# 实验基本交代信息

Based on our proposal outlined above, we implemented an agent that learns by using a DRL (see Sec. II B) to manage the environments described in Sec. II D. The agent is trained for a maximum number of 104 episodes, where the learning success is evaluated every 50 episodes. Single simulation experiments can be carried out on standard notebook computers in a reasonable computing time (one to two hours on a single machine). Using a tuned hyperparameter set (see Table I in the Appendix for details), we can formulate three key findings of this work that is outlined below. First, we find that learning in terms of increasing rewards in the environments is indeed possible. Second, we investigate the specific pathways found by the learner and observe that the agent acts with great farsightedness. Moreover, we see a general strategy behind the detected trajectories that the learner has developed. Third, we explore that the agent also achieves good performance in scenarios in which the state space is only partially observable to the agent.

根据上面概述的提案，我们实施了一种通过使用DRL学习（请参阅第II b）来管理SEC中描述的环境的代理。 II D.对代理人进行最多104集的训练，其中每50集对学习成功进行评估。单个仿真实验可以在合理的计算时间（一台机器上一到两个小时）在标准笔记本计算机上进行。使用调谐的超参数集（有关详细信息，请参见附录中的表I），我们可以制定以下概述的这项工作的三个关键发现。首先，我们发现，在环境中的奖励增加确实是可能的。其次，我们研究了学习者发现的特定途径，并观察到代理人的远视性能。此外，我们看到了学习者发展的轨迹背后的一般策略。第三，我们探讨了代理在状态空间仅部分可观察到的代理人的场景中也可以实现良好的性能。

1. Training and stability

In order to verify the overall applicability of our algorithm, we first analyze the learning behavior in general. Unlike in supervised learning, where one can evaluate the performance of an algorithm by evaluating it on a set of test data, it is not obvious how to evaluate accurately the training progress an agent makes in RL problems. Here, we stick to the method used by Mnih et al.25 visualizing the training properly. We plot the total reward the agent collects during one run over the number of learning episodes. Each value is computed as a running average over 50 episodes. Each curve is the average of 100 independent simulations.

为了验证算法的总体适用性，我们首先分析学习行为。与监督学习不同，在该学习中可以通过在一组测试数据上评估算法的性能，在一组测试数据上进行评估，如何准确评估代理在RL问题中所做的培训进度。在这里，我们坚持MNIH等人使用的方法。25正确可视化训练。我们将代理商在一次学习情节数量中收集的总奖励绘制了总奖励。每个值计算为50集以上的平均值。每条曲线是100个独立模拟的平均值。

As a result, we see that, after a certain number of episodes, the average reward per episode significantly increases in our environments (see Fig. 2). Obviously, the agent finds trajectories that reveal a high reward. In other words, it learns to manage the environment. We conclude that management can indeed be learned by the agent.

结果，我们看到，经过一定数量的发作，我们环境的平均奖励显着增加（见图2）。显然，经纪人发现轨迹揭示了很高的奖励。换句话说，它学会了管理环境。我们得出的结论是，代理商确实可以学习管理。

Furthermore, we observe that the learning of the agent is stabilized by using the extensions to DQN-Learning as outlined in Sec. II C. The plot suggests that the usage of dueling network architectures combined with double-Q-learning (DDQN + Duel) and prioritized experience replay with importance sampling (PER IS) significantly increases the performance of our DQN-Agent. The positive effect of PER IS can be explained by the observation that, in both environments, we find states in the resulting trajectories, where the dynamics significantly changes (as it will be outlined below). Experiences containing these states will be privileged in the learning process by PER IS. This is in good agreement with the results in Ref. 52. Therefore, all results outlined below are achieved by using our best performing agent (DDQN + Duel + PER IS), if not stated otherwise. Moreover, this is in qualitatively good accordance with other comparisons of different learning architectures, as, e.g., presented in Ref. 34, and the learning curves show a similar shape as seen in other DRL applications.24,25,34

