时间尺度上各框架不同变量的 PIT 离散度空间分布图。对于全球大多数地区的每个变量，DENS 都显著过度离散，PRW 在大部分热带海盆上除外。AIMIP 的局地校准优于 DENS，达到显著水平的区域少得多。然而，AIMIP 存在离散不足，尤其是在 TAS（图3a）和 Z500（图3e）中；而在极地 PRW 中则表现为过度离散，这或许是因为干旱极地地区的气候态 PRW 极低。训练期内，AIMIP PRW 的离散度在空间上并不均匀；而在测试期内，AIMIP 的 PRW 离散不足分布得更均匀，同时 TAS 和 Z500 仍然离散不足（见补充图 S9 ）。

全球 PIT 直方图见图4，D 指标见补充表S3。边际校准表现为均匀的 PIT 直方图，其中

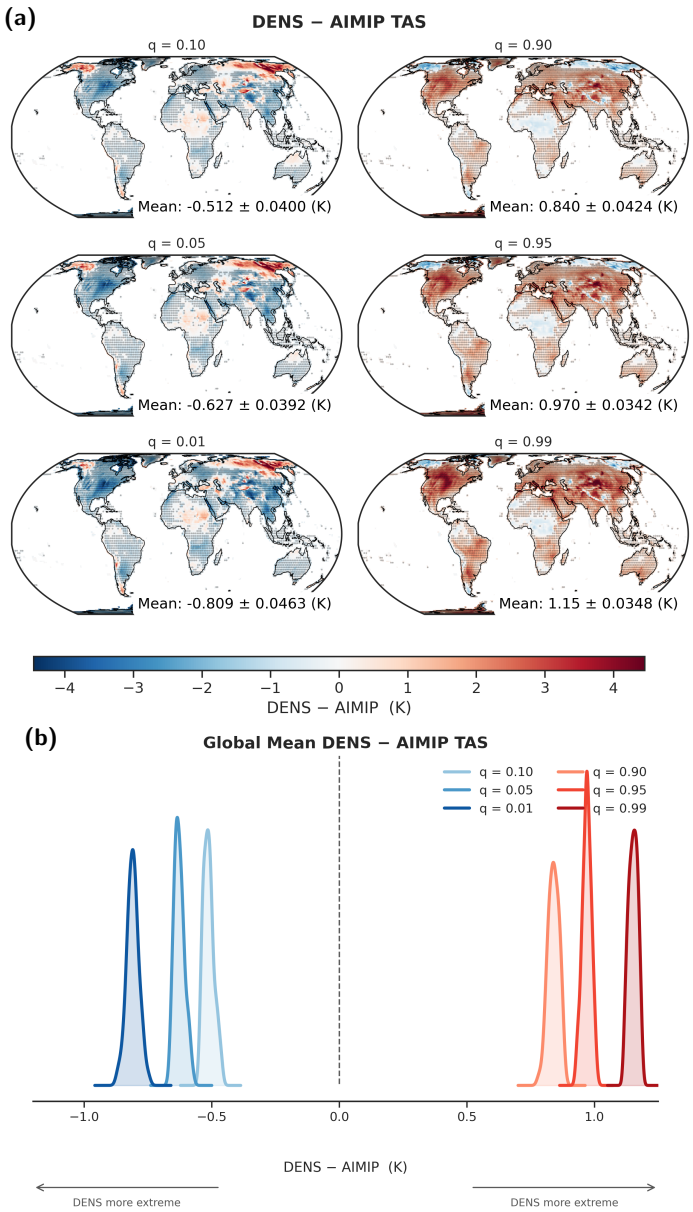

图 2 (a) DENS 与 AIMIP 在 TAS 的第 1、5、10、90、95 和 99 分位数上的全球平均差异。点状阴影表示，在所有自助法迭代中，DENS 的值都比 AIMIP 的值更极端。(b) 面板 (a) 中差异的全球平均值。分布和不确定性（2σ）根据自助法分布计算得出。

每个箱包含相同比例的 ERA5 样本。对于离散度不足的集合，ERA5 样本将不成比例地集中在集合的极端尾部

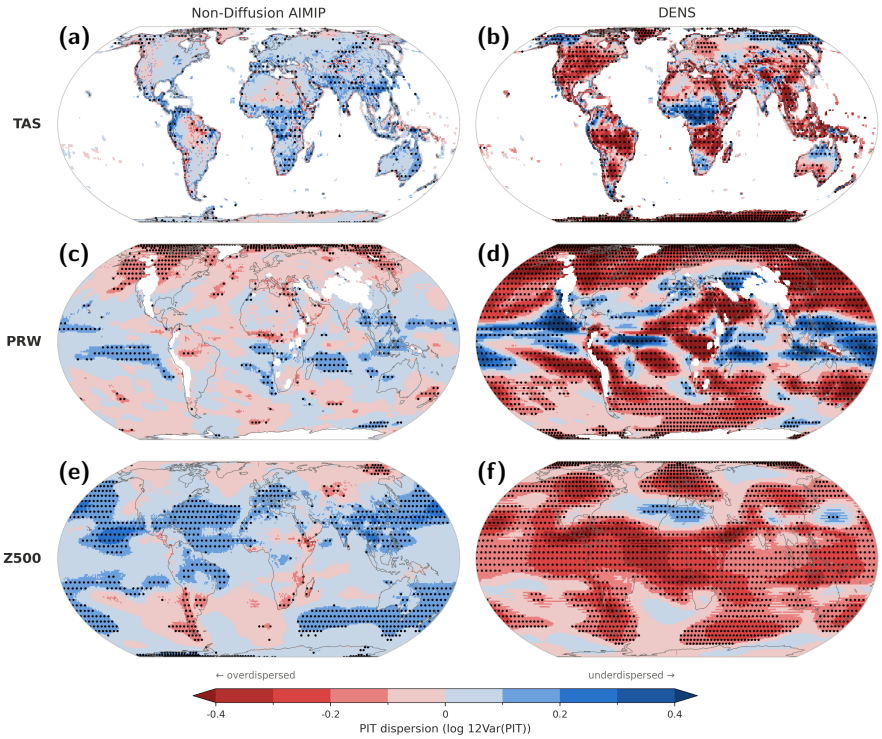

图 3 训练期内 (a) AIMIP TAS、(b) DENS TAS、(c) AIMIP PRW、(d) DENS PRW、(e) AIMIP Z500 和 (f) DENS Z500 的 PIT 离散度比。点状阴影表示显著性 ($p \leq 0.05$)，通过双尾蒙特卡洛检验确定；零分布是通过计算 N_{eff} 个标准均匀分布样本的方差，并重复 1000 次迭代生成的； N_{eff} 表示在给定网格点处估计 PIT 值方差时的有效样本数，并考虑了时间自相关 [37]。

边缘分布会形成 U 形直方图；相反，过度离散集合的 PIT 直方图会呈倒 U 形。

全球月度 PIT 直方图（图 4）与空间分布图呈现出相同的结果：DENS 在全球范围内表现出过度离散（Z500 的 D 指标显著），而 AIMIP 则表现出相对较弱的欠离散（TAS 的 D 指标显著）。对于 Z500，DENS 在高分位数和低分位数上的偏差相对对称，但 TAS 和 PRW 的偏差则偏向于过度代表暖干极端事件（图4b）。在测试期内，相较于 ERA5，AIMIP 非扩散模式在三个变量上均表现出欠离散。然而，由于相较于训练期而言样本量有限，D 指标未达到显著性（见补充材料部分 S6）。

相较于训练期，样本量有限，因此 D 指标未达到显著性（见补充材料部分 S6）。

在年时间尺度上，AIMIP 在 TAS 和 Z500 中仍表现为欠离散，并且在全球陆地的绝大多数区域，TAS 均呈现欠离散（见补充图 S10）。因此，在 AIMIP 的年时间尺度上，三个变量的全球PIT直方图均显著偏离均匀性（补充图S12）。相反，在全球大部分地区，DENS 中的年平均 500 hPa 位势高度和 PRW 呈过度离散；然而，DENS 中 TAS 的离散信号在空间上比月时间尺度下异质性更强（补充图 S10），且在全球范围内未达到显著性（补充图 S12）。

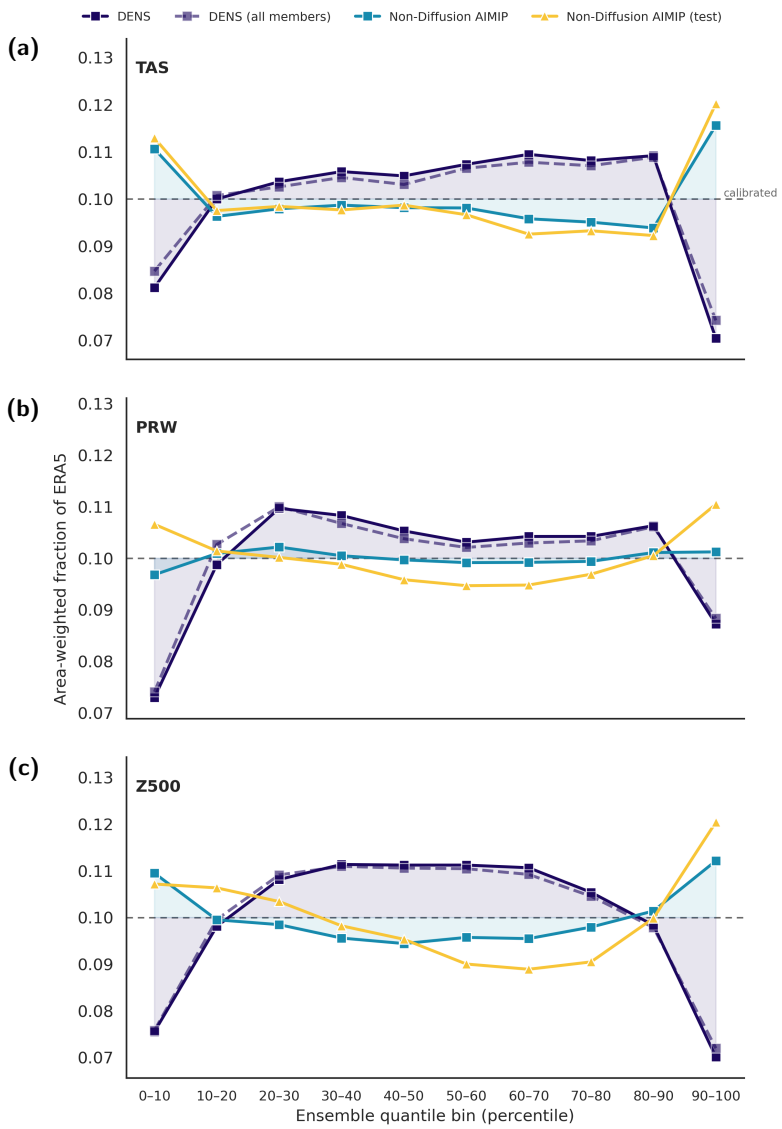

图 4 DENS 和非扩散型AIMIP模型中 TAS、PRW 和 Z500 的全局概率积分变换（PIT）直方图。校准完美的模式会生成均匀分布的 PIT 直方图。

4 讨论与结论

在本研究中，我们首次系统比较了人工智能气候模型的 AMIP风格多模式集合（AIMIP）之间的不确定性表征 [23]

