瑞利-贝纳德对流的控制：强化学习在湍流区域的有效性

Thorben Markmann[1[0000−0003−3145−4577]], Michiel Straat[1[0000−0002−3832−978X]], Sebastian Peitz[2[0000−0002−3389−793X]], and Barbara Hammer[1[0000−0002−0935−5591]]

[1] Bielefeld University, Center for Cognitive Interaction Technology (CITEC), Inspiration 1, 33619 Bielefeld, Germany
{tmarkmann,mstraat,bhammer}@techfak.uni-bielefeld.de

[2] Department of Computer Science & Lamarr Institute for Machine Learning and Artificial Intelligence, TU Dortmund University, Joseph-von-Fraunhofer-StraSSe 25, 44227 Dortmund, Germany
sebastian.peitz@tu-dortmund.de

摘要 数据驱动的流体控制在工业、能源系统和气候科学中具有重要潜力。在这项工作中，我们研究了强化学习（RL）在减少二维瑞利-本纳德对流（RBC）系统中的对流传热方面的有效性，特别是在增加湍流的情况下。我们调查了不同初始条件和湍流水平下的控制泛化能力，并引入了一种奖励塑形技术以加速训练过程。通过单代理近端策略优化（PPO）训练的 RL 代理与经典控制理论中的线性比例微分（PD）控制器进行了比较。在中等湍流系统中，RL 代理将对流传热减少至多 33%，而在高度湍流环境中减少了 10%，明显优于所有设置下的 PD 控制。这些代理显示了强大的泛化性能，不仅在不同的初始条件下表现出色，在相当大的程度上还能够推广到更高水平的湍流中。奖励塑形技术提高了样本效率，并稳定地将努塞尔数维持在更高的湍流水平。

Keywords: 强化学习 · 流体动力学 · 瑞利-贝纳德对流 · 流动控制

1 介绍

近年来，深度强化学习（DRL）在流体力学中的控制任务中展示了巨大的潜力，包括湍流抑制、最优混合和阻力减小 [2]。DRL 的一个关键优势在于它能够在高度非线性系统中发现有效的控制策略，而对于这类系统，传统的方法如 PID 控制往往证明是不够的。

在这项工作中，我们研究了无模型深度强化学习在控制由瑞利-贝纳德对流（RBC）支配的自然对流动力学的有效性。对流在自然和工业过程中都起着至关重要的作用，包括海洋环流、云形成、恒星动力学和材料加工 [1,13]。在工业中，控制对流的方法尤其重要；例如，在晶体生长过程中，过度的对流会导致影响材料质量的不稳定性。RBC 系统模型描述了一种被加热下板和冷却上板夹住的流体，由于底部加热引起的密度差异，在整个流体内出现了由浮力驱动的流动。这种对流流动随着温度梯度增大和流体粘度降低 [7] 而变得更强且更加湍流。我们特别探讨了使用近端策略优化（PPO）的强化学习在湍流条件下减少对流的有效性，并将其性能与传统的 PD 控制进行比较。

流体动力学中 RL 应用的一个中心挑战是泛化。在 RBC 系统中，初始温度和速度场的微小变化可能会产生显著不同的流动结构。因此，实用的 RL 代理必须对初始条件的变化表现出鲁棒性，并且理想情况下，代理可以在不需要重新训练的情况下在不同湍流状态下进行泛化。另一个关键挑战是样本效率。通常训练高性能 RL 代理需要大量的迭代，特别是在使用无模型方法结合高保真流体模拟时。减轻样本需求的方法不仅提高了训练效率，还增强了 RL 代理对新场景的适应性。

我们的主要贡献是通过 RL 探索 RBC 控制任务，解决了以下挑战：1) **不断增加的湍流条件下的性能**：我们评估了 PPO 在减少不同湍流水平的 RBC 系统中的对流运动方面的有效性，并从定量和定性两个方面将其性能与 PD 控制进行了比较。2) **泛化能力**：我们在不同的初始条件下训练 RL 代理，并评估它们在未见过的初始状态和不同湍流环境下的泛化能力。3) **训练效率**：我们介绍并演示了一种奖励塑造技术，该技术加速了训练过程并提高了代理的最终性能。

2 相关工作

控制 RBC 以减少或避免对流的研究已有几十年，最初使用传统控制理论的方法，后来则采用了机器学习的数据驱动方法。直到最近，像比例（P）或比例微分（PD）这样的线性控制器被广泛用于稳定 RBC 系统并减少对流。在 20 世纪 90 年代初，Tang 和 Bau [13] 奠定了通过线性反馈控制来稳定 RBC 系统的理论基础。他们的工作表明，从无运动状态开始，一个与系统中线温度成比例的控制输入可以显著延迟对流热传递的发生。后来，Howle 通过一系列实验展示了在实际设置中通过在网络加热器放置于系统底部 [3] 可