此外，我们观察到，通过使用SEC中概述的DQN学习的扩展，对试剂的学习进行了稳定。 II C.该图表明，对决网络体系结构与双Q学习（DDQN DUEL）相结合，并优先考虑使用重要性采样（PER IS）的优先经验重播（PER）大大提高了我们的DQN代理的性能。 Per IS的积极作用可以通过以下观察结果来解释：在这两个环境中，我们在所得轨迹中发现了状态，在该轨迹中，动态发生了显着变化（如下所述）。每个国家的经验将在学习过程中享有特权。这与参考文献中的结果非常吻合。 52。因此，下面概述的所有结果都是通过使用我们的最佳性能代理（DDQN DUEL PER IS）来实现的，如果没有另行说明。此外，这在定性上与不同学习体系结构的其他比较相符，例如，参考文献中提出。 34，学习曲线的形状与其他DRL应用相似。24,25,34

# 不同算法单一奖励的曲线图

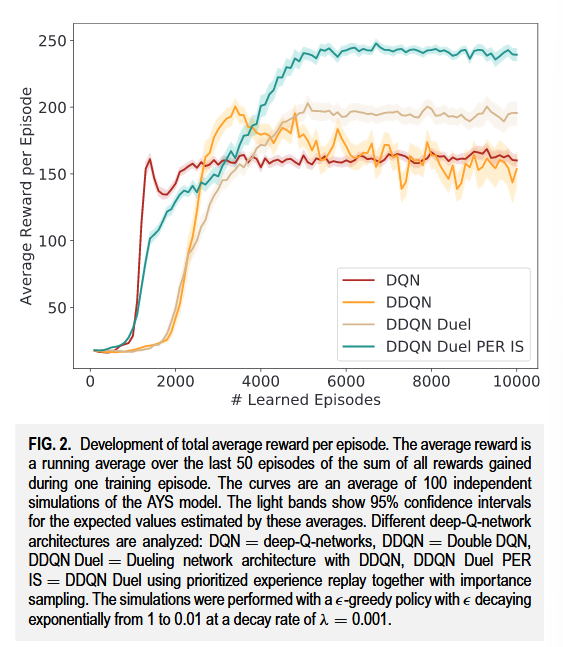


FIG. 2. Development of total average reward per episode. The average reward is a running average over the last 50 episodes of the sum of all rewards gained during one training episode. The curves are an average of 100 independent simulations of the AYS model. The light bands show 95% confidence intervals for the expected values estimated by these averages. Different deep-Q-network architectures are analyzed: DQN = deep-Q-networks, DDQN = Double DQN, DDQN Duel = Dueling network architecture with DDQN, DDQN Duel PER IS = DDQN Duel using prioritized experience replay together with importance sampling. The simulations were performed with a -greedy policy with decaying exponentially from 1 to 0.01 at a decay rate of λ = 0.001

FIG。2 .每集总平均奖赏的发展。平均奖赏是在一个训练集中获得的所有奖赏的总和的最后50个片段的运行平均值。曲线是AYS模型100次独立模拟的平均值。光带显示了这些平均值估计的期望值的95 %置信区间。分析不同的深度Q网络架构：DQN = deep - Q网络，DDQN =双层深度Q网络，DDQN Duel = Dueling网络架构，DDQN Duel PER IS = DDQN Duel，采用优先级经验重放和重要性采样。模拟采用λ = 0.001的衰减率从1指数衰减到0.01的"贪婪"策略