以及动力学模型。我们发现，与动力学模型相比，人工智能集合在月和年时间尺度上的强迫变率较弱，内部变率和模型不确定性也更低。因此，在TAS、Z500和PRW方面，AIMIP中的气候极值弱于动力学集合，热带海洋上的可降水量极端值除外。变率受抑意味着AIMIP模型产生的气候状态范围比动力学模型更窄，同时对样本外的理想化变暖情景也表现出不符合物理规律的响应，这可能会限制其在气候风险评估中的效用。

与月时间尺度上的ERA5相比，AIMIP的Z500和TAS气候分布呈现轻微欠离散。相比之下，动力学集合的三个变量均表现出过度离散，且其离散程度高于AIMIP的欠离散信号。在年时间尺度上，与月时间尺度相比，AIMIP的三个变量仍均呈欠离散；而动力学集合中，TAS的校准失准并不显著，但PRW和Z500则呈现显著的过度离散。AIMIP相对于ERA5表现出更优的边际校准或许并不意外，因为人工智能模型的设计目标是在训练期内模拟ERA5。此外，AIMIP在测试期内呈现的欠离散程度与训练期相近或更强。因此，两种框架的不确定性表征都存在校准失准，但表现方式和程度各不相同，这意味着在当前的人工智能模拟器与动力学模型之间做出选择，取决于具体用途。

人工智能气候模型相较于物理模型具有明显优势：它们以较低的气候偏差实现了对厄尔尼诺—南方涛动的真实刻画，而计算成本仅为后者的一小部分 [23]。相反，不同领域的人工智能模型 [38–40] 可能会对训练数据过拟合。此前研究发现，人工智能气候模拟器会低估历史变暖趋势 [23, 41, 42] ，且难以应对未来气候情景 [14]。在这项研究中，我们表明，尽管其校准通常优于动力

模式（相对于ERA5而言），人工智能模型多模式集合（AIMIP）仍会低估变率和不确定性。

当前人工智能模型主要针对提升预测能力和降低平均偏差进行优化。然而，其不确定性量化需要经过校准并值得信赖。随着社区不断推动将人工智能气候模型整合到研究和业务流程中，本研究表明，我们需要优先投入精力，尤其是在多模式集合背景下，改进人工智能模型对内部变率和模型不确定性的表征。

数据可用性

AIMIP 模型输出及配套的 ERA5 数据可按照以下说明获取：https://github.com/ai2cm/AIMIP/blob/main/evaluations/README.md。DENS 模型输出取自 Pangeo CMIP6 Google Cloud Storage AMIP 实验，可通过 Intake-ESM（https://intake-esm.readthedocs.io/en/stable/index.html）访问。

代码可用性

本文所述所有分析所依据的代码将在发表后公开。

参考文献

[1] Hawkins, E. & Sutton, R. 缩小区域气候预测不确定性的潜力。《美国气象学会通报》90，1095–1108（2009）。网址 https://journals.ametsoc.org/view/journals/bams/90/8/2009bams2607_1.xml。

[2] Mariotti,A. 等。从次季节到季节及更长期精准预报的机遇窗口。《美国气象学会通报》**101**，E608–E625（2020）。网址 https://doi.org/10.1175/BAMS-D-18-0326.1。

[3] Deser, C., Phillips, A., Bourdette, V. & Teng, H. 气候变化预测中的不确定性：内部变率的作用。《气候动力学》**38**, 527–546 (2012). 网址 https://link.springer.com/article/10.1007/s00382-010-0977-x。

[4] 政府间气候变化专门委员会。《*2021*年气候变化：自然科学基础》：政府间气候变化专门委员会第六次评估报告第一工作组贡献（剑桥大学出版社，2023年）。

[5] Eyring,V. 等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**, 1937–1958 (2016). 网址 https://gmd.copernicus.org/articles/9/1937/2016/。

[6] Frankcombe, L. M., England, M. H., Kajtar, J. B., Mann, M. E. & Steinman, B. A. 在存在内部变率的情况下，选择集合平均来估算强迫信号。《气候杂志》**31**, 5681–5693 (2018). 网址https://doi.org/10.1175/JCLI-D-17-0662.1。

[7] Lehner,F. 等 利用多个大型集合与 CMIP5/6 对气候预测不确定性进行划分。《地球系统动力学》 **11**, 491–508（2020）。网址 https://esd.copernicus.org/articles/11/491/2020/。

[8] Shaw, T. A. 与 Stevens, B. 另一场气候危机。《自然》 **639**, 877–887（2025）。网址 https://www.nature.com/articles/s41586-025-08680-1。

[9] Bordoni, S., Kang, S. M., Shaw, T. A., Simpson, I. R. & Zanna, L. 气候建模的未来。《*npj*气候与大气科学》**8**，99（2025）。网址 https://www.nature.com/articles/s41612-025-00955-8。

[10] Mahesh,A. 等大型集合——第 2 部分：利用球面傅里叶神经算子生成的回报试验大型集合的性质。《地球科学模型开发》**18**，5605–5633（2025）。网址 https://gmd.copernicus.org/articles/18/5605/2025/。

[11] Lam, R. 等 学习高技巧的中期全球天气预报。《科学》 **382**,1416–1421 (2023)。网址https://www.science.org/doi/10.1126/science.adi2336。

[12] Bi, K. 等 使用三维神经网络进行精准的中期全球天气预报。《自然》 **619**, 533–538 (2023)。网址https://www.nature.com/articles/s41586-023-06185-3。

[13] Pathak,J. 等 FourCastNet：一种使用自适应傅里叶神经算子的全球数据驱动高分辨率天气模式（2022）。网址 https://arxiv.org/abs/2202.11214。 arXiv:2202.11214。

[14] Watt-Meyer, O. 等 ACE2：精准学习从次季节到年代际的大气变率与受迫响应。《*npj*气候与大气科学》**8**,205 (2025)。

[15] Kochkov, D.等用于天气与气候的神经环流模型。《自然》**632**, 1060–1066（2024年）。

[16] Couairon, G., Singh, R., Charantonis, A., Lessig, C. & Monteleoni, C. ArchesWeather & ArchesWeatherGen：一种用于高效机器学习天气预报的确定性生成模型。*arXiv* 电子预印本arXiv:2412.12971（2024年）。

[17] Clyne,G.,Couairon,G., Gastineau, G., Monteleoni, C. & Charantonis, A. Arch- esClimate：基于流匹配的概率性年代际集合生成。*arXiv* 预印本 arXiv:2509.15942（2025）。

[18] Cresswell-Clay, N. 等。 一种用于高效模拟观测气候的深度学习地球系统模式。 *AGU Advances* **6**，e2025AV001706（2025）。网址 https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2025AV001706。 E2025AV001706 2025AV001706。

[19]Brenowitz, N. D. 等。 《瓶中气候：迈向公里尺度全球大气生成式基础模型》。 *arXiv* 预印本arXiv:2505.06474（2025年）。

[20] Hall, K. J. C. & Molina, M. J.《Monthly Diffusion v0.9：用于首届 AI-MIP 的潜在扩散模型》。*arXiv* 预印本 arXiv:2604.13481（2026年）。

[21] Allen,A. 等。 端到端数据驱动天气预报。 《自然》 **641**，1172–1179（2025年）。 网址：https://www.nature.com/articles/s41586-025-08897-0。

[22] Selz, T. 等。 关于人工智能天气预报模型的有效分辨率。 地球系统人工智能（2026年）。网址：https://doi.org/10.1175/AIES-D-25-0121.1。在线先行发表。

[23] Henn, B. 等AIMIP 第一阶段：人工智能天气与气候模型的系统评估。*arXiv* 预印本arXiv:2605.06944（2026年）。

[24] Pallotta, G.等用于评估深度学习地球系统模型的受 PMP 启发的评估框架。 *arXiv* 预印本arXiv:2604.06567（2026年）。

[25] Dunne，J. P. 等 不断发展的第七阶段耦合模式比较计划（CMIP7）及其快速通道，以支持未来气候评估。 《地球科学模型开发》 **18**，6671–6700（2025）。网址 https://gmd.copernicus.org/articles/18/6671/2025/。

[26] Hersbach，H. 等 ERA5全球再分析。 英国皇家气象学会季刊 **146**，1999–2049（2020）。

[27] G´orski, K. M. 等 HEALPix：在球面上对分布数据进行高分辨率离散化和快速分析的框架。《天体物理学杂志》**622**, 759–771（2005）。

[28] Padull´es,R. 等 利用无线电掩星研究热带降水与自由对流层水汽的关系。 《大气科学杂志》 **79**, 1585 – 1600（2022）。网址：https://journals.ametsoc.org/view/journals/atsc/79/6/JAS-D-21-0052.1.xml。

[29] Northrop, P. J. & Chandler, R. E. 量化未来气候预测中的不确定性来源。 《气候杂志》 **27**, 8793 – 8808（2014）。网址：https://journals.ametsoc.org/view/journals/clim/27/23/jcli-d-14-00265.1.xml。

[30] Gneiting, T., Balabdaoui, F. 和 Raftery, A. E. 概率预报、校准与锐度。英国皇家统计学会会刊*B*辑：统计方法学**69**，243–268（2007）。网址 http://dx.doi.org/10.1111/j.1467-9868.2007.00587.x。

[31] Gneiting, T. 和 Ranjan, R. 预测分布的组合（2011）。网址 https://arxiv.org/abs/1106.1638。

[32] Nipen, T. 和 Stull, R. 校准数值天气预报集合的概率预报。《*Tellus A*》**63**, 858–875（2011）。网址 https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0870.2011.00535.x。

[33] Trenberth、K. E.、Smith、L.、Qian、T.、Dai、A. 和 Fasullo、J. 利用观测数据和模式数据估算全球水量平衡及其年循环。《水文气象学杂志》**8**，758–769（2007）。网址：http://dx.doi.org/10.1175/JHM600.1。

[34] Timmermann, A. 等厄尔˜尼诺—南方涛动的复杂性。《自然》**559**，535–545（2018）。网址http://dx.doi.org/10.1038/s41586-018-0252-6。

[35] Donohoe, A. & Battisti, D. S. 大气加热和温度的季节循环。《气候杂志》**26**, 4962–4980 (2013).