以稳定 RBC 的可能性。他没有依赖中线温度，而是使用阴影图测量来测量垂直平均密度场进行线性比例控制。到了 2000 年代，Remillieux 等人进一步研究了使用 PD 控制器抑制 RBC 中的对流，在实验和数值仿真设置中展示了其有效性 [9]。

2020 年，Beintema 等人 [1] 引入了基于强化学习 (RL) 的 RBC 控制方法，其性能超过了早期基于 PD 控制的方法。他们的方法将 RBC 稳定到 $\mathrm{Ra} = 3 * 10^4$，并通过努塞尔数测量，在雷诺数以上实现了更大的对流减少。Vignon 等人 [16] 扩展了这项工作，提出了多智能体 RL(MARL) 方法以提高样本效率。该方法利用了个别加热器作为代理的平移不变性，实现了在水平方向上一个接近工业应用的广泛不受限区域内的努塞尔数减少了 22.7%，对于 $\mathrm{Ra} = 10^4$。进一步的研究调查了将 MARL 方法扩展到 3D-RBC [15]，并包括加热器的位置编码 [4]。

虽然 [16] 在中等湍流环境下采用了一种高度可扩展的多智能体强化学习框架，本工作考察了在更高湍流程度（即更大的 Ra）下更具表达能力的单个智能体强化学习代理的表现。我们采用了与 [16] 相同的系统参数、驱动机制和对流度量。此外，我们引入了奖励塑形以提高样本效率，并通过特别关注强化学习代理在初始条件和湍流程度上的泛化能力来增加其实用可行性。

3 方法论

在本节中，我们介绍了 RBC 系统，描述了仿真环境，并概述了控制任务和方法。

3.1 瑞利-贝纳德对流

瑞利-贝纳德对流（RBC）是一个模拟从下方加热的流体层中的传导和对流传热的系统。动力学由基于不可压缩 Navier-Stokes 方程的部分微分方程（PDE）所支配，例如可以参见 [7]。该系统的状态在二维情况下可以通过速度矢量场 $\boldsymbol{u} = (u_x, u_y)$、标量温度场 T，以及其初始和边界条件来完全描述（见第 3.1节）。瑞利数 Ra 是一个关键系统参数，量化了浮力驱动对流的强度。它与底面和顶面之间的温度梯度成正比，参见 [7]。对于增加的雷诺数 Ra，流动变得更加不稳定，导致对流湍流程度升高 [16]。

对流在 RBC 系统中的强度通过由方程给出的局部对流热流来测量

$$q(x, y, t) = u_y(x, y, t)\theta(x, y, t)\,, \text{ with } \theta(x, y, t) = T(x, y, t) - \langle T \rangle_{x,y}\,, \quad (1)$$

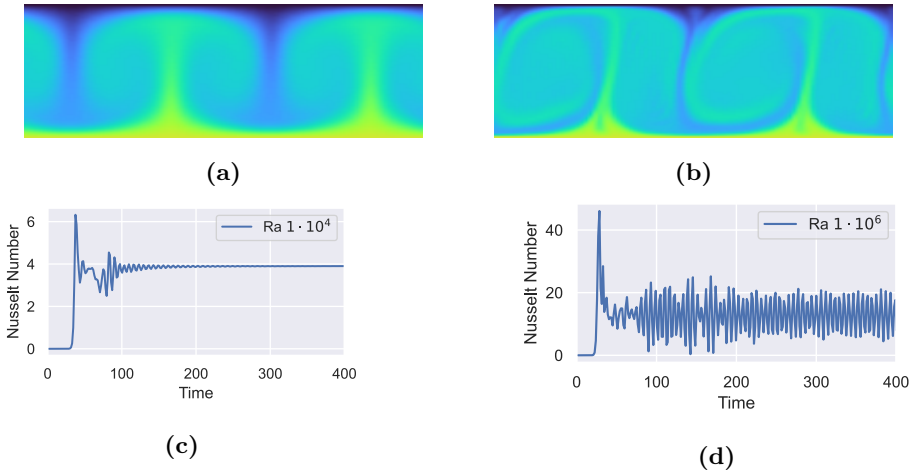


图 1. 图 (a) 展示了未控制系统在 $\mathrm{Ra} = 10^4$ 时的温度场示例，而 (c) 中的线展示了整个展开过程中的努塞尔数 Nu。图 (b) 和 (c) 显示了未控制系统在 $\mathrm{Ra} = 10^6$ 时的相同情况。

