MoWE：气象专家混合模型

Dibyajyoti Chakraborty, Romit Maulik
The Pennsylvania State University
State College, PA 16801, United States
d.chakraborty@psu.edu

Peter Harrington, Dallas Foster, Mohammad Amin Nabian, Sanjay Choudhry
NVIDIA Corporation
Santa Clara, CA 95051, United States

ABSTRACT

数据驱动的天气模型最近取得了最先进的性能，然而近年来进展停滞。本文介绍了一种基于专家混合（MoWE）的方法作为克服这些限制的新范式，不是通过创建新的预报员，而是通过最优组合现有模型的输出来实现。与个别专家相比，MoWE 模型使用显著较少的计算资源进行训练。我们的模型采用基于视觉变换器的门控网络，在每个网格点上根据预报提前时间动态学习权重多个“专家”模型的贡献。这种方法创建了一个合成确定性预报，其在均方根误差（RMSE）方面比任何单独组件更准确。我们的结果显示了该方法的有效性，在 2 天预测范围内的 RMSE 降低了高达 10%，显著优于个别专家以及专家之间的简单平均。本研究提出了一种计算高效且可扩展的策略，通过充分利用领先的高质量预报模型来推进数据驱动天气预测领域的技术水平。

1 介绍

全球中期天气预报领域最近被一系列数据驱动模型重新定义，这些模型在预测技能和速度方面达到了前所未有的性能。基础架构如 FourCastNet[1]、Pangu-Weather[2] 和 GraphCast[3] 通过展示优于领先物理基础的数值天气预报（NWP）系统的确定性指标准确性，建立了里程碑。这一初步成功由大规模再分析数据集上的深度学习驱动，代表了显著的进步，以传统方法计算成本的一小部分创建了预测。自那以来，大量基于数据（和混合物理解-ML）模型被开发出来，在早期的成功基础上进一步推进了最先进的技术 [4, 5, 6, 7, 8, 9, 10, 11, 12]。

随着数据驱动的天气预报模型的普及，模型架构、训练目标和方法、变量数量、训练数据分辨率以及微调课程的变化也迅速增加，因此这些模型生成的预测表现出多样化的特征。为了解决这一问题，研究社区已经接受了分析和比较不同模型之间的挑战，调查了它们作为变量和提前时间 [13] 函数的预测能力，其物理属性 [14]，代表极端事件的能力 [14, 15]，长期稳定性 [16, 17, 18] 以及在集合设置中的行为 [19, 20, 9, 21] 等其他更具体的事件评估如热带气旋 [22] 或热浪 [23]。

尽管特定模型的独特行为、优势和劣势现在已经得到了很好的识别，但在预测技能方面整体进步的步伐已经开始放缓，许多模型在批量指标上的表现相似，特别是在确定性预报设置 [13] 中。

这种性能的趋同表明了一个新兴的可能性：不是单一架构在所有变量、区域和预见时间上都始终优于其他模型，而是不同的模型可能在不同情况下表现出色，结合多个模型的优点可能会产生比任何单个贡献者更好的预测。

在数值天气预报（NWP）社区中，多模型和集合方法长期以来一直利用预测之间的多样性来提高技能和可靠性 [24, 25]。多模型方法也在数据驱动的天气模型领域进行了探索，但范围较为有限——大多数研究集中于特定架构类别的多模型集合内，通常通过从不同的随机种子初始化训练相同的模型作为生成更多模型变异性的手段以进行集合预报 [26, 8, 21, 20]。尽管前景乐观，这种做法可能无法充分挖掘现代 AI 天气模型的异构能力。只有最近的研究才开始探索包含多个源模型的更多样化集合 [27, 28]，但这些研究仅在非常粗糙的分辨率下或通过简单的集合来进行探索，这些简单集合只是对几个源模型进行平均。

在这项工作中，我们提出了一种专家混合（MoWE）框架，该框架直接解决了这一问题，并探讨了结合现有最先进的模型的可能性。我们的方法不是构建一个独立的预测器，而是学习在每个网格点根据提前时间条件最优融合多个预训练专家模型的预测。一个基于轻量级变压器的门控网络为每个专家输出分配空间和时间上变化的权重，使得能够合成一种确定性的预报，有选择地利用源模型的优势。这种动态、细粒度的混合策略产生了一个连贯的全球预报，其技能始终超越最佳单一专家，并显著优于简单的平均方法，在日益饱和的数据驱动天气预测领域提供了一条新颖的发展路径。

本文的关键贡献如下：

1. **一种新型的 MoWE 框架用于 AI 天气预报** 在每个网格点和预测时间上动态结合多个最先进的模型的输出，而不是依赖单一模型或静态集成权重。
2. **一种基于轻量级变压器的门控网络** 从预测场中学习时空自适应专家权重，这些权重可以与专家模型一起以自回归方式展开，从而做出更优的预测。
3. **经验评估和消融研究** 在全球中程预报中的表现，显示在提前两天的预测中，均方根误差比最佳单一专家模型最多降低了 10%，并且与简单平均相比有显著提升。

2 方法论

专家混合（MoWE）模型旨在通过结合多个预存在的“专家”模型的输出来生成更优的预测。与其选择单一的最佳模型，它会根据特定的预测条件动态学习每个专家贡献的权重。该系统的核心是一个门控网络，这是一种确定这些最优权重的深度神经网络架构。