[36] Barnston, A. G., Chelliah, M. & Goldenberg, S. B. 赤道太平洋厄尔尼诺—南方涛动相关性极高的海表温度区域说明：研究札记。《大气—海洋》**35**, 367–383 (1997). 网址 http://dx.doi.org/10.1080/07055900.1997.9649597。

[37] Vehtari, A., Gelman,A., Simpson, D., Carpenter, B. & B¨urkner, P.-C. 秩归一化、折叠和局部化：用于评估马尔可夫链蒙特卡罗收敛性的改进版 rˆ（附讨论）。《贝叶斯分析》**16**, 667–718 (2021). 网址http://dx.doi.org/10.1214/20-BA1221。

[38] Ying, X. 过拟合概述及其解决方案。《物理学杂志：会议系列》**1168**，022022（2019）。网址：https://doi.org/10.1088/1742-6596/1168/2/022022。

[39]Kernbach, J. M. & Staartjes, V. E. Staartjes, V. E., Regli, L. & Serra, C.（编）基于机器学习的临床预测建模基础：第二部分——泛化与过拟合。（编者：Staartjes, V. E., Regli, L. & Serra, C.）《临床神经科学中的机器学习》，15–21（施普林格国际出版，查姆，2022）。

[40] Ma, C.等。理解并缓解视觉语言模型提示微调中的过拟合问题。《*IEEE*视频技术电路与系统汇刊》**33**，4616–4629（2023）。

[41] Singh,R. 等人 在多年代气候模拟中评估 ArchesWeather 和 ArchesWeather-ergen 的技能与稳定性（2026年）。网址https://arxiv.org/

abs/2605.29976。arXiv:2605.29976。

[42] Landsberg, J. B. & Barnes, E. A. 用昨日的气候预测未来：人工智能天气与气候模型中的温度偏差。《地球物理研究快报》**53**，e2025GL119740（2026年）。网址 https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2025GL119740。E2025GL119740 2025GL119740。

[43] Shapley, L. S. *17. n*人博弈的一个值，307–318（普林斯顿大学出版社，1953年）。网址http://dx.doi.org/10.1515/9781400881970-018。

[44] Rozemberczki, B.等。机器学习中的 *Shapley*值，IJCAI-2022，5572–5579（国际人工智能联合会议组织，2022年）。网址 http://dx.doi.org/10.24963/ijcai.2022/778。

[45] Yang, R.等 天气与气候预测中的可解释机器学习：综述。《大气环境》**338**, 120797（2024年）。网址 http://dx.doi.org/10.1016/j.atmosenv.2024.120797。

[46] Jansma, A. 使用 Shapley 值和高效采样进行气候模型因子分解（2026年）。网址https://arxiv.org/abs/2609.11948。

[47] Lunt, D. J. 等 数值模拟的多变量因子分解。地球科学模式发展 **14**, 4307–4317（2021年）。网址 http://dx.doi.org/10.5194/gmd-14-4307-2021。

[48] Lunt,D. J. 等 上新世中期暖期和极地放大的成因。《地球与行星科学快报》**321-322**, 128–138（2012年）。网址http://dx.doi.org/10.1016/j.epsl.2011.12.042。

[49] Karczmarz, A., Mukherjee, A., Sankowski, P. 和 Wygocki, P. 树模型特征重要性的改进计算：Shapley 与班扎夫（2021）。网址 https:

//arxiv.org/abs/2108.04126.

致谢

本研究部分由全球水安全中心通过 DoD HQ00342420014、美国能源部资助项目 #DE-SC0024212、美国国家科学基金会资助项目 #AGS-2210068、美国国家海洋和大气管理局（NOAA）资助项目 #NA24OARX431C0022，以及协调模型诊断与比较计划（PCMDI）地球系统建模项目提供支持。该项目由美国能源部科学办公室的地球与环境系统建模计划资助。

作者贡献

J.B.L.、D.M.M. 和 M.A.F. 参与了本项工作的构思、方法设计、数据采集与处理、分析和撰写。E.A.B. 参与了撰写与编辑，并对分析提出了建议。

人工智能披露

本项目中所述实验的代码以及用于制作出版物图表的脚本，均借助 Claude Code（Anthropic）生成。作者审阅并核实了代码仓库和相关手稿文本的内容，并对其准确性和正确性负责。

利益冲突

作者声明不存在任何竞争性利益。

补充材料：人工智能与动力气候集合中的不确定性比较

Jacob B. Landsberg（jlandsbe@bu.edu）、Devin M. McAfee、M.A. Fernandez 和 Elizabeth A. Barnes

S1 模型偏差 29

表 S1 - TAS、PRW 和 Z500 的平均偏差 .29

S2 内部变率与模型不确定性 30表 S2 - 按 AIMIP 模式划分的内部变率 .30图 S1 - 内部变率方差图 .30图 S2 - 模式不确定性图 .31

S3 自助法收敛 32

图 S3 - 模式间和模式内收敛 .32

S4 极值 33

图 S4 - 月度 PRW 极端差异的地图和分布 . . 33图 S5 - 月度 Z500 极端差异的地图和分布 . . 34图 S6 - 年度 TAS 极端差异的地图和分布 . . . 35图 S7 - 年度 PRW 极端差异的地图和分布 . . . 36图 S8 - 年度 Z500 极端差异的地图和分布 . . . 37

S5 补充 PIT 离散度图 38

图 S9 - AIMIP测试时段的 PIT 离散度 .38图 S10 - 12个月平滑的 PIT 离散度 .39

S6 PIT D度量显著性 40

表 S3 - PIT D度量自助法结果 .40

S7 补充 PIT 直方图 41图 S11 - AIMIP 扩散模型的全球 PIT .41图 S12 - AIMIP 非扩散模式的全球 12 个月 PIT 42

S8 方差分解 43

S9 小样本偏差校正 44图 S13 - 偏差校正方差分量 .45

S10分位数差异归因 46

S10.1 Shapley结果 .46图 S14 - 月度全球平均Shapley值 .47图 S15 - 月度Shapley值地图 .47图 S16 - 与图 S14 相同，但针对年度平滑数据。 . . . . . 48图 S17 - 与图 S15 相同，但针对年度平滑数据。 . . . . . 48S10.2 Shapley值 .49S10.3 与因子分离的关系 .49S10.4 实现 .50

S1 模式偏差

表 S1 DENS 和 AIMIP 相对于 ERA5 的 TAS、PRW 和 Z500 平均偏差。

Model	TAS bias (K)	PRW bias (mm)	Z500 bias (m)
DENS			
NorCPM1	-0.16	+1.17	+10.8
MIROC6	+2.82	+1.78	-11.8
GISS-E2-1-G	+0.34	+1.21	-30.0
CESM2	+0.07	+0.38	+6.2
IPSL-CM6A-LR	-0.86	+1.02	-40.3
mean ± σ	**+0.44 ± 1.40**	**+1.11 ± 0.50**	**-13.0 ± 22.2**
AIMIP train			
NeuralGCM-HRD	+0.07	+0.27	-7.2
MD1.5	-0.47	+0.36	+1.3
DLESyM	-0.19	—	-4.2
ACE2.1-ERA5	+0.09	+0.05	+0.4
ArchesWeather-V2	+0.14	+0.13	+0.3
cBottle-1-3	+0.09	+0.10	-1.2
ArchesWeatherGen-V2	+0.16	+0.14	+0.2
mean ± σ	**-0.02 ± 0.23**	**+0.17 ± 0.12**	**-1.5 ± 3.1**
AIMIP test			
NeuralGCM-HRD	-0.08	+0.30	-9.5
MD1.5	-1.15	+0.05	-8.5
DLESyM	-0.40	—	-7.4
ACE2.1-ERA5	-0.60	-0.29	-11.0
ArchesWeather-V2	-0.07	-0.01	-4.3
cBottle-1-3	-0.65	-0.26	-11.3
ArchesWeatherGen-V2	-0.18	-0.01	-5.5
mean ± σ	**-0.45 ± 0.39**	**-0.04 ± 0.22**	**-8.2 ± 2.6**

S2 内部变率与模型不确定性

表 S2 训练期（1979–2014）、测试期（2015–2024）和 +2K 模拟中各 AIMIP 模式的月平均全球平均内部变率，以时间自助抽样的平均值表示。粗体值表示模式训练期与测试期或 +2 K 模拟的自助抽样范围（最小值至最大值）不重叠。

Model	TAS (K²)	PRW (mm²)	Z500 (m²)
ACE2.1-ERA5	2.15 (2.16, **3.93**)	1.98 (2.03, **4.06**)	1079 (1090, **1721**)
ArchesWeather-V2	2.56 (2.42, 2.45)	1.89 (1.98, **2.19**)	1249 (1235, 1239)
ArchesWeatherGen-V2	2.43 (2.35, 2.48)	1.85 (1.92, **2.17**)	1210 (1199, 1210)
DLESyM	2.36 (2.28, **1.88**)	—	1043 (1093, 1008)
NeuralGCM-HRD	2.37 (2.32, 2.23)	2.10 (**2.26**, **2.87**)	1257 (1239, 1192)
MD1.5	0.678 (0.668, 0.630)	0.231 (0.228, 0.237)	498 (488, **458**)
cBottle-1-3	1.14 (1.20, —)	0.704 (**0.769**, —)	381 (405, —)

Train (Test, +2K)