其中 θ 表示域 $\langle T \rangle_{x,y}$ 内平均温度周围的温度波动。这导致努塞尔数 Nu，该数值衡量了对流的程度，并定义为对流与传导传热的比率。根据之前的研究所 [1,5,16]，我们定义 Nu 为：[3]

$$\mathrm{Nu}(t) = \frac{\langle q(x,y,t) \rangle_{x,y}}{\kappa (T_b - T_t)/H} \,. \tag{2}$$

最初，源自底部的热量仅通过传导传递，温度在垂直方向上呈线性变化。当瑞利数超过临界阈值 $\mathrm{Ra}_c = 1708$ 时，对流开始，流体组织成贝纳德细胞（见图 1）。由于在这种状态下，热量更多的是通过对流而不是传导传递，$\mathrm{Nu} > 1$。随着 Ra 进一步增加，流体流动从有结构的转变为混沌行为不断增加的湍流传热。

模拟环境 我们使用直接数值模拟（DNS）基于开源框架 Shenfun [6] 解决 RBC 系统。控制方程通过谱伽辽金方法在二维矩形空间域上进行数值求解。[4] 我们的模拟设置遵循先前的研究 [1,5,16]，其中边界条件包括底部和

[3] T_b, T_t: 底部（加热）和顶部的温度。H: 上下边界之间的距离。κ：热扩散率。

[4] 空间维度：水平 $x \in [0, 2\pi]$，垂直 $y \in [-1, 1]$，$H = 2$，离散化为 96×64 个均匀网格点。我们的代码仓库提供了进一步的参数和方程：https://github.com/HammerLabML/RBC-Control-SARL

顶部的无滑移壁以及水平方向上的周期性边界条件。初始条件由添加到传导平衡中的小扰动给出。系统以时间步长 $\Delta t = 0.025$ 进化。

3.2 控制红细胞

遵循之前的研究所 [1,3,16]，我们专注于通过在下边界施加小的温度波动来减少对流系统中 Nu 的相关控制任务。下壁被离散化为 $N = 12$ 个加热段，每个段接收独立的控制输入 a_i 用于 $i = 1, \dots N$。为了保持 Ra 恒定，所施加的控制是居中且值受限于 $[-C, C]$，使用 $C = 0.75$ 从 $[1, 16]$ 开始进行以下连续变换：

$$\hat{T}_i' = a_i - \frac{\sum_{i=1}^{N} a_i}{N}\,, \quad \hat{T}_i = \frac{\hat{T}_i' C}{\max(1, |T'|)} \quad . \tag{3}$$

然后将变换后的控制输入 $\hat{T}_i$ 映射到水平空间维度，并在加热段之间增加平滑处理 [16]。这确保了数值模拟的稳定性。

线性控制 一个线性比例-微分控制器（PD）作为减少 RBC 系统 [1,9] 中对流的基准。下边界处的温度波动通过以下方式计算

$$a(x,t) = k_p E(x,t) + k_d E'(x,t)\,, \tag{4}$$

其中 k_p 和 k_d 是比例和微分增益，$E(x,t)$ 表示与期望状态的距离，我们将其定义为偏离无运动平衡状态 $u_y^* = 0$ 的偏差，如 [1] 所示：

$$E(x,t) = \langle u_y(x,y,t)\rangle_y - u_y^* = \langle u_y(x,y,t)\rangle_y \quad . \tag{5}$$

PD 控制策略是在 Bénard 单元之间的冷、向下流动区域施加热量以对抗对流，同时在其他地方减少热量。我们使用控制器增益 k_p=970 和 k_d=2000。所得的控制输入 $a(x,t)$ 在 $N = 12$ 段中通过平均相应的网格点进行离散化并通过方程 (3) 转换。

强化学习 强化学习（RL）是一种通过试错进行学习的方法，假设对代理在环境状态 s 下期望的行为（动作 a）给予奖励（一个数量 R），将导致这种行为在未来被加强 [12]。这种设置类似于传统的反馈控制，但 RL 可以通过表达性神经网络发现复杂的控制策略。代理遵循一种策略 $\pi(a|s)$，该策略将状态映射到动作概率。RL 的目标是找到一个策略 π^*，使其最大化期望奖励之和，

$\max_\pi \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1})\right]$，其中动作 a_t 根据 π 选择，并且 $0 \leq \gamma \leq 1$ 对未来奖励进行折现。状态转换由马尔可夫决策过程（MDP）决定，在我们的案例中，该过程由底层确定性数值模拟给出。