# 三维单个算法单个reward的轨迹分析图

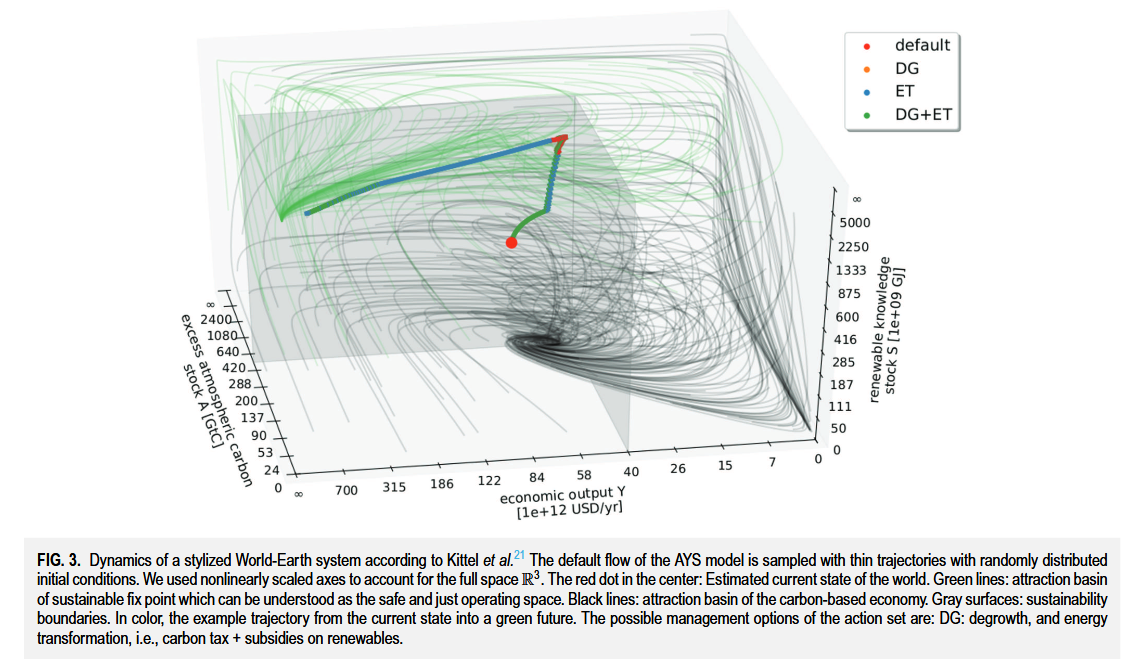


FIG. 3. Dynamics of a stylized World-Earth system according to Kittel et al.21 The default flow of the AYS model is sampled with thin trajectories with randomly distributed initial conditions. We used nonlinearly scaled axes to account for the full space R3. The red dot in the center: Estimated current state of the world. Green lines: attraction basin of sustainable fix point which can be understood as the safe and just operating space. Black lines: attraction basin of the carbon-based economy. Gray surfaces: sustainability boundaries. In color, the example trajectory from the current state into a green future. The possible management options of the action set are: DG: degrowth, and energy transformation, i.e., carbon tax + subsidies on renewables.

FIG . 3 .根据Kittel et al21，一个典型的世界-地球系统的动力学AYS模型的缺省流是用随机分布的初始条件的细轨迹采样的。我们使用非线性比例轴来描述整个空间R3。中心的红点：估计当前世界的状态。绿线：可持续固定点的吸引盆，可理解为安全、公正的操作空间。黑线：碳基经济的吸引盆。灰色表面：可持续性边界。在颜色上，该示例轨迹从当前状态进入绿色未来。行动集的可能管理选择是：DG：去增长和能源转型，即碳税+对可再生能源的补贴。

【具体单个算法单个reward对应的3维轨迹，结果分析】

## 【具体单个算法单个reward对应的3维轨迹，结果分析】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

我们发现，即使环境的动态对智能体事先是未知的，它也能够在可持续边界(见图3)中找到轨迹，这在另一项基于使用状态空间离散化的可行性理论算法的研究中被认为是不可能的。21

【交代框架中转换的细节过程】

## 【交代框架中转换的细节过程】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

【对于高维轨迹中的细节交代，边界情况的特点】

## 【对于高维轨迹中的细节交代，边界情况的特点】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

图3中，在靠近边界的区域，能量转换( energy transformation，ET )选项(代表一种能源税或补贴)在短时间内交替开启和关闭，实质上达到了持续实施较小税收/补贴的效果。