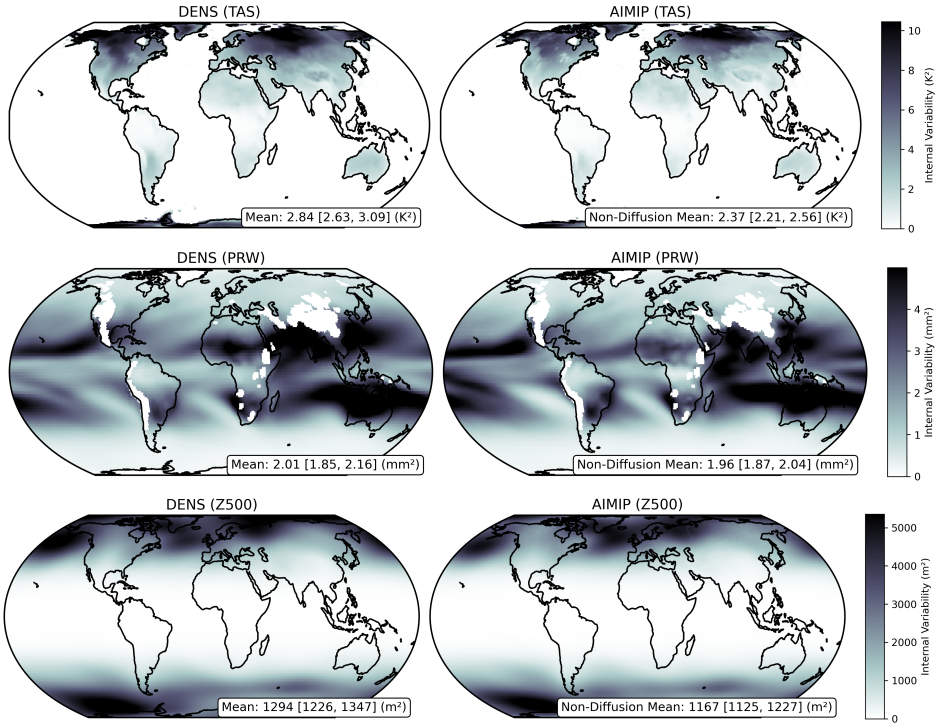

图 S1 DENS 和 AIMIP 模型的 TAS、PRW 和 Z500 变量内部变率。对于 TAS，已屏蔽海洋区域，以避免模型强迫产生过大影响；模型强迫主要由指定海表温度驱动。对于所有变量，DENS 的内部变率均大于 AIMIP。

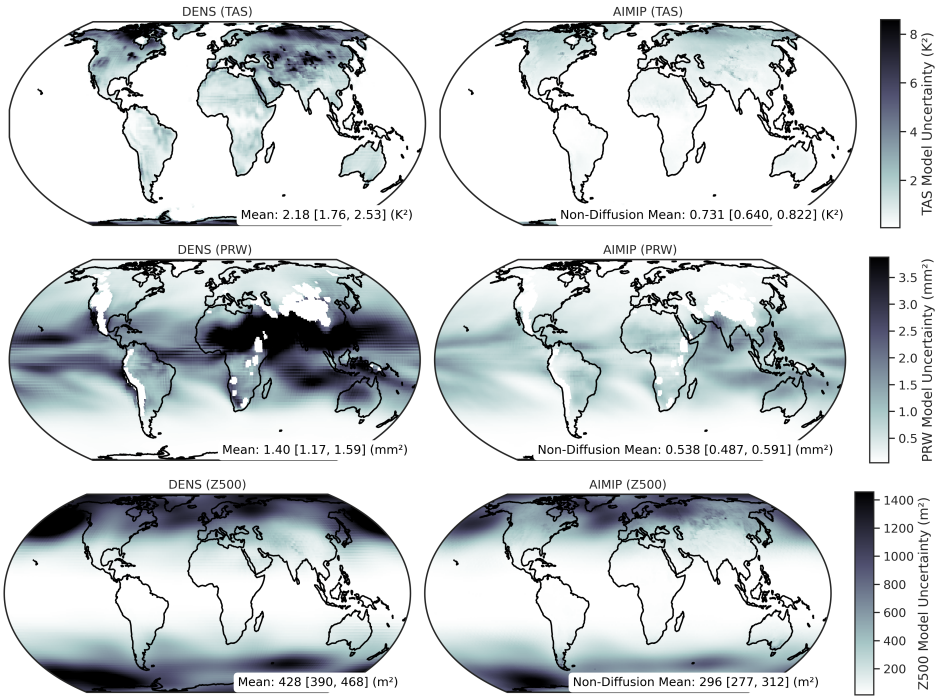

图 S2 DENS 和 AIMIP 模型在 TAS、PRW 和 Z500 变量上的模式不确定性。对于 TAS，已屏蔽海洋区域，以避免模型强迫造成过大的影响；模型强迫主要由指定海表温度驱动。对于所有变量，DENS 的模式不确定性均大于 AIMIP。

S3 自助法收敛

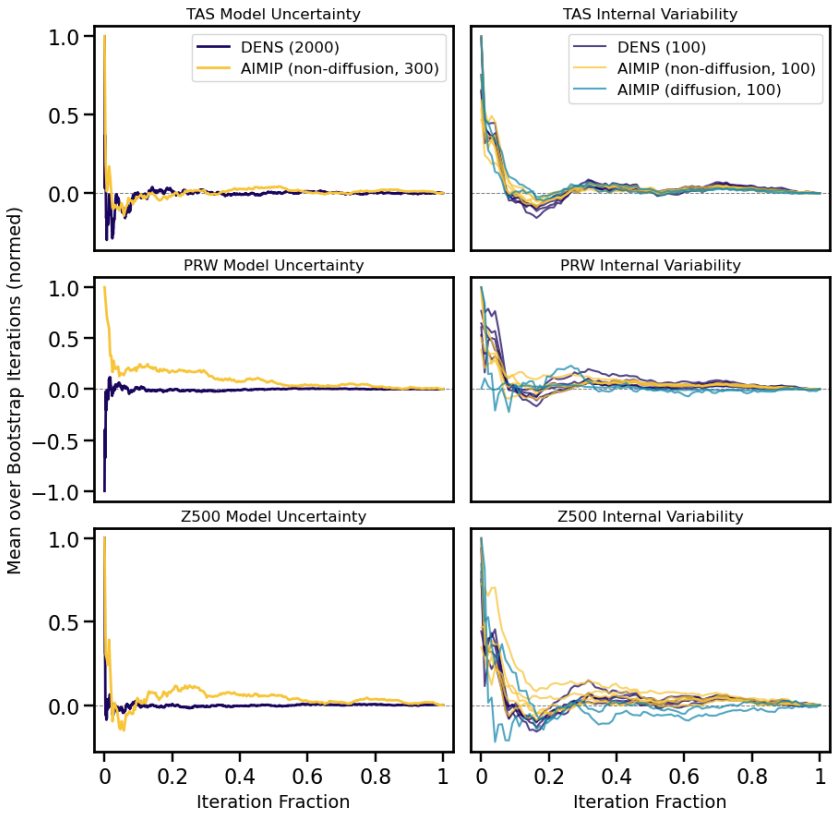

图 S3 模式间方差和模式内方差自助法的收敛情况。对于每个变量（TAS、PRW、Z500），图中显示了截至该次迭代的所有自助法迭代的平均值（方括号中为所示迭代次数）。在相应分析中，数值均按时间、纬度、经度和模式取平均。左侧面板显示 DENS 模式的模式间方差收敛情况，其中包含一个额外的子抽样步骤，以考虑每个模式成员数量的差异；同时还显示 AIMIP 非扩散模式的结果。右侧面板显示模式内方差收敛情况，不同颜色对应不同的模式样本。