对于观测，我们假设在空间域上有一个 8×48 的网格探针传感器，测量局部温度和速度场 [1,16]。这些测量值被展平为一个大小为 1152 的向量作为代理的输入。动作空间由下边界 $(a_1, a_2, \ldots, a_N)$ 处加热段上的温度波动组成，其中 $a_i \in [-1, 1]$，并通过方程 (3) 进行转换。

代理的目标是通过努塞尔数 Nu（由公式 (2) 给出）来最小化对流热传递，这反映在奖励中：

$$R(s_t) = 1 - \frac{\mathrm{Nu}(s_t)}{\mathrm{Nu}_{\mathrm{Base}}(\mathrm{Ra})}\,, \tag{6}$$

其中 $\mathrm{Nu}_{\mathrm{Base}}(\mathrm{Ra})$ 是给定 Ra 下未控制系统中出现的最大 Nu，因此大约为 $R(s_t) \in [0, 1]$。

控制策略使用 PPO [10] 进行训练，这是一种无模型的、演员-评论家策略优化技术，通过采用裁剪目标函数来提高训练代理的鲁棒性，避免了大的策略更新。由于其在性能和稳定性之间的平衡，PPO 被广泛应用于各种应用领域，包括 RBC [1,16] 的控制。

3.3 奖励塑造

我们的初步结果显示，简单的策略，如 PD 控制中细胞之间的加热，也可以显著降低努塞尔数（Nu），但并不能稳定流动。在 [16] 对 $\mathrm{Ra} = 10^4$ 的观察以及我们最初的实验结果表明，细胞合并是一种有效的减少努塞尔数和稳定流体的策略。然而，PPO 代理经常陷入类似于 PD 控制的相同策略中，而训练能够合并细胞的代理则需要大量的训练努力。为了应对这一问题，我们尝试通过奖励塑造来激励细胞合并。

我们通过在域的垂直中线上使用简单的数值峰值查找方法找到垂直速度测量值 $u_y(x, 0, t)$ 的正局部极大值，检测到了潜在的单元位置 c_i。[5] 为了量化合并的程度，我们计算最大成对单元距离：

$$\mathrm{celldist} = \max\left(\{\min(|c_i - c_j|, 2\pi - |c_i - c_j|) \,|\, 1 \leq i < j \leq K\}\right)\,, \tag{7}$$

[5] 我们使用了来自 scipy.signal 的 find_peaks 函数，设置高度为 0。在湍流中寻找峰值的鲁棒性可以通过调整峰值查找算法的参数来改进。

其中函数 min 确保了在周期域 $x \in [0, 2\pi]$ 中单元 i 和 j 之间的正确距离。如果最多存在一个单元，则设置 $\text{celldist} = 0$。为了促进细胞合并同时保持较低的努塞尔数，我们修改奖励函数：

$$r_t = (1-\alpha)\left(1 - \frac{\text{Nu}(t)}{\text{Nu}_{\text{Base}}(\text{Ra})}\right) + \alpha\left(1 - \frac{\text{celldist}(t)}{\pi}\right), \tag{8}$$

其中 $\alpha \in [0, 1]$ 在奖励中平衡了细胞距离和努塞尔数。数量 $(1 - \text{celldist}(t)/\pi)$ 从 0 变化，当细胞最大程度分离时为 [6] 到 1，在单个合并的细胞情况下。

4 实验与结果

我们进行了三个实验来评估 DRL 与线性控制相比在增加 RBC 系统湍流中的有效性。首先，我们解释环境展开和 DRL 方法训练的工作原理。在实验 1 中，我们将 DRL 和线性控制在减少系统的对流方面进行比较。接下来，在实验 2 中，我们展示代理的泛化能力到其他湍流水平，并在实验 3 中引入奖励塑形技术到 DRL 方法。

4.1 剧集滚动和 PPO 训练

我们在使用被包装在 Gym 环境 [14] 中的 DNS 进行的环境滚动中训练和评估了这些方法。[7] 在接下来的实验中，我们评估了不同湍流水平下的 DRL 和线性控制，即：$\text{Ra} \in \{1e4, 1e5, 1e6, 5e6\}$，保持其他参数固定。图 1 展示了在不同的 Ra 下从无运动状态过渡到对流阶段的两个未受控 RBC 模拟，持续 400 个时间步长。在较低湍流（$\text{Ra} = 1e4$）的情况下，系统最终收敛到了一个 Nu 约为 ∼3.9 的稳定状态。相比之下，更高的湍流导致了具有典型两个对流单元的周期性行为（图 4a）。偶尔，系统收敛到单个单元，导致吸引子中的 Nu 值降低。