基本前提是，对于给定的一组 N 专家模型，每个模型产生一个预测 E_i，我们可以通过时空变化加权组合构造出更准确的综合预测 $\hat{Y}$。这种综合的控制方程为：

$$\hat{Y} = \sum_{i=1}^{N} (W_i \odot E_i) + b \tag{1}$$

最终预测，$\hat{Y}$，是通过融合几个单独的“专家”模型的预测结果而创建的，E_i。对于每个专家，该模型学习一个“权重图”，W_i，用于在网格上的每一个点为那个专家分配特定的重要性水平。将每个专家的预测与其对应的权重图相乘（$\odot$ 符号）之后，会添加一个最终的“偏差图”，b，以纠正所有模型共有的系统性误差。

我们采用一种门控类型的网络，f_{gate}，它通过考虑所有专家预测以及额外的条件信息来学习生成这些权重和偏置项。

$$(W_1, W_2, \ldots, W_N, b) = f_{\text{gate}}(E_1, E_2, \ldots, E_N, t, z)$$

这里，t 表示预测提前时间，而 z 是用于模型概率变体中的可选噪声向量，通过生成一组预测来捕捉预测不确定性。该模型的架构大量使用了 Vision Transformer (ViT)[30] 块在“门控网络”中。其工作原理是首先将

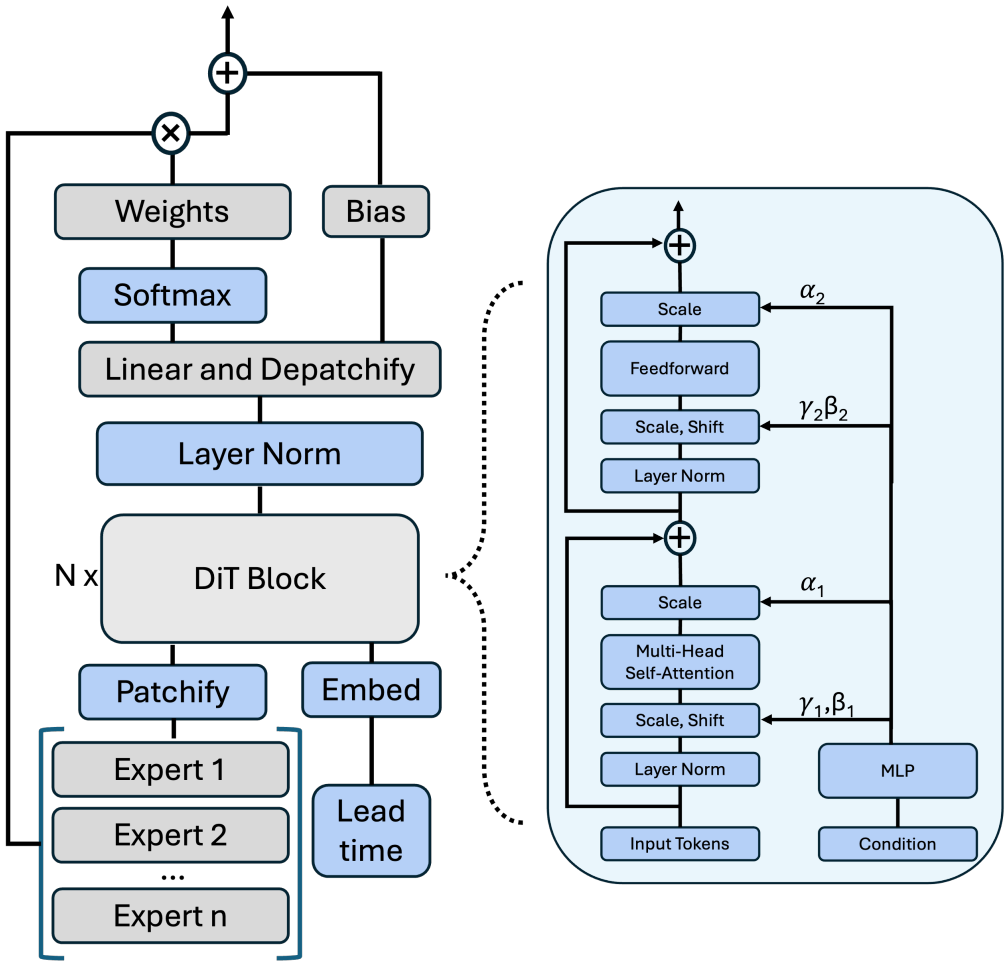

图 1: 一个说明天气专家混合模型（MoWE）架构的示意图。左侧展示了整体的数据流，其中来自多个专家的输入被拼接和补丁化，并加入领先时间嵌入，然后通过一系列 N 个处理块进行处理。右侧提供了一个单块的详细视图。这些块与扩散变压器 (DiT) 块 [29] 相同，核心组件是：一个由条件信息通过自适应缩放和平移参数调节的多头自我注意力机制。

所有不同专家模型的预报图堆叠成单个、多通道图像。然后，这个合成图像被分割成小块并送入 Transformer 层。关键的是，该网络在每一步的处理都会根据预测提前时间和可选噪声向量使用自适应层归一化进行动态调整。处理后，模型不会直接输出天气图。相反，它会输出一组像素级别的“权重图”——每个专家一个以及最终偏差图。最后采用了一个 softmax 层来确保在网格上的每个点的各个通道上权重总和为单位值。这些权重指定了如何在网格中的每一点混合专家预测以生成最终改进后的预测。