【对于边界情况特点的总结】

## 【对于边界情况特点的总结】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

因此，提供不同的税收水平作为个体选项可能会进一步提高学习成功。

# 过渡句

# 高维轨迹分析图

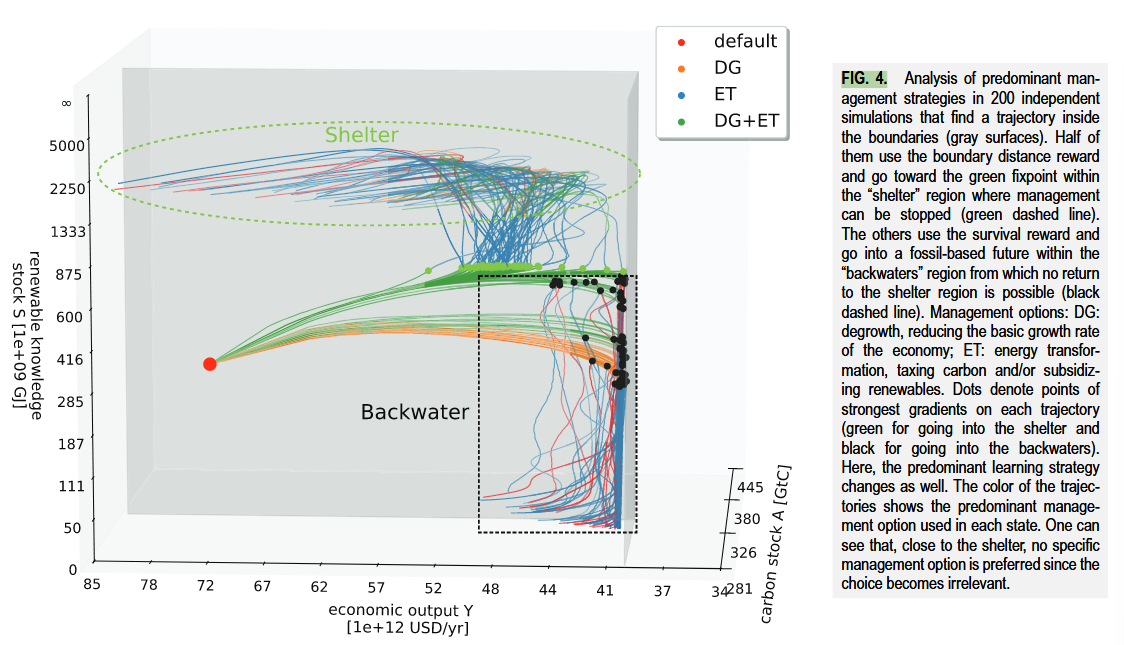


FIG. 4. Analysis of predominant management strategies in 200 independent simulations that find a trajectory inside the boundaries (gray surfaces). Half of them use the boundary distance reward and go toward the green fixpoint within the “shelter” region where management can be stopped (green dashed line). The others use the survival reward and go into a fossil-based future within the “backwaters” region from which no return to the shelter region is possible (black dashed line). Management options: DG: degrowth, reducing the basic growth rate of the economy; ET: energy transformation, taxing carbon and/or subsidizing renewables. Dots denote points of strongest gradients on each trajectory (green for going into the shelter and black for going into the backwaters). Here, the predominant learning strategy changes as well. The color of the trajectories shows the predominant management option used in each state. One can see that, close to the shelter, no specific management option is preferred since the choice becomes irrelevant.