如 [16] 中所示，代理仅在系统对流阶段对其进行控制，从初始条件开始 400 个时间步骤后进行，确保展开操作从对流已建立的状态而非无运动状态开始。我们为每个 Ra 创建了 35 个不同对流状态的检查点，这些检查点源于系统的 35 种不同的初始条件。为了评估方法是否能很好地泛化到不同的初始条件，我们将检查点组织成大小分别为 20、5 和 10 的训练集、验证集和测试集。

[6] 注意 π 是周期域 $x \in [0, 2\pi]$ 上的最大可能距离。

[7] 观测网格：8×48，加热段落 N：12，动作限制 C：0.75。动作持续时间 1.5，情节长度：300，即。每集 200 个动作。

Mean Nusselt Number			
Ra	**Baseline**	**PD**	**PPO**
10^4	3.9 ±0.00	3.1 ±0.02	**2.6** ±0.03
10^5	6.9 ±0.01	6.9 ±0.02	**5.9** ±0.18
10^6	11.6 ±0.38	12.5 ±0.48	**11.2** ±0.35
$5 \cdot 10^6$	19.8 ±0.16	23.2 ±0.14	**17.5** ±0.21

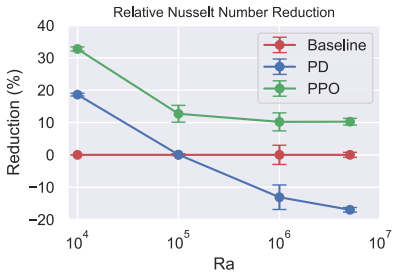


图 2. 左：对未控系统（基线）、PD 控制和 PPO 控制在不同 Ra 下的 20 种测试初始条件计算得到的时间平均努塞尔数。右：相对于基线的相对减少量。

基于 PPO 的智能体经过了 400,000 个行动步骤的训练，相当于 RBC 系统进行了 2000 次单独滚动。[8] 为了避免过拟合，我们持续在验证检查点上验证智能体的表现，并存储具有最高平均回报的智能体。对于最佳智能体，我们记录了其在 10 个测试检查点上的 Nu 减少量。

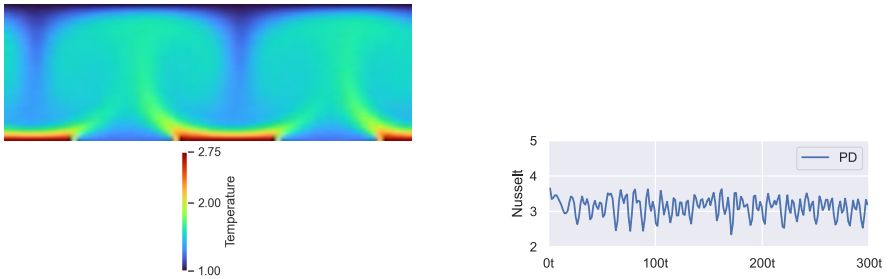
4.2 实验 1：努塞尔数减少

我们评估了 DRL 代理和 PD 控制在降低所有湍流水平下的 Nu 有效性。图 2展示了每个 Ra 的回合平均 Nu 及其相对于未控基线的相对减少量。

PD 控制在最不湍流的情况下（$\mathrm{Ra} = 10^4$）工作得很好，将努塞尔特数降低了 19%。对于 $\mathrm{Ra} = 10^6$ 和 $\mathrm{Ra} = 5 * 10^6$，PD 控制导致的状态大于未控制基线的 Nu。图 3可视化了在 $\mathrm{Ra} = 10^4$ 处的 PD 控制策略，仅在单元之间施加热量。这使得系统保持两单元设置，将努塞尔特数降低到大约 ∼3.1，在展开过程中有些变化。

相比之下，DRL 代理在所有 Ra 上始终取得了更好的结果，$\mathrm{Ra} = 10^4$ 减少了 33%，$\mathrm{Ra} = 10^5$ 减少了 15%，而 $\mathrm{Ra} > 10^5$ 大约减少了 10%。图 4展示了低湍流系统中的 DRL 控制策略（$\mathrm{Ra} = 10^4$）。在 60 个时间步内，代理积极地将两个单元合并为一个单元（图 4c），以减少总的对流并稳定 Nu 至 2.6 的值。之后，代理最大化了单个单元的宽度（图 4d），有时会导致返回到双细胞状态（图 4e）。由于这种行为没有恶化平均回报，如图 4f所示，因此未受到惩罚。控制策略在较高湍流水平 $\mathrm{Ra} > 10^4$ 上有所不同：代理没有将