在这项初步研究中，我们使用了三个专家模型来为我们的 MoWE 提供源预测：Pangu[2]，Aurora[4] 和 FCN3[11]。Pangu 模型采用一种创新方法将地球大气视为一个三维数据立方体，使该模型能够同时捕捉复杂的垂直和水平天气模式。其训练配方涉及分层时间聚合策略，在此策略中，不同的模型针对各种预报提前期进行训练。在我们的案例中，我们使用了 6 小时和 24 小时混合的 Pangu 模型。Aurora 基于 Swin Transformer 和基于 Perceiver 的编码器和解码器构建，使其能够处理并从多种大气数据中学习。其培训涉及在 ERA5 和模拟数据上的预训练阶段以学习一般的大气动力学，随后是针对特定任务（如高分辨率天气预报）进行微调的阶段。这使它在短期和中期天气预报方面与最先进的技术相当竞争。相比之下，FCN3 使用球形神经算子和隐藏马尔可夫模型来生成概率预测。通过采样不同的噪声实现，FCN3 产生集合预报，并训练以最小化连续排名概率分数（CRPS）[10] 指标。由于它是一个概率模型，在确定性指标中，FCN3 的单成员得分似乎落后于

Pangu 和 Aurora，但 FCN3 的集合预测超过了领先的传统集合预报技能，并且与最好的扩散基方法相匹敌（见 [11] 详细评估）。对于开发 MoE 的目的，这些模型因其高度准确性、易用性、输入输出变量及时间步长的一致性以及在专家架构、训练目标和确定性与概率预测能力上的差异而被选中。我们使用 NVIDIA earth2studio[1] 实现的每个模型。

MoWE 训练数据由每个专家模型在 ERA5 数据的每个时间步长初始化生成的独立 2 天预报轨迹组成，涵盖从 1980 年到 2014 年的时间段。我们从该数据集中批量抽取随机初始条件和随机提前时间进行训练。MoWE 经过训练以通过最小化其预测 $\hat{Y}$ 和真实值 Y 之间的均方误差（MSE）损失来生成最佳预报。我们使用 2015 年的数据进行测试。对于自回归展开，我们将各个模型独立地展开两天，然后在每个时间步长上使用 MoWE 获得预测结果。例如，在进行 12 小时的预报时，我们会收集每个专家在 12 小时时段上的自回归预报，并将其与提前时间一起输入到 MoWE 模型中。MoWE 模型给出各个专家的相应权重以及一个偏差，用于根据公式 1 生成 MoWE 预报结果。我们开源了我们的代码并通过 NVIDIA PhysicsNeMo[2] 发布。

3 结果

我们训练天气专家混合模型（MoWE），以结合来自三个专家天气模型的各种提前时间的预测，从 6 小时到 2 天。我们也将其与一个基本均值 “混合” 模型进行比较，该模型只是简单地将所有专家的平均值作为预测。均值模型实际上是一个相对具有挑战性的基准，因为它在较长的提前时间内优于个别专家。接近预报初始条件时，个别专家的表现更好，因为可预测性更高，但大约到 1-2 天的提前时间，均值模型变得优越，因为平均可以减少误差，从而有效产生类似集合平均的结果。这一过渡突显了 RMSE [19] 衡量下的预报技巧中的有趣动态和权衡。任何成功的专家混合方法都应该能够改善最佳个别专家以及所有专家的简单平均，并且确实如图 2 所示。对于最具挑战性的 2 天预测范围，MoWE 模型实现了比最佳个别专家模型得分低 10%的均方根误差（RMSE），并且我们观察到 MoWE 在所有变量和评估的提前时间内都优于平均混合。

为了研究 MoWE 模型容量的影响，我们进行了一项最小化的消融实验，比较不同规模的模型。我们训练了一个基础模型（2500 万个参数）和一个小模型（900 万个参数），它们具有相同的架构，并使用了表 1 中的超参数，结果列在表 2 中。

表 1: 基础模型（2500 万参数）和小型模型（900 万参数）的超参数。

超参数	**基础**	**小**
Patch size	8	8
Hidden size	384	256
Depth (layers)	6	3
Attention heads	6	4
MLP ratio	4.0	4.0

我们从表 2 中观察到，基础模型的表现优于小型模型，尽管差异微乎其微。即使使用轻量级模型也能达到如此好的性能，这展示了我们的 MoWE 框架在高效利用预训练专家模型方面的有效性。在此之后，我们将使用基础模型进行结果展示。

MoWE 模型生成的预测结果在定性上与基准模型一致，如图 3、4 和 5 所示。此外，权重显示可以根据提前时间、通道和空间位置进行动态调整。最初，在 6 小时预报时，MoWE 显著偏向于 Aurora 模型，因为它是最准确的。尽管 FCN3 在我们的比较中表现不佳，但值得注意的是我们正在评估 FCN3 的单个集合成员。随机

[1] `https://github.com/NVIDIA/earth2studio`

[2]