S4 极值

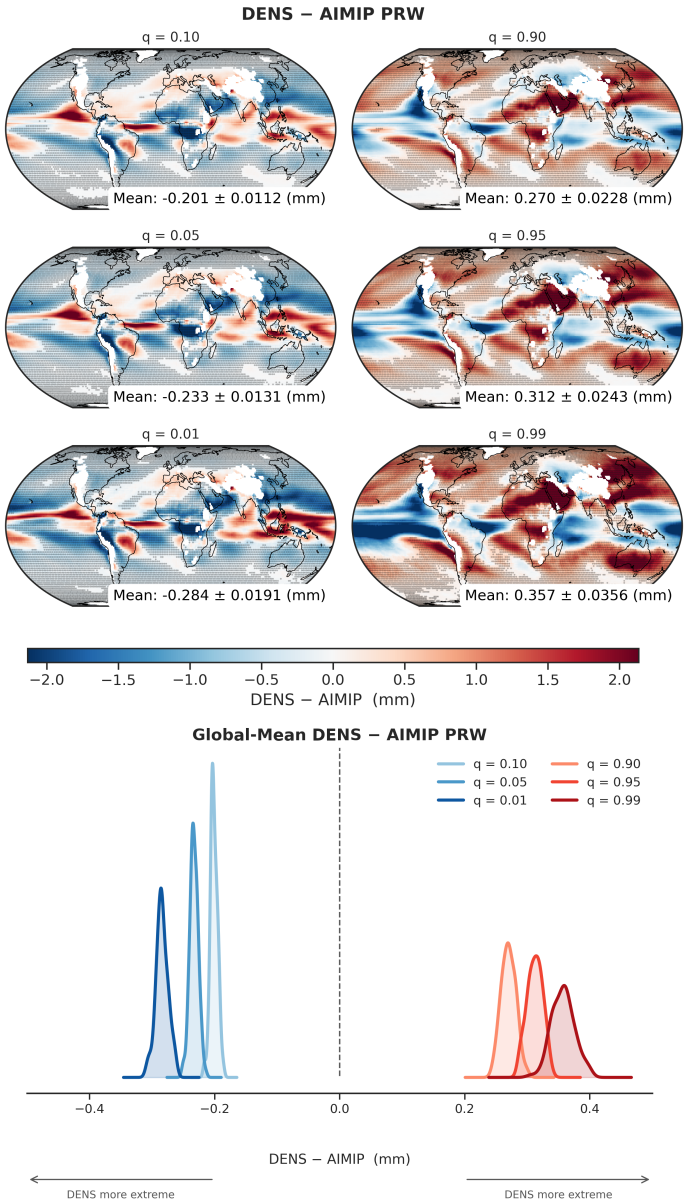

图 S4 与图 2相同，但针对 PRW。

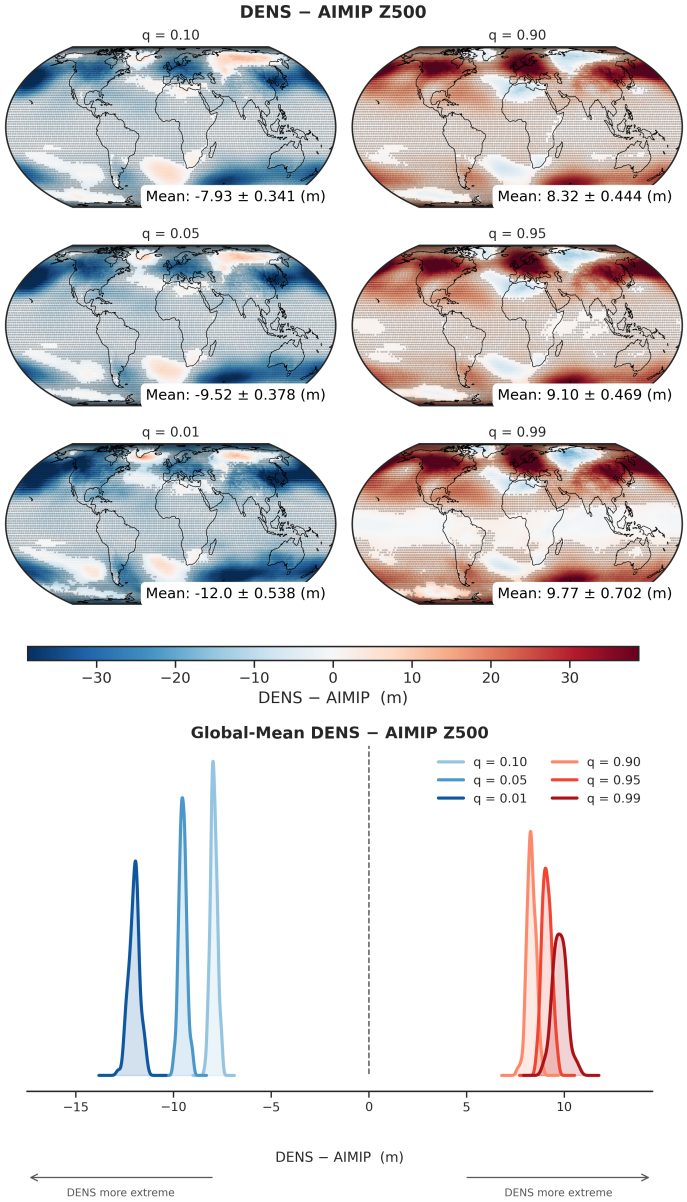

图 S5 与图 2相同，但针对 Z500。

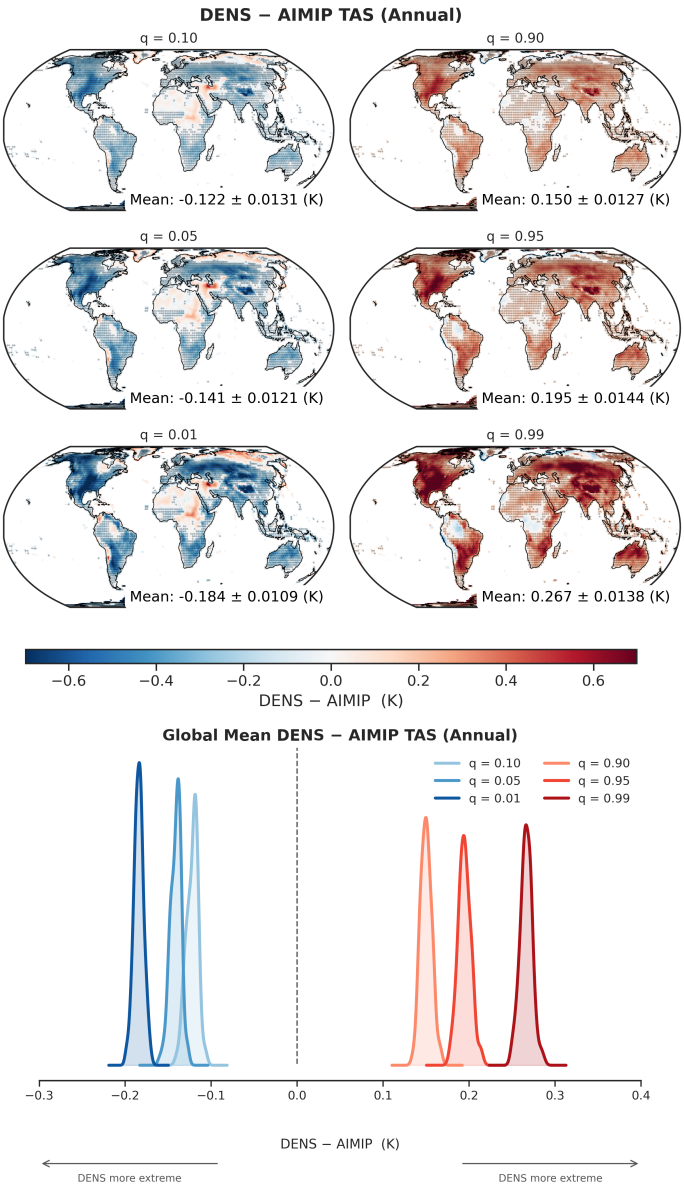

图 S6 与图 2相同，但针对年度 TAS。

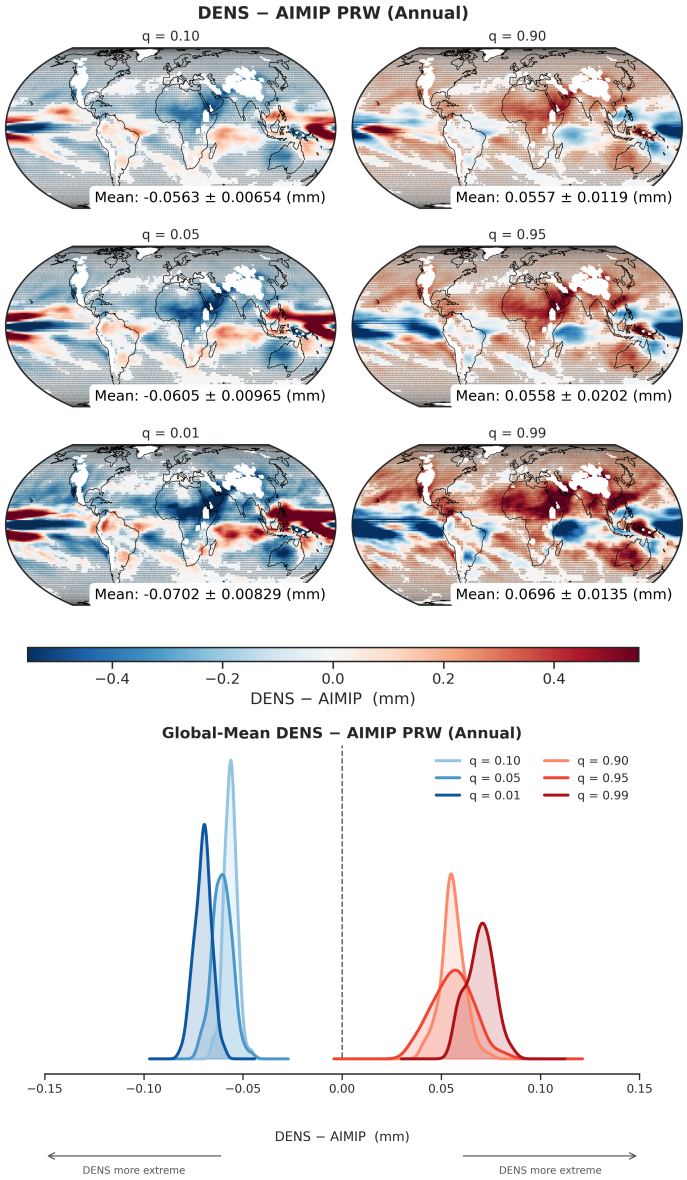

图 S7 与图 2相同，但针对年度 PRW。

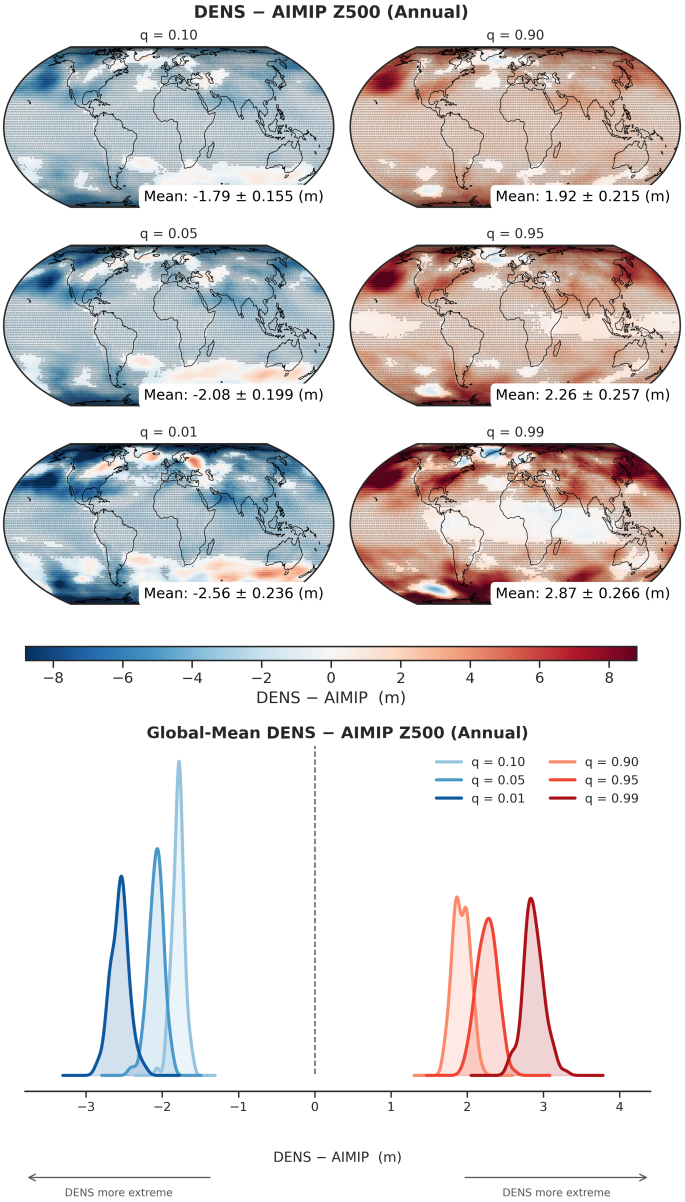

图 S8 与图 2相同，但针对年平均 500 hPa 位势高度。

S5 补充 PIT 离散度图

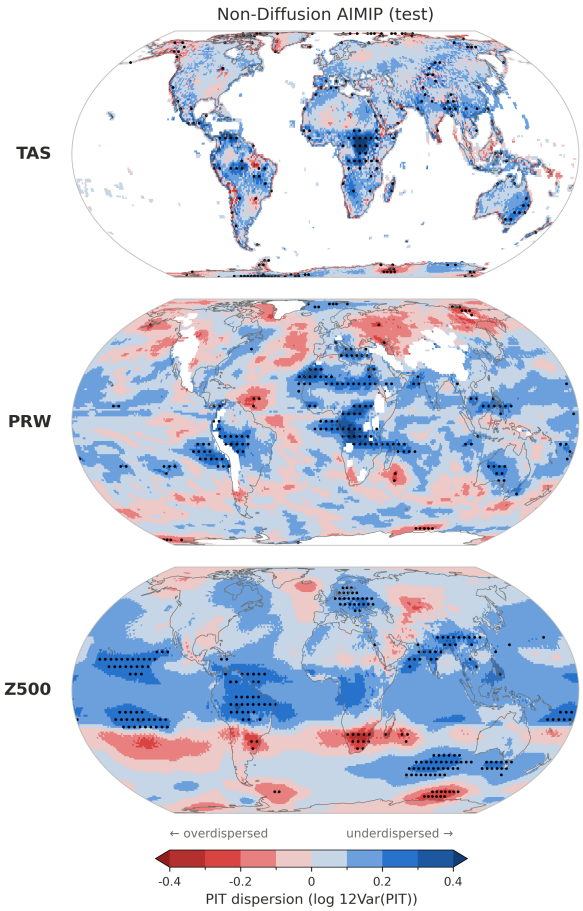

图 S9 与图 3相同，但展示测试期（2015–2024）内非扩散型AIMIP模型的结果。

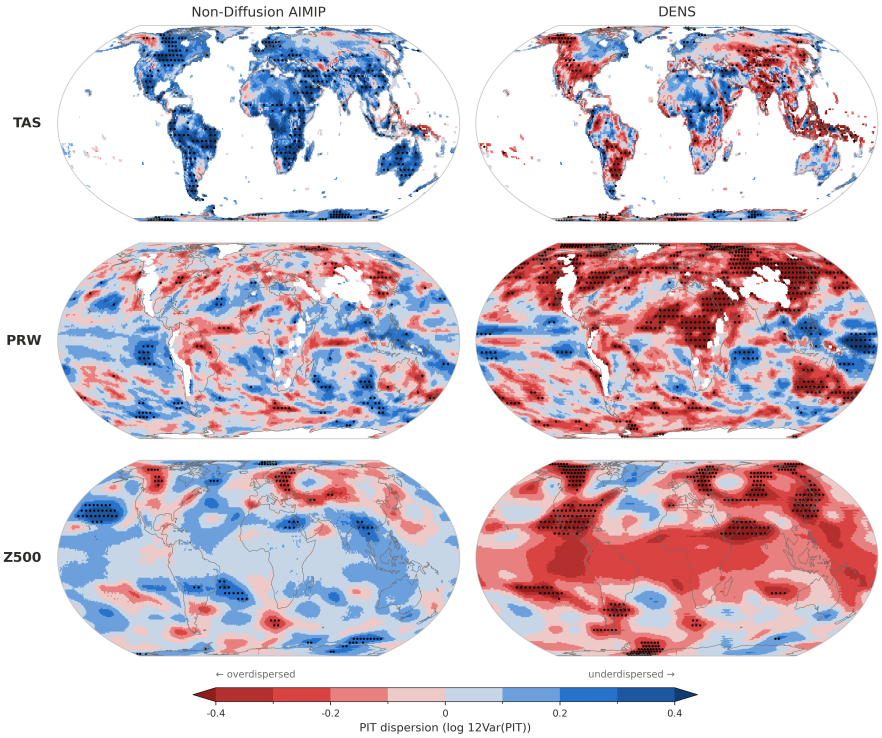

图 S10 与图3相同，但使用训练期（1979–2014）的12个月平滑数据。

S6 PIT D度量显著性

为衡量边际校准的偏差，我们报告 D 指标（参见 Nipen 和 Stull [32] 以及公式 S1）。

$$D = \sqrt{\frac{1}{B}\sum_{i=1}^{B}\left(\frac{b_i}{\|T\|} - \frac{1}{B}\right)^2}, \tag{S1}$$

这里， B 是分箱数（本例中为 10）， b_i 是分配到分箱 i的 ERA5 输出数， $\|T\|$ 是 ERA5 样本总数。计算 $\frac{b_i}{\|T\|}$ 时，我们采用余弦纬度加权。

为评估计算得出的 D 指标是否显著超出抽样噪声的影响，我们通过选取单个成员作为伪观测值，构建 D 指标的零分布。我们依据排除该成员后由其框架形成的十分位数对该成员进行分箱，计算 D，并对所有成员重复此过程。如果观测到的 D 超过该零分布的第90百分位数，则以粗体标示。

表 S3 图 4中 PIT 直方图对应的偏差度量。粗体数值表示，在 90% 或更多的自助法迭代中，偏差大于抽样噪声所预期的偏差。

Statistic	TAS	PRW	Z500
DENS			
D_{obs} (1-month)	1.26×10^{-2}	1.08×10^{-2}	$\mathbf{1.44 \times 10^{-2}}$
D_{obs} (12-month)	1.73×10^{-3}	9.27×10^{-3}	$\mathbf{1.62 \times 10^{-2}}$
AIMIP train — non-diffusion			