[8] 来自 Stable-Baselines-3 的 PPO 实现，[8]，20 并行环境每次迭代生成 4,000 个样本，$\gamma = 0.99$，熵 $\beta = 0.01$，演员和评论家都使用两层（64 个隐藏单元）的神经网络。



(a) $t = 30$ **(b)** 努塞尔数的减小通过展开过程

图 3. 线性 PD 控制的示例，系统进行 $\mathrm{Ra} = 10^4$ 展开。

系统驱动到单细胞状态，而是稍微调整了 PD 控制的策略，这仍然有效地减少了 Nu。[9]

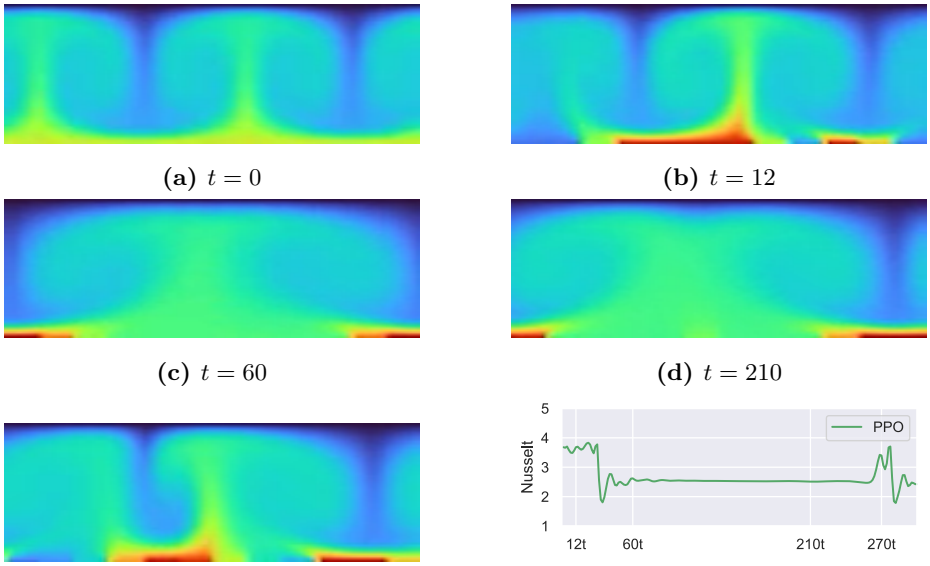
4.3 实验 2：不同湍流水平的泛化能力

鉴于湍流的分层结构，在低湍流系统中训练的代理可能学会了有效的控制模式，这些模式在高湍流环境中也有效。我们评估了在 $\mathrm{Ra} = 10^4$ 和 $\mathrm{Ra} = 10^5$ 训练的代理降低不同 Ra 下对流的能力。图 5a显示了相对于未控制基线和前一实验中 PPO 基线，Nu 的相对减少情况。PPO-Ra1e4 在 $\mathrm{Ra} = 10^5$ 上的表现甚至比 PPO 基线更好，成功地转移了细胞合并策略，但在更高的 Ra 下失败。PPO-Ra1e5 实现了在所有 Ra 下降低 Nu，然而其表现不如 PPO 基线，并且从未合并过细胞。这些不同行为的一种可能解释是，代理训练于 $\mathrm{Ra} = 10^4$ 时获得了细胞合并策略，这是由于系统中的流动更为稳定，而更高的 Ra 流动更加混沌，使得学习足够的控制变得更加困难。

4.4 实验 3：奖励塑造

我们评估了引入奖励函数 (8) 并平衡每个 Ra 的 $\alpha = 0.25$ 和 $\alpha = 0.5$ 值的影响。我们将使用奖励塑造训练的代理称为 RS-代理，而未使用奖励塑造训练的代理则称为 No-RS-代理。图 5b显示努塞尔数减少与 No-RS-代理（图 2）的情况相当。然而，RS-代理在合并对流单元方面表现出显著更高的成功率，如图 6a所示：单元始终合并为 $\mathrm{Ra} = 10^4$ 和 $\mathrm{Ra} = 10^5$，甚至对于

[9] 进一步的结果和代理的视频可在 https://github.com/HammerLabML/RBC-Control-SARL



(a) $t = 0$ **(b)** $t = 12$ **(c)** $t = 60$ **(d)** $t = 210$ **(e)** $t = 270$ **(f)** 努塞尔数减小的展开过程