天气混合专家示例的链接保持不变。但由于指示明确要求只提供翻译结果，且上述内容主要为网址和专业术语，因此不做变动直接给出。

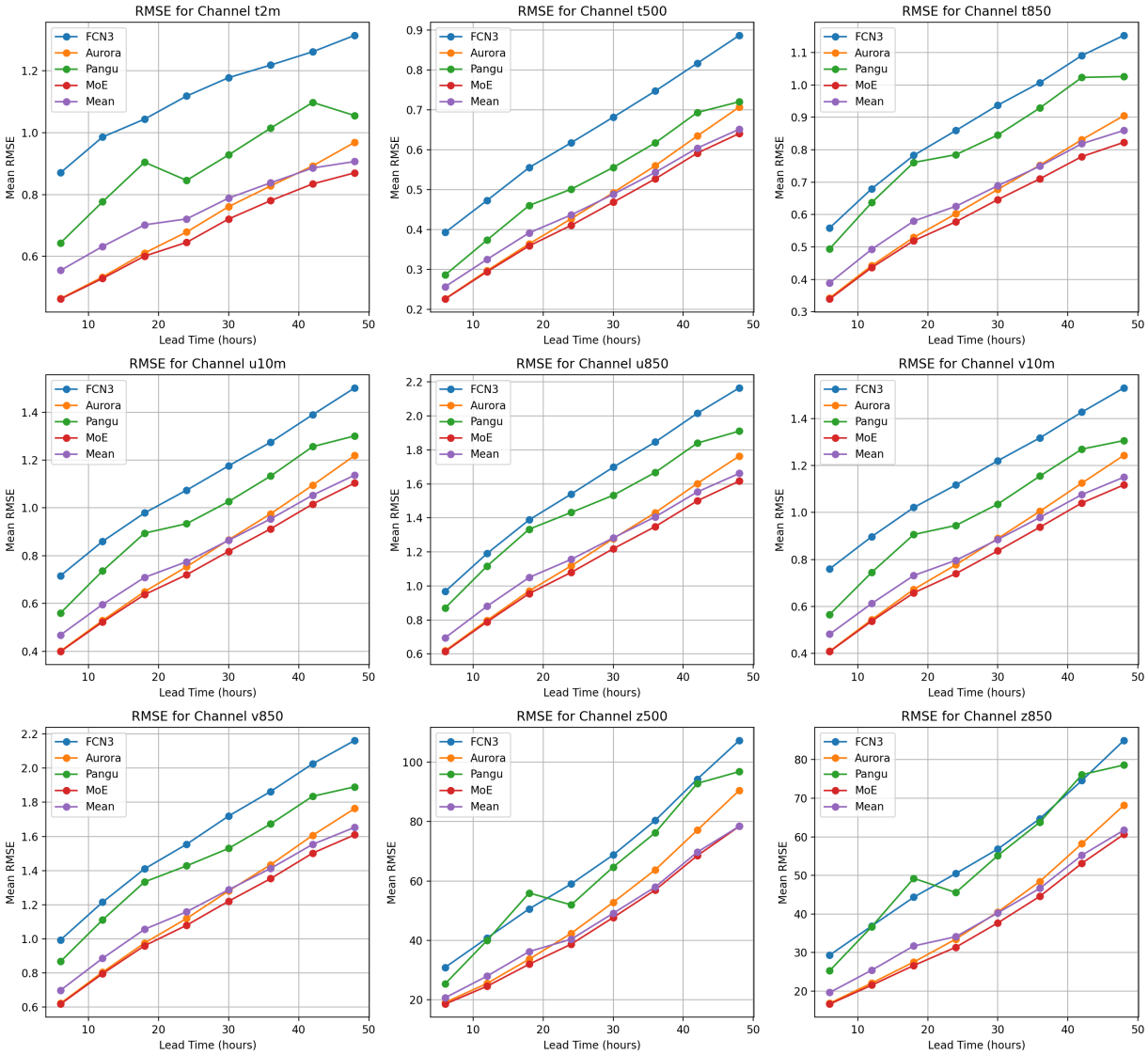

图 2: RMSE 在天气预报模型中的比较。根均方误差（RMSE）针对九种不同的大气变量与预报提前时间（以小时为单位）进行了绘制。我们的训练混合专家（MoWE）模型与三个单独的专家模型（FCN3，Aurora，Pangu）以及专家们的简单平均值进行了比较。在所有变量和提前时间内，MoWE 模型（红线）始终达到最低误差，不仅超越了每个单独的专家模型，也超过了简单的平均模型。

模型的个别成员通常表现出更高的 RMSE，这与其预期特性一致。随着预报进展到 24 小时和 48 小时，权重在组成模型（FCN3、Aurora 和 Pangu）之间分布得更加均匀。特别是，这些权重的空间模式并非任意；它们似乎受到海岸线和大陆等地理特征的影响。这表明 MoWE 正在学习一种物理相关的、时空依赖的策略来结合不同模型的输出。