D_{obs} (1-month)	$\mathbf{6.80 \times 10^{-3}}$	1.44×10^{-3}	5.80×10^{-3}
D_{obs} (12-month)	2.28×10^{-2}	3.11×10^{-3}	7.97×10^{-3}
AIMIP test — non-diffusion			
D_{obs} (1-month)	$\mathbf{8.74 \times 10^{-3}}$	4.89×10^{-3}	9.26×10^{-3}
D_{obs} (12-month)	$\mathbf{1.90 \times 10^{-2}}$	5.47×10^{-3}	1.63×10^{-2}

S7 补充 PIT 直方图

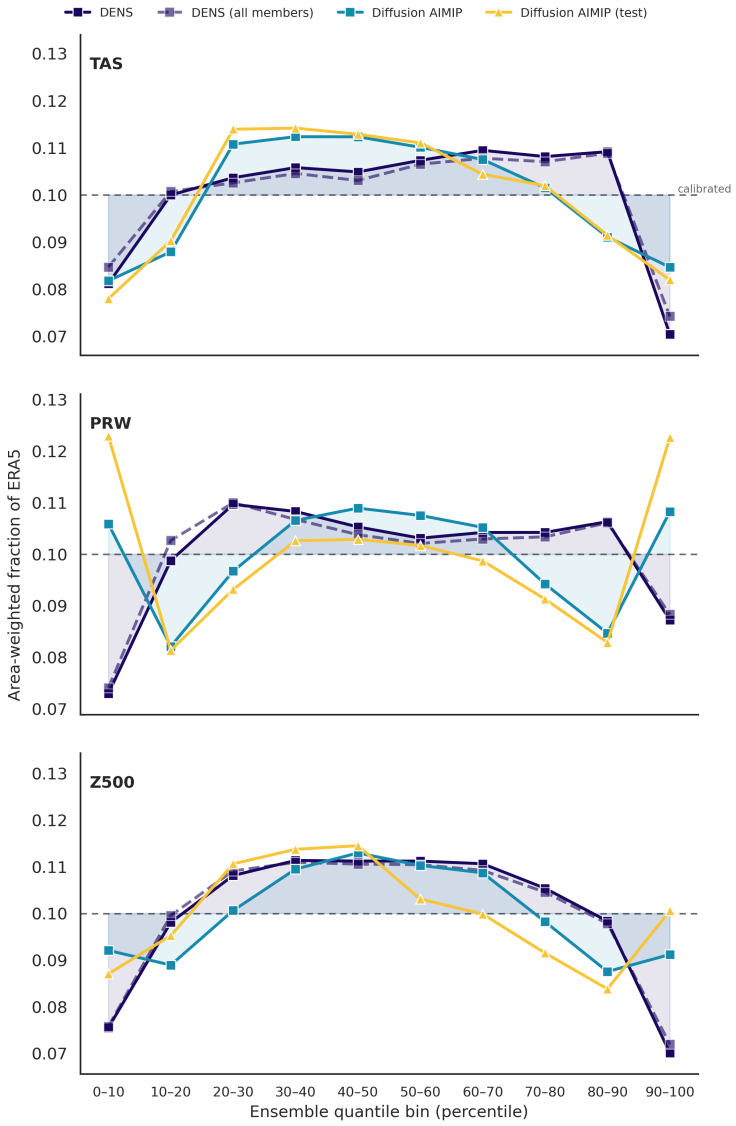

图 S11 与图 4相同，但针对 AIMIP 扩散模型。

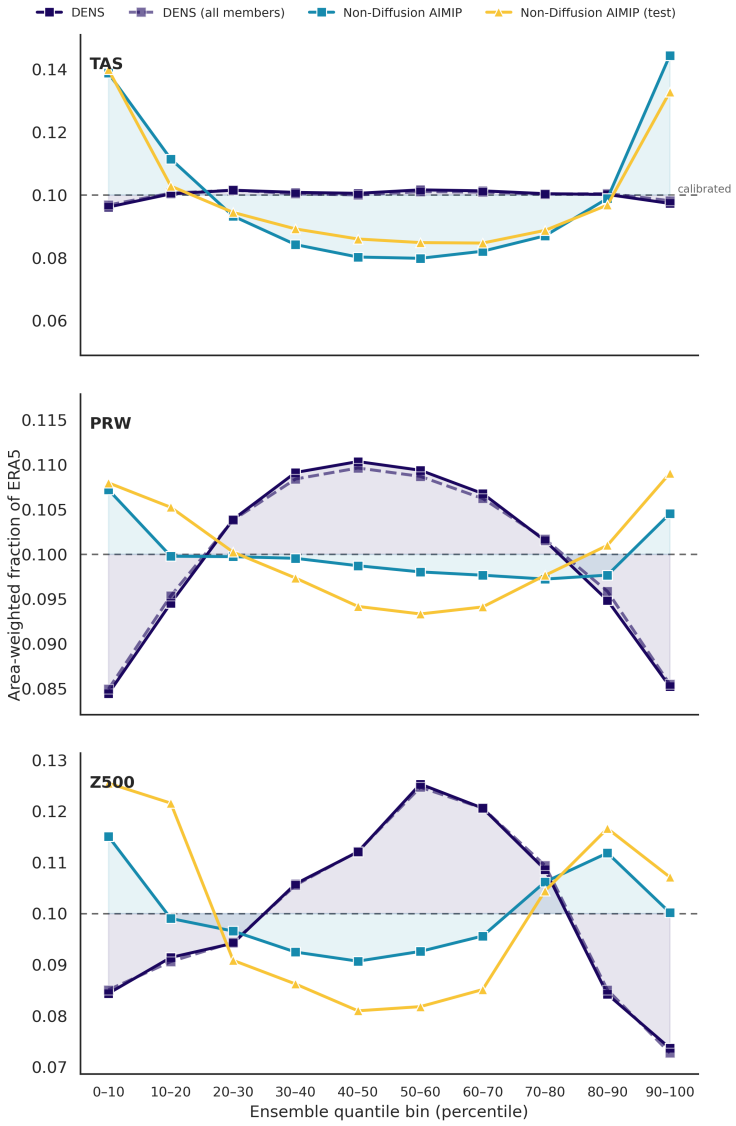

图 S12 与图 4相同，但使用12个月平滑数据。

S8 方差分解

设 Y 表示多模式集合中任意位置的某个气候变量输出。通俗地说，我们利用全方差定律，将 Y 在不同模式、成员和时间上的总方差分解为以下几项之和：模式内方差的模式平均和时间平均（内部变率）、时间平均的模式间方差（模型不确定性），以及多模式平均的时间方差（强迫变率）：

$$
\begin{aligned}
\mathrm{Var}(Y) &= \mathbb{E}[\mathrm{Var}(Y \mid T)] + \mathrm{Var}(\mathbb{E}[Y \mid T]) \\
&= \mathbb{E}[\mathbb{E}[\mathrm{Var}(Y \mid M, T) \mid T]] + \mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid M, T] \mid T)] + \mathrm{Var}(\mathbb{E}[Y \mid T]) \\
&= \mathbb{E}[\mathrm{Var}(Y \mid M, T)] + \mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid M, T] \mid T)] + \mathrm{Var}(\mathbb{E}[Y \mid T]) \\
&= \mathbb{E}[\text{intramodel variance}] + \mathbb{E}[\text{intermodel variance}] + \text{forced variability}
\end{aligned}
\tag{S2}
$$