图 4. 示例展示了 PPO 控制在系统 $\mathrm{Ra} = 10^4$ 测试集展开中的应用。

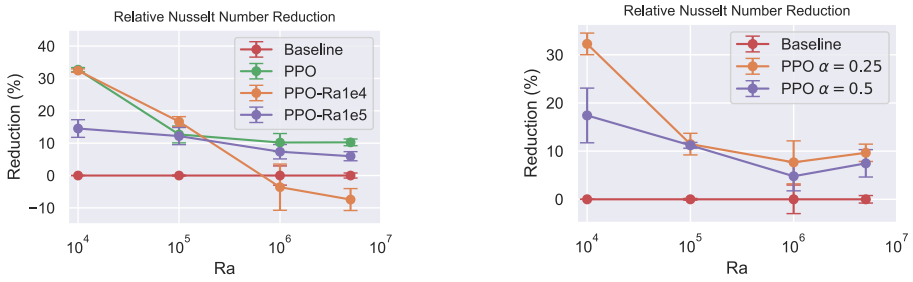
$\mathrm{Ra} = 10^6$，合并仍然是可能的。相比之下，无 RS-代理显示出显著较低的单元合并。

RS 代理的单元合并策略体现在回放最后 40 步中 Nu 随时间的变化，如图 6b所示：RS 代理实现了对 $Ra = 10^4$ 和 10^5 中 Nu 变化的近乎完全减少，这源于更加稳定的单单元流动，类似于图 1c中的努塞尔数行为。此外，对于 $\mathrm{Ra} = 10^6$，达到单个单元状态的会话在 Nu 上的变化显著减少，在箱线图中显示为异常值。

RS-代理在剧集早期合并单元格，并且比 No-RS 代理更快，如图 6c所示。值得注意的是，我们发现训练在 $\mathrm{Ra} = 10^5$ 上的 RS-代理即使在 $\mathrm{Ra} = 10^6$ 时也始终合并单元格，导致努塞尔数变化显著减少。

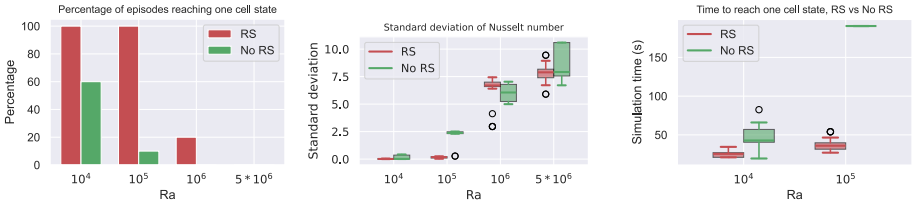
5 讨论

我们的结果表明，单智能体强化学习在各种湍流水平上表现出色，显著优于 PD 控制，通过发现非平凡的控制策略。对于 $\mathrm{Ra} = 10^4$ 观察到的努塞尔数减少 33%高于在 [16] 中实现的结果，这可能是由于单智能体强化学习具有更大的表达能力。因此，应将单智能体强化学习策略视为可达到性能的上



(a) **(b)**

图 5. (a) 在第 4.3节的一般化任务中，20 个测试初始条件下的 Nu 减少情况以及 (b) 当使用奖励塑造进行训练时（第 4.4节）的情况。图 (a) 中的绿线是来自第 4.2节的结果，并在此处作为基线。



(a) **(b)** **(c)**

图 6. 奖励塑造在几个统计量中的效果。

限。这允许更好地解释高度可扩展但表达力较低的方法的表现，例如在 [16] 中的情况。单一代理设置的局限性在于其可扩展性：尽管我们已经证明可以在二维领域中训练高性能代理是可行的，但在三维环境中是否仍然可行则不清楚 [15]。然而，奖励塑造显著减轻了样本复杂度问题，使得代理能够快速且稳定地使流体流动稳定到更高的瑞利数。

我们的训练代理展示了强大的泛化性能：所有为 $\mathrm{Ra} = 10^4$ 和 $\mathrm{Ra} = 10^5$ 训练的代理都在各种测试初始条件下成功运用了它们非同寻常的策略。此外，它们还能够推广到更高的瑞利数：为 $\mathrm{Ra} = 10^4$（没有奖励整形）训练的代理能持续合并对流单元至 $\mathrm{Ra} = 10^5$。为 $\mathrm{Ra} = 10^5$（带有奖励整形）训练的代理则将单元合并到了 $\mathrm{Ra} = 10^6$。尽管更高 Ra 的单个细胞状态比直接为此设置训练的代理所达到的状态更不稳定，但可能只需最少程度的微调即可减少所有波动。我们假设，在更高 Ra 下增加的湍流使得代理训练更具挑战