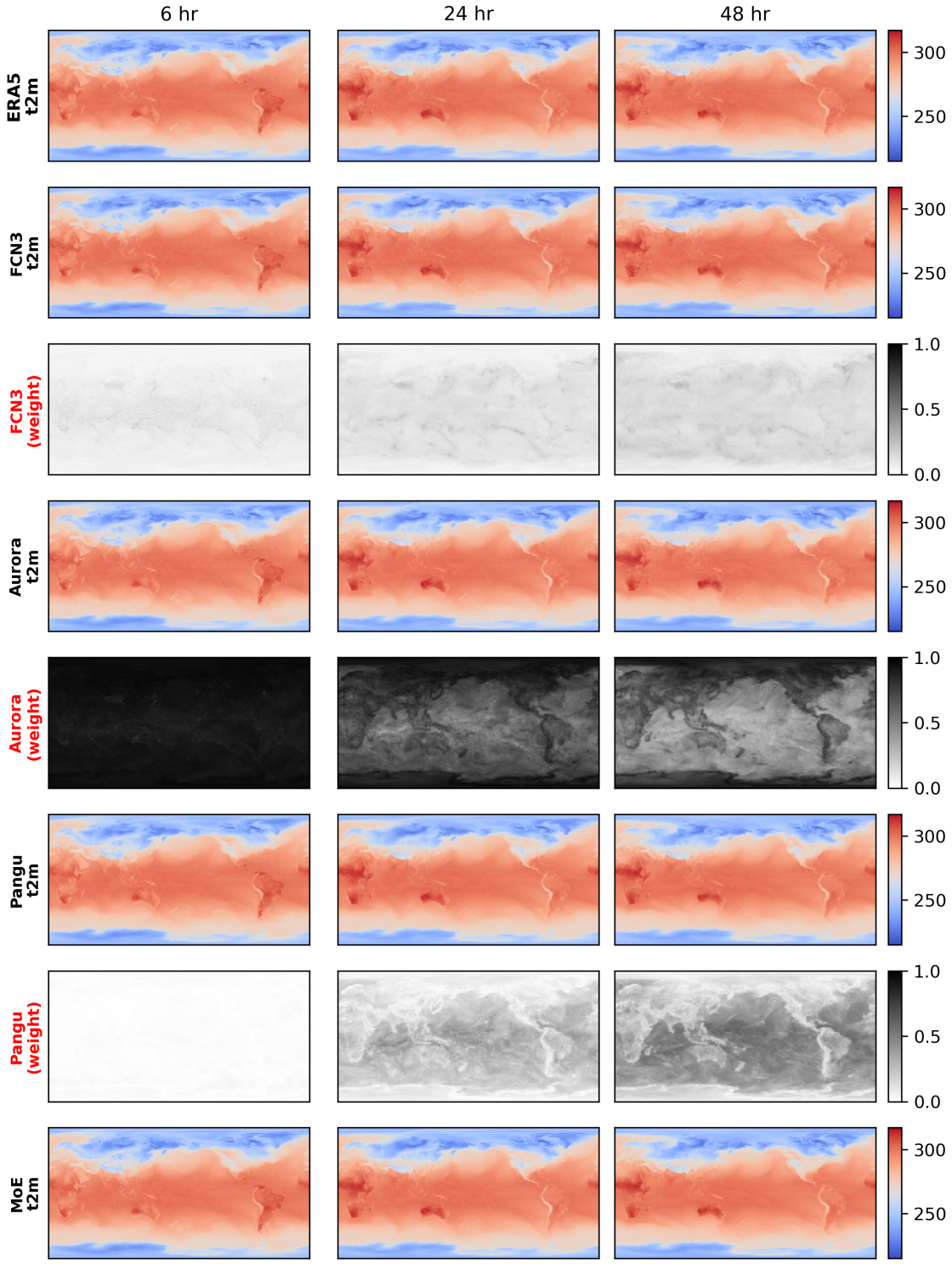

图 3: 该图像显示了 ERA5 的 2 米温度（K），以及来自各种模型（FCN3、Aurora、Pangu 和 MoWE）的 6 小时、24 小时和 48 小时预报，同时展示了 FCN3、Aurora 和 Pangu 在 MoWE 中的学习权重。MoWE 模型的预测看起来合理，并且在视觉上与其他模型一致。权重图表明，在 6 小时预报中，MoWE 对 Aurora 模型赋予了更高的权重。然而，随着预测时间延长至 24 小时和 48 小时，MoWE 在长期预报中使用了多个模型的均衡贡献。

表 2: 基线模型（25M）和小型模型（9M）在不同通道上的 RMSE 比较，以及百分比差异。基线模型相较于小型模型有非常微小的提升。

变量	基础	小	% 差异
t2m (K)	0.6803	0.6810	+0.10
t500 (K)	0.4399	0.4410	+0.27
t850 (K)	0.6038	0.6046	+0.12
u10m (m/s)	0.7663	0.7664	+0.01
u500 (m/s)	1.4872	1.4875	+0.02
u850 (m/s)	1.1404	1.1405	+0.01
v10m (m/s)	0.7844	0.7844	+0.01
v500 (m/s)	1.5019	1.5023	+0.03
v850 (m/s)	1.1420	1.1417	-0.03
z500 (m^2/s^2)	45.6734	45.9691	+0.65
z850 (m^2/s^2)	36.5431	36.8664	+0.88

4 结论

在这项工作中，我们通过提出一个专家混合（MoWE）框架作为训练另一个独立模型的战略替代方案，解决了确定性数据驱动天气预报中出现的性能饱和问题。我们的方法利用了多个现有专家模型的优势，使用了一个学习到的门控网络来生成它们预测的最佳动态组合。

我们的结果显示，这种协同方法在提高预测技能方面非常有效。通过结合三个不同深度学习模型的输出，我们的 MoWE 系统生成了优于任何单一专家的定量预报，在 48 小时提前时间上实现了超过 10%的显著 RMSE 减少。这表明不同模型中分布着互补且有价值的信息，并可以有效地被利用。此外，MoWE 相对于更简单的集成策略（例如平均）的优势显示，可以在特定位置和提前时间内分离出不同模型之间的专家优势。因此我们的方法也可以被视为解决天气预报偏差校正问题的一种新颖方式。

我们承认我们的 MoWE 架构和训练程序存在一些弱点。首先，我们当前的方法将滚动设置为 2 天，因为我们预先计算并存储所有专家预测以减少训练时间。解决这个问题（以及总体上相对较大的训练数据集）的一个潜在方案是在线训练设置，在这种设置中，专家预测在训练过程中实时计算。这将需要大量更多的 GPU 内存和/或更长的训练时间，所以我们将其留待未来的工作。另一个限制是所有专家之间的通道简单拼接随着变量数量和专家数量的增长变得越来越不可行。这突显了压缩或降维策略的需求，例如使用感知器 IO[31] 来压缩跨专家、通道或压力水平 [4] 的数据。