其中，M 表示模式，T 表示时间步。在第 2 行和第 3 行之间，我们利用全期望定律，将 $\mathbb{E}[\mathbb{E}[\mathsf{Var}(Y \mid M, T) \mid T]] = \mathbb{E}[\mathsf{Var}(Y \mid M, T)]$ 写为。

由于我们去除了每个模式的时间平均值，因此可以将模式间方差项和强迫方差项合并为一个单独的项，表示各模式单模式强迫方差的平均值。

$$
\begin{aligned}
\mathrm{Var}(Y) &= \mathbb{E}[\mathrm{Var}(Y \mid M)] + \mathrm{Var}(\mathbb{E}[Y \mid M]) \\
&= \mathbb{E}[\mathrm{Var}(Y \mid M)] \\
&= \mathbb{E}[\mathbb{E}[\mathrm{Var}(Y \mid M, T) \mid M]] + \mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid M, T] \mid M)] \\
&= \mathbb{E}[\mathrm{Var}(Y \mid M, T)] + \mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid M, T] \mid M)] \\
&= \mathbb{E}[\text{intramodel variance}] + \mathbb{E}[\text{single-model forced variance}]
\end{aligned}
\tag{S3}
$$

S9 小样本偏差校正

在本研究中，我们使用有限集合的样本估计值，其中有 m 个模式，每个模式有 $n = 5$ 个成员（$m = 5$，但 AIMIP PRW 除外，因为省略了 $m = 4$，如 DLESyM）。将内部变率、模型不确定性和强迫变率的样本估计值分别记为 I、M 和 F，在平衡随机效应模式下，它们的期望值为

$$\begin{aligned} \mathbb{E}[\hat{I}] &= \sigma_I^2, \\ \mathbb{E}[\hat{M}] &= \sigma_M^2 + \frac{\sigma_I^2}{n}, \\ \mathbb{E}[\hat{F}] &= \sigma_F^2 + \frac{\sigma_M^2}{m} + \frac{\sigma_I^2}{mn}, \end{aligned} \tag{S4}$$

其中，σ_I^2、σ_M^2 和 σ_F^2 分别是总体内部变率、模型不确定性和强迫变率。另一方面，$\frac{\sigma_I^2}{n}$ 和 $\frac{\sigma_M^2}{m} + \frac{\sigma_I^2}{mn}$ 是有限集合中抽样偏差对应的项。

扣除抽样误差偏差后的方差分量如图S13所示。校正降低了两个框架中的模型不确定性，并略微降低了强迫变率。正文中的定性结论未变：DENS 的内部变率、模型不确定性和强迫变率均大于 AIMIP。

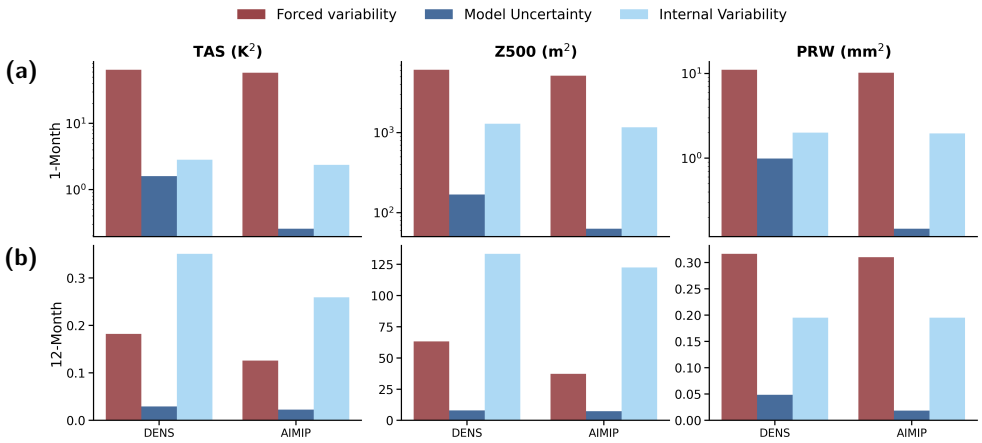

图 S13 与图 1相同，但对模型不确定性和强迫变率进行了小样本偏差校正（第 S9节）。展示了训练期内（a）月度和（b）年度（12个月平滑）结果。

S10 分位数差异归因

需要指出的是，由于图 2 中的分位数差是相对于各框架的模式平均分位数而言的，因此，下文 Shapley 归因中所考虑的“强迫变率”与单模式强迫方差的平均值有关。这取决于多模式平均的变率，也会受到模型不确定性的影响。也就是说，下文强迫变率的 Shapley 贡献不能直接与图 1所示的强迫变率相比较，因为它们也受到模型不确定性的影响。预期的单模式强迫方差大致相当于月时间尺度上的强迫变率，因为在月时间尺度上，强迫变率总体上是占主导地位的项（图 1）。然而，对于年度平滑数据，模型不确定性和强迫变率都会影响观测到的单模式强迫方差，尤其是 TAS 和 Z500。

在 S10.1 节中，我们展示了月和年时间尺度上强迫变率与内部变率的全球平均及空间 Shapley 值。我们在 S10.2 节介绍 Shapley 值的背景，并在 S10.3 节说明其与因子分离的关系。我们在 S10.4 节详细介绍分位数差异归因的实现。

S10.1 Shapley结果

由于我们通过构建从 DENS 到 AIMIP 的路径来进行归因（见补充材料第 S10.4 节），所得的 Shapley 值与 AIMIP 减去 DENS 的分位数差相关，这与图 2 中的符号相反。因此，负的 Shapley 值表示该组分——无论是外强迫还是内部变率——对图 2 中的正分位数差有贡献；而正的 Shapley 值表示其产生了负贡献。