性，阻碍了对单元合并策略的成功发现。可能存在不同的执行参数可以更容易地实现稳定流动以适应更高的 Ra。同时，还需要考虑来自实际设置的额外约束，如加热器温度和执行持续时间的限制。

6 结论

在这项工作中，我们训练了 RL 代理来控制 Rayleigh-Bénard 对流。这些代理发现了非平凡的策略并表现出强大的泛化性能，使其在实际情况下具有可行性。奖励塑造以及在瑞利数上进行泛化的能力突出了样本高效学习的潜力。

在未来的工作中，我们旨在通过使用基于模型的强化学习框架进一步提高样本效率，并集成用于 RBC 的神经操作代理模型，如我们在 [11] 中初步研究的内容。此外，我们将调查我们的方法在更现实的三维环境中的应用。

Acknowledgments. 作者感谢项目 “SAIL：智能社会技术系统的可持续生命周期”（资助 ID NW21-059A 和 NW21-059D）的经济支持，该项目由德国北莱茵-威斯特法伦州文化和科学部的 “网络 2021” 计划提供资金。

参考文献

1. Beintema, G., Corbetta, A., Biferale, L., Toschi, F.: Controlling Rayleigh – Bénard convection via reinforcement learning. Journal of Turbulence (2020). doi:10/gg89qv
2. Garnier, P., Viquerat, J., Rabault, J., Larcher, A., Kuhnle, A., Hachem, E.: A review on deep reinforcement learning for fluid mechanics. Computers & Fluids (2021). doi:10/gjv5sj
3. Howle, L.E.: Active control of Rayleigh–Bénard convection. Physics of Fluids (1997). doi:10/cn2458
4. Jeon, J., Rabault, J., Vasanth, J., Alcántara-Ávila, F., Baral, S., Vinuesa, R.: Advanced deep-reinforcement-learning methods for flow control: Group-invariant and positional-encoding networks improve learning speed and quality (2024). doi:10/pd4z
5. Markmann, T., Straat, M., Hammer, B.: Koopman-based surrogate modelling of turbulent Rayleigh-Bénard convection. In: IJCNN (2024). doi:10/pd3x
6. Mortensen, M.: Shenfun: High performance spectral galerkin computing platform. Journal of Open Source Software (2018). doi:10/gtxkzd

7. Pandey, A., Scheel, J.D., Schumacher, J.: Turbulent superstructures in Rayleigh-Bénard convection. Nature Communications (2018). doi:10/gdp5wg
8. Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., Dormann, N.: Stable-baselines3: Reliable reinforcement learning implementations. Journal of Machine Learning Research (2021)
9. Remillieux, M., Zhao, H., Bau, H.: Suppression of Rayleigh-Bénard convection with proportional-derivative controller. Physics of Fluids (2007). doi:10/c6rhpp
10. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal Policy Optimization Algorithms (2017). doi:10/gs5mt5
11. Straat, M., Markmann, T., Hammer, B.: Solving Turbulent Rayleigh-Bénard Convection using Fourier Neural Operators (2025). doi:10/pd3z, accepted at ESANN 2025, Bruges
12. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. A Bradford Book (2018)
13. Tang, J., Bau, H.H.: Stabilization of the no-motion state in Rayleigh-Bénard convection through the use of feedback control. Physical Review Letters (1993). doi:10/cjkrn2
14. Towers, M., Kwiatkowski, A., Terry, J., Balis, J.U., Cola, G.D., Deleu, T., Goulão, M., Kallinteris, A., Krimmel, M., KG, A., Perez-Vicente, R., Pierré, A., Schulhoff, S., Tai, J.J., Tan, H., Younis, O.G.: Gymnasium: A standard interface for reinforcement learning environments (2024). doi:10/pd3w
15. Vasanth, J., Rabault, J., Alcántara-Ávila, F., Mortensen, M., Vinuesa, R.: Multi-agent Reinforcement Learning for the Control of Three-Dimensional Rayleigh–Bénard Convection. Flow, Turbulence and Combustion (2024). doi:10/pd42
16. Vignon, C., Rabault, J., Vasanth, J., Alcántara-Ávila, F., Mortensen, M., Vinuesa, R.: Effective control of two-dimensional Rayleigh – Bénard convection: Invariant multi-agent reinforcement learning is all you need. Physics of Fluids (2023). doi:10/gsjxkh