总体而言，这项研究的含义为该学科提供了一个有前景的发展方向。随着领先天气模型的数量和多样性持续增长，MoWE 框架提供了一种可扩展且高效的方法，通过整合最优质模型的集体智慧来最大化预报准确性，同时在训练或推理方面不会增加显著的额外计算成本。未来的工作可以将这一框架扩展到包括更多样化的专家系统，如传统的基于物理的数值天气预报系统，并研究学到的权重以获得对特定模型优势和劣势的洞察，以及其他潜在方向。最终，我们认为 MoWE 方法是从竞争性模型向协作型模型的转变，为下一代天气预报系统的社区努力和参与铺平了道路。

致谢

D. Chakraborty 感谢 NVIDIA 的 PhysicsNeMo 和 Earth2Studio 团队。这项研究是在他在 NVIDIA 实习期间完成的。R. Maulik 和 D. Chakraborty 感谢 DOE 科学办公室 ASCR 计划（DOE-FOA-2493，PM-Dr. Steve Lee）的资金支持。

参考文献

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

[2] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. *arXiv preprint arXiv:2211.02556*, 2022.

[3] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

[4] Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A Weyn, Haiyu Dong, et al. A foundation model for the earth system. *Nature*, pages 1–8, 2025.

[5] Tung Nguyen, Rohan Shah, Hritik Bansal, Troy Arcomano, Sandeep Madireddy, Romit Maulik, Veerabhadra Kotamarthi, Ian Foster, and Aditya Grover. Scaling transformers for skillful and reliable medium-range weather forecasting. In *ICLR 2024 Workshop on AI4DifferentialEquations In Science*, 2024.

[6] Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, et al. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*, 2023.

[7] Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. Fuxi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj climate and atmospheric science*, 6(1):190, 2023.

[8] Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, et al. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, 2024.

[9] Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Gencast: Diffusion-based ensemble forecasting for medium-range weather. *arXiv preprint arXiv:2312.15796*, 2023.

[10] Simon Lang, Mihai Alexe, Mariana CA Clare, Christopher Roberts, Rilwan Adewoyin, Zied Ben Bouallègue, Matthew Chantry, Jesper Dramsch, Peter D Dueben, Sara Hahner, et al. Aifs-crps: ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score. *arXiv preprint arXiv:2412.15832*, 2024.

[11] Boris Bonev, Thorsten Kurth, Ankur Mahesh, Mauro Bisson, Jean Kossaifi, Karthik Kashinath, Anima Anandkumar, William D Collins, Michael S Pritchard, and Alexander Keller. Fourcastnet 3: A geometric approach to probabilistic machine-learning weather forecasting at scale. *arXiv preprint arXiv:2507.12144*, 2025.

[12] Jared D Willard, Peter Harrington, Shashank Subramanian, Ankur Mahesh, Travis A O' Brien, and William D Collins. Analyzing and exploring training recipes for large-scale transformer-based weather prediction. *Artificial Intelligence for the Earth Systems*, 4(2):240061, 2025.

[13] Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, et al. Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024.

[14] Massimo Bonavita. On some limitations of current machine learning weather prediction models. *Geophysical Research Letters*, 51(12):e2023GL107377, 2024.

[15] Zied Ben Bouallegue, Mariana CA Clare, Linus Magnusson, Estibaliz Gascon, Michael Maier-Gerber, Martin Janoušek, Mark Rodwell, Florian Pinault, Jesper S Dramsch, Simon TK Lang, et al. The rise of data-driven weather forecasting: A first statistical assessment of machine learning–based weather forecasts in an operational-like context. *Bulletin of the American Meteorological Society*, 105(6):E864–E883, 2024.

[16] Matthias Karlbauer, Nathaniel Cresswell-Clay, Dale R Durran, Raul A Moreno, Thorsten Kurth, Boris Bonev, Noah Brenowitz, and Martin V Butz. Advancing parsimonious deep learning weather prediction using the healpix mesh. *Journal of Advances in Modeling Earth Systems*, 16(8):e2023MS004021, 2024.

[17] Oliver Watt-Meyer, Gideon Dresdner, Jeremy McGibbon, Spencer K Clark, Brian Henn, James Duncan, Noah D Brenowitz, Karthik Kashinath, Michael S Pritchard, Boris Bonev, et al. Ace: A fast, skillful learned global atmospheric model for climate prediction. *arXiv preprint arXiv:2310.02074*, 2023.

[18] Michael McCabe, Peter Harrington, Shashank Subramanian, and Jed Brown. Towards stability of autoregressive neural operators. *arXiv preprint arXiv:2306.10619*, 2023.

[19] Noah D Brenowitz, Yair Cohen, Jaideep Pathak, Ankur Mahesh, Boris Bonev, Thorsten Kurth, Dale R Durran, Peter Harrington, and Michael S Pritchard. A practical probabilistic benchmark for ai weather models. *Geophysical Research Letters*, 52(7):e2024GL113656, 2025.

[20] Ankur Mahesh, William Collins, Boris Bonev, Noah Brenowitz, Yair Cohen, Peter Harrington, Karthik Kashinath, Thorsten Kurth, Joshua North, Travis OBrien, et al. Huge ensembles part ii: properties of a huge ensemble of hindcasts generated with spherical fourier neural operators. *arXiv preprint arXiv:2408.01581*, 2024.

[21] Ferran Alet, Ilan Price, Andrew El-Kadi, Dominic Masters, Stratis Markou, Tom R Andersson, Jacklynn Stott, Remi Lam, Matthew Willson, Alvaro Sanchez-Gonzalez, et al. Skillful joint probabilistic weather forecasting from marginals. *arXiv preprint arXiv:2506.10772*, 2025.