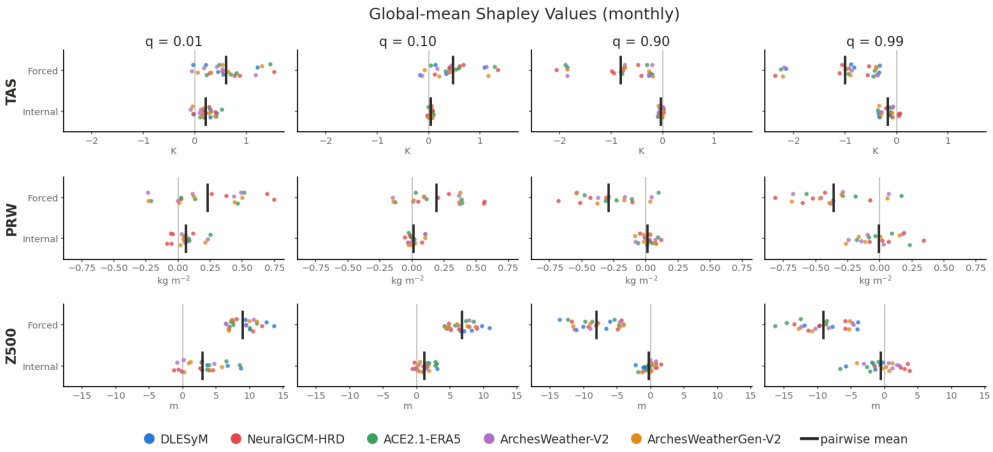

图 S14 月度强迫变率和内部变率的全球平均成对 Shapley 值，相对于 AIMIP-DENS 分位数差异（在训练期内），以及成对平均 Shapley 值。

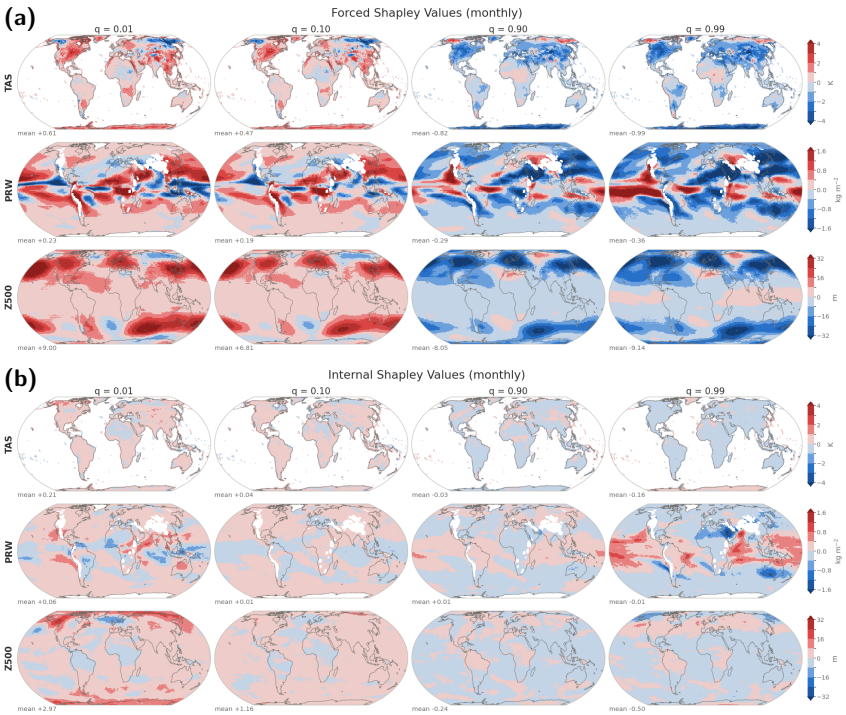

图 S15 成对　　平均月度 (a) 强迫变率和 (b)　　int部变率。

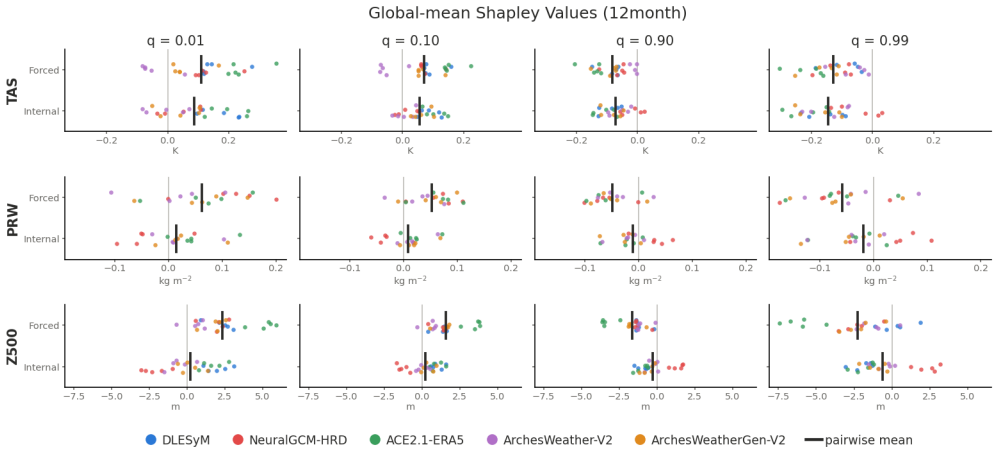

图 S16 同图 S14，但使用年度平滑数据。

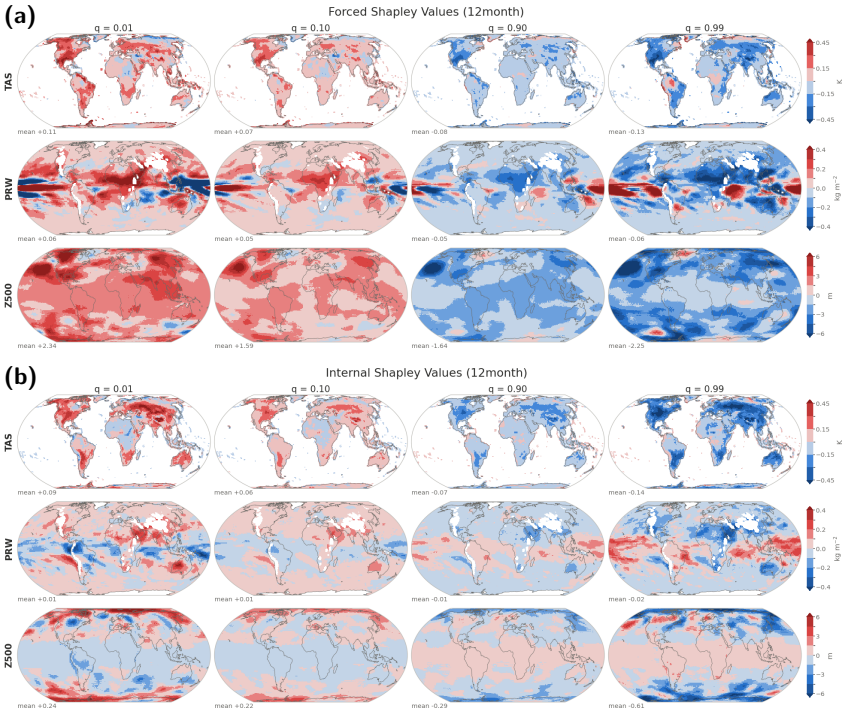

图 S17 同图 S15，但使用年度平滑数据。

S10.2 Shapley值

在图2 中，我们发现 AIMIP 产生的极端值弱于 DENS。在这里，我们试图将观察到的分位数差异归因于这两个框架之间内部变率和强迫变率的差异。为此，我们使用合作博弈论中的 Shapley值 [43]。合作博弈由一组参与者 $\mathcal{N}$（称为大联盟）和一个价值函数 $v(S)$ 组成，该价值函数为参与者联盟 $S \subseteq \mathcal{N}$分配收益。参与者 i对联盟 S 的边际贡献，是将参与者 i 加入联盟后所增加的价值， $v(S \cup \{i\}) - v(S)$。

Shapley值是一种解概念，或用于确定如何在参与者之间分配大联盟的收益 $v(\mathcal{N})$ 的规则。Shapley值 ϕ_i考虑参与者的所有可能排列，并根据参与者 i 在每种排列中对其前序联盟 i 的平均边际贡献给予相应收益：

$$\phi_i(v) = \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \frac{|S|!\ (|\mathcal{N}| - |S| - 1)!}{|\mathcal{N}|!}\ (v(S \cup \{i\}) - v(S)) \tag{S5}$$

也就是说，要计算某个参与者的 Shapley值，我们会考察所有不包含该参与者的联盟，并计算该参与者对每个联盟的边际贡献。随后，我们对这些贡献取平均值，并根据该联盟在所有排列中位于该参与者之前的频率，对其进行加权， $\frac{|S|!\ (|\mathcal{N}|-|S|-1)!}{|\mathcal{N}|!}$。

Shapley值是合作博弈论中唯一满足四项公平性公理的解概念：效率、对称性、虚设参与者和线性 [详见 43，了解更多详情]。Shapley值已广泛用于可解释人工智能中的公平特征归因 [44]，包括天气与气候领域的各种应用 [例如，45]。

S10.3 与因子分离的关系

Shapley值与数值气候建模中的因子分离[46]密切相关。因子分离是一类用于将气候模式输出的变化归因于模式组成部分变化的方法

，相对于基准状态而言， [47]这些组成部分以二元因子的形式呈现，例如，规定的二氧化碳来自工业化前基准还是变暖情景。

因子分离需要针对各个因子子集运行模拟。归因时，考虑从基准状态到扰动状态的非回返路径，每次加入一个因子。对于给定路径，每个因子的贡献都以该路径上相对于前一个因子子集的模式输出变化来衡量。对于 N 个因子，Lunt等人 [47] 提出的因子分离线性求和方法会针对所有 $N!$条路径计算贡献。每个因子的归因值定义为其在所有路径上的平均贡献[47], ，这等同于该因子的Shapley值 [46]；此处，价值函数表示相对于基准状态的模式输出变化。线性求和方法是对Lunt等人 [48], 所述技术的扩展；该技术同样会对其余因子的所有子集上的因子贡献取平均，但对每个子集赋予相同权重。这等同于班扎夫值；对于 $N = 2$个因子，该值与Shapley值相等，但通常不满足效率公理 [43, 49]。对于 $N > 2$ 而言，效率的缺失正是Lunt等人 [47] 扩展该技术的动机，从而得到一种修订后的、等同于Shapley值的方法。

我们的实现不同于传统的因子分离方法，因为我们不会针对每个因子子集重新运行 AIMIP 或 DENS 模拟。相反，如下文所述，我们会事后交换两个集合中的变异性成分，创建合成混合集合，以构建从 DENS 到 AIMIP 的路径。

S10.4 实现

对于 q给定网格点处的分位数，我们将模式平均分位数差 ΔQ改写为所有 AIMIP-DENS 模型对之间成对差异的平均值。请注意，这与图2中所示的分位数差互为相反数。

$$\Delta Q = \frac{1}{N_A N_D} \sum_{i=1}^{N_A} \sum_{j=1}^{N_D} \left(Q_A^{(i)} - Q_D^{(j)} \right) \tag{S6}$$

其中， N_A 表示 AIMIP 模型的数量（TAS 和 Z500 为 5，PRW 为 4），N_D 表示 DENS 模型的数量（每个变量均为 5）， $Q_A^{(i)}$ 和 $Q_D^{(j)}$ 分别表示第 i个 AIMIP 模式和第 j个 DENS 模式的第 q个分位数。

我们用 $Q(F, I)$ 表示将内部变率 I 加到强迫变率 F 后得到的时间序列的第 q 分位数。具体而言， F 表示给定模式的集合平均，该模式属于 AIMIP 或 DENS；而 I 则是同一模式或不同模式的成员残差。如果 F 属于与 I 相同的模式，那么 $Q(F, I)$ 就是该模式的第 q 分位数。否则， $Q(F, I)$ 就是第 q 分位数，它来自一个混合集合：将一个模式的成员平均与另一个模式的成员残差相结合。

我们考虑一个有两名参与者的博弈 $\mathcal{N}_\mathcal{A} = \{F_A, I_A\}$，其中 $F_A = \{F_A^{(i)}\}_{i=1}^{N_A}$ 和 $I_A = \{I_A^{(i)}\}_{i=1}^{N_A}$ 分别表示每个 AIMIP 模式的强迫变率和内部变率。对于方程 S6 中的每一组成对比较，我们定义一个博弈 $(\mathcal{N}_\mathcal{A}, v_{ij})$，其中 v_{ij} 定义为

$$v_{ij}(S) = Q(F, I) - Q(F_D^{(j)}, I_D^{(j)}) \tag{S7}$$

$$\text{where} \begin{cases} F = F_A^{(i)} \text{ if } F_A \in S \text{ else } F_D^{(j)} \\ I = I_A^{(i)} \text{ if } I_A \in S \text{ else } I_D^{(j)} \end{cases}$$

价值函数 v_{ij} 衡量联盟 S 形成的混合集成与 DENS 模式 j之间的分位数差。对于每个成对博弈，我们计算 F_A 和 I_A的 Shapley 值。

$$\phi_{F_A}(v_{ij}) = \frac{1}{2} \left(v_{ij}(\{F_A\}) + v_{ij}(\{F_A, I_A\}) - v_{ij}(\{I_A\}) \right) \tag{S8}$$

$$\phi_{I_A}(v_{ij}) = \left(Q_A^{(i)} - Q_D^{(j)}\right) - \phi_{F_A}(v_{ij}) \tag{S9}$$

方程S9 由 Shapley 值的效率性质推出 [43], 该性质要求 F_A 和 I_A 对于一个模型对，其 Shapley 值之和等于两个模式之间的分位数差。

合并方程S6 和S7，我们定义一个价值函数，用于表示联盟S 与DENS形成的混合多模式集合之间的模型平均分位数差。

$$v(S) = \frac{1}{N_A N_D} \sum_{i=1}^{N_A} \sum_{j=1}^{N_D} v_{ij}(S) \tag{S10}$$

根据Shapley值的线性性质 [43], , 相对于AIMIP-DENS分位数差 ΔQ， F_A 和I_A 的Shapley值等于成对Shapley值的平均值。

$$\phi_{F_A}(v) = \frac{1}{N_A N_D} \sum_{i=1}^{N_A} \sum_{j=1}^{N_D} \phi_{F_A}(v_{ij}) \tag{S11}$$

$$\phi_{I_A}(v) = \frac{1}{N_A N_D} \sum_{i=1}^{N_A} \sum_{j=1}^{N_D} \phi_{I_A}(v_{ij}) \tag{S12}$$