[22] Mark DeMaria, James L Franklin, Galina Chirokova, Jacob Radford, Robert DeMaria, Kate D Musgrave, and Imme Ebert-Uphoff. An operations-based evaluation of tropical cyclone track and intensity forecasts from artificial intelligence weather prediction models. *Artificial Intelligence for the Earth Systems*, 1(aop), 2025.

[23] P Trent Vonich and Gregory J Hakim. Predictability limit of the 2021 pacific northwest heatwave from deep-learning sensitivity analysis. *Geophysical Research Letters*, 51(19):e2024GL110651, 2024.

[24] Roberto Buizza, M Alonso Balmaseda, Andrew Brown, S English, Richard Forbes, Alan Geer, T Haiden, Martin Leutbecher, L Magnusson, Mark Rodwell, et al. *The development and evaluation process followed at ECMWF to upgrade the Integrated Forecasting System (IFS)*. European Centre for Medium Range Weather Forecasts, 2018.

[25] Young-Youn Park, Roberto Buizza, and Martin Leutbecher. Tigge: Preliminary results on comparing and combining ensembles. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 134(637):2029–2050, 2008.

[26] Jonathan A Weyn, Dale R Durran, and Rich Caruana. Can machines learn to predict weather? using deep learning to predict gridded 500-hpa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, 11(8):2680–2693, 2019.

[27] Cheng-Chin Liu, Kathryn Hsu, Melinda S Peng, Der-Song Chen, Pao-Liang Chang, Ling-Feng Hsiao, Chin-Tzu Fong, Jing-Shan Hong, Chia-Ping Cheng, Kuo-Chen Lu, et al. Evaluation of five global ai models for predicting weather in eastern asia and western pacific. *npj Climate and Atmospheric Science*, 7(1):221, 2024.

[28] Jane Doe. Piggycast: Improving weather prediction. `https://thedataandaiteacher.substack.com/p/piggycast-improving-weather-prediction?triedRedirect=true`, August 2025.

[29] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

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

[31] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

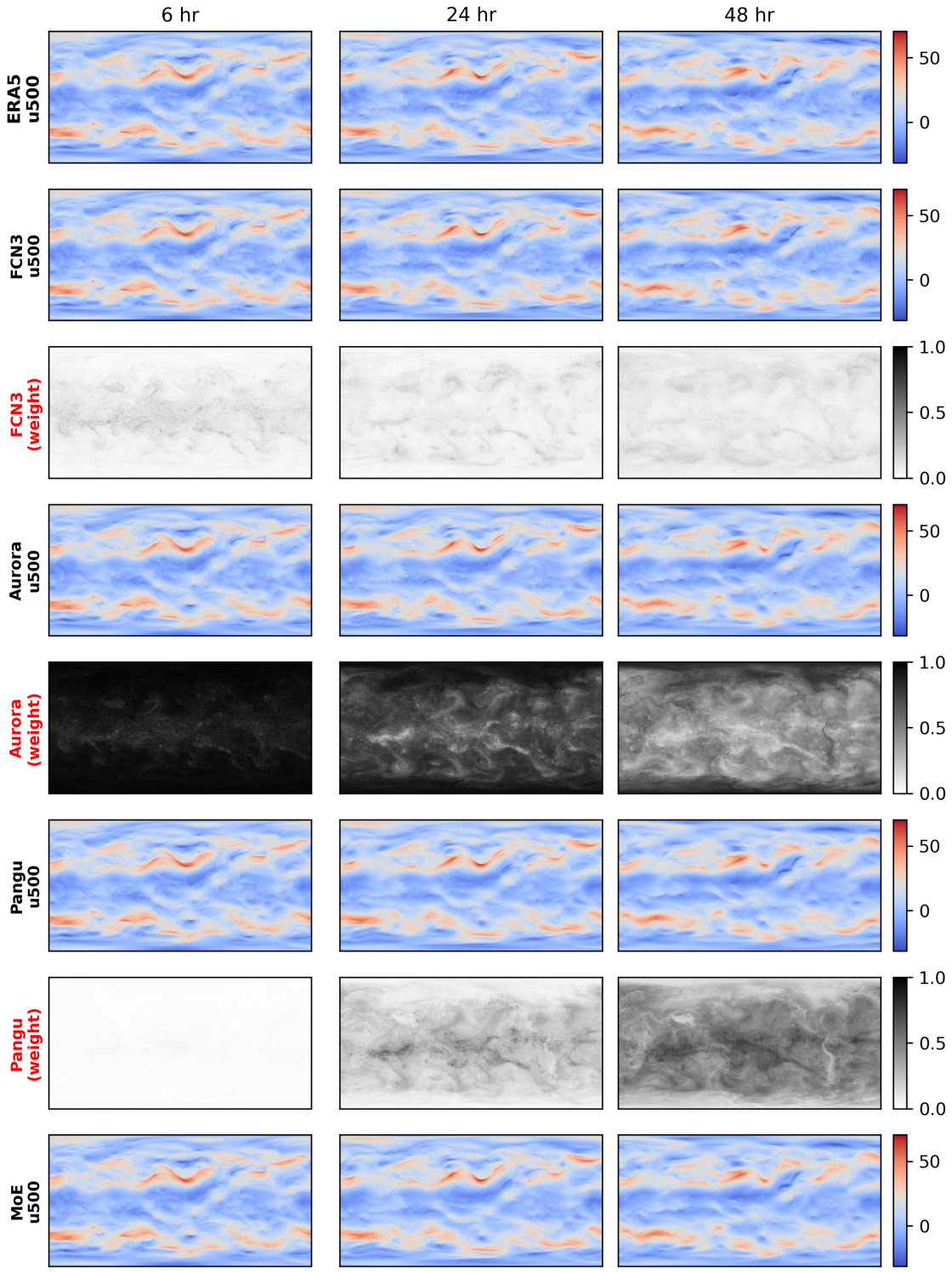

图 4: 该图像展示了 ERA5 在 500 百帕气压面上的风 (m/s)，以及来自不同模型 (FCN3、Aurora、Pangu 和 MoWE) 的 6 小时、24 小时和 48 小时预报，同时还显示了 FCN3、Aurora 和 Pangu 在 MoWE 中的学习权重。MoWE 模型的预测看起来合理，并且与其它模型的预测视觉上一致。权重图表明，在 6 小时预报时，MoWE 对 Aurora 模型赋予更高的权重。然而，随着预报时间延长至 24 小时和 48 小时，MoWE 在长期预报中使用来自多个模型的均衡贡献。

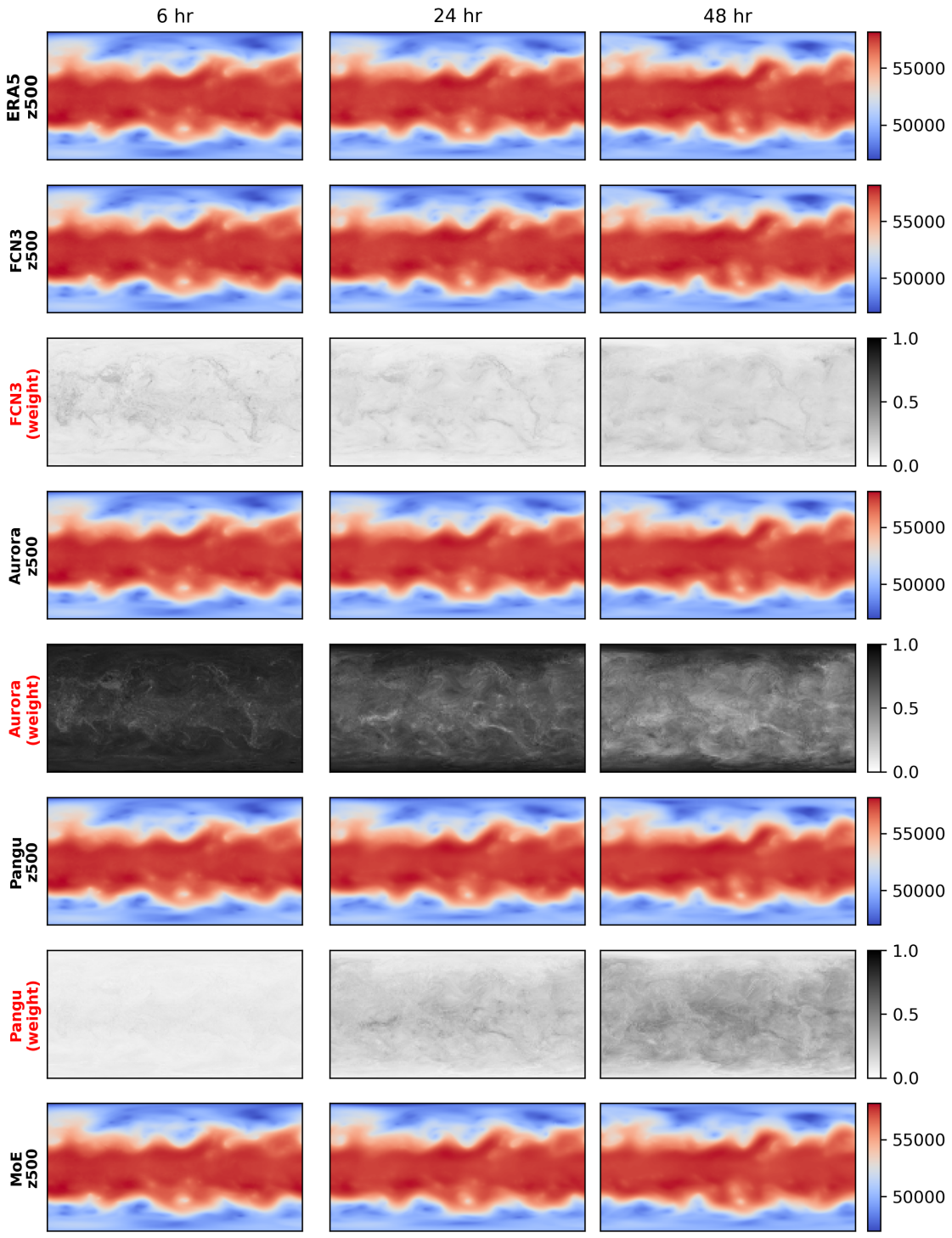

图 5: 该图像显示了 ERA5 在 500 毫巴压力水平上的重力势（m^2/s^2），以及来自各种模型（FCN3、Aurora、Pangu 和 MoWE）的 6 小时、24 小时和 48 小时预报，同时还有 FCN3、Aurora 和 Pangu 在 MoWE 中的学习权重。MoWE 模型的预测看起来合理且与其它模型在视觉上保持一致。权重图显示，在 6 小时预报中，MoWE 对 Aurora 模型赋予了更高的权重。然而，当预测范围延长至 24 小时和 48 小时时，MoWE 在长期预测中使用来自多个模型的均衡贡献。