如图。 4。对200个独立模拟中主要管理策略的分析，这些策略在边界内找到轨迹（灰色表面）。他们一半使用边界距离奖励，然后朝着可以停止管理的“庇护所”区域内的绿色固定点（绿色虚线）。其他人则使用生存奖励，并进入“回死者”地区的基于化石的未来，从中不可能返回庇护所地区（黑色虚线）。管理选择：DG：降解，降低经济的基本增长速度； ET：能源转化，征税碳和/或补贴可再生能源。点表示每个轨迹上最强梯度的点（进入庇护所的绿色，黑色以陷入死水）。在这里，主要的学习策略也改变了。轨迹的颜色显示了每个状态中使用的主要管理选项。可以看到，在靠近庇护所的地方，由于选择无关紧要，因此没有任何特定的管理选项。

【分析不同奖励函数下， 单一算法下结果不同】

## 【分析不同奖励函数下， 单一算法下结果不同】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

如果选择边界距离奖励，经过足够长的学习，智能体总是在( A , Y , S) = ( 0 ,∞,∞)处找到一条通往"绿色不动点"的路径，其中到边界的距离是最大化的。对于生存奖励，代理人只对停留在边界内感兴趣。因此，它找到了通向绿色固定点的途径，以及通向靠近S = 0边界的区域的途径，然后在那里它设法保持。

【大量分析，得出共性特点】

## 【大量分析，得出共性特点】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

虽然学习者找到了许多可行的路径，但是我们发现的学习策略可以泛化。我们分析了代理人在轨迹的不同部分使用最多的管理选项。

【大量试验下-不同reward下对应的高维轨迹实验结果特点-图4】

## 【大量试验下-不同reward下对应的高维轨迹实验结果特点-图4】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

如图4所示。这些不同区域的主要管理选择现在被用于下面的讨论。

【引入高维分析中label特点的理论部分】

## 【引入高维分析中label特点的理论部分】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

如图4所示的不同区域可以用一个关于具有管理选项和期望状态的动态系统的状态空间的定性拓扑的数学理论来分析，称为可持续管理的拓扑( TSM )。

【对高维轨迹分析中的标记来进行解释分析】

## 【对高维轨迹分析中的标记来进行解释分析】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

有趣的是，这些区域可以被看作大致对应于TSM框架中的一些概念，特别是"庇护所"和"回水区"的概念。这些区域的大致位置在图4中用虚线表示。

# 高维轨迹分析图-分析视角2

【从另一个角度来细节陈述-不同reward下对应的高维轨迹实验结果特点-图4】

## 【从另一个角度来细节陈述-不同reward下对应的高维轨迹实验结果特点-图4】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

我们确定代理商使用的一种通用策略。从当前状态开始，我们发现，为了永远在边界范围内，一开始只使用一种能量转换（ET）或降解（DG）的单个管理选项就不足以保持。相反，必须应用ET和DG以确保将系统保持在未来的可持续性边界内。要了解这种行为，必须回顾两个可能的管理选项DG和ET的效果（有关详细信息，我们参考附录）。

【PB分析】

## 【PB分析】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

两个边界Ā和Ȳ对于学习者来说都是危险的。仅使用选项ET会导致可再生知识的增加，但违反了Ā边界。

【具体分析多维度PB中的冲突特点】

## 【具体分析多维度PB中的冲突特点】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

应用选项DG一方面会尊重̄A边界，但另一方面将击中Ȳ边界。

【概括性总结agent策略的组合特点】

## 【概括性总结agent策略的组合特点】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

学习者发现的策略是两种选择的组合：

【根据时间线来分析agent策略的action变化过程】

## 【根据时间线来分析agent策略的action变化过程】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

首先，它使用选项ET DG到达Ā边界的一定距离。但是，y limim的限制临近。

【对特点的时间点来进行分析】

## 【对特点的时间点来进行分析】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

在特定的时间点，代理必须将其主要的管理策略更改为ET，以使可再生知识库存增加速度更快，并且代理避免违反Ȳ的边界价值。

【结合实验结果来对特定现象进行分析-图4】

## 【结合实验结果来对特定现象进行分析-图4】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

在DG ET向ET变化的特定点上，轨迹的急剧转弯发生（见图4）。如果S在这一点上足够大，转弯是“向上的”，并且一段时间后，无论选择的管理选择如何，现在每个轨迹都通向无限的经济产出和可再生知识的地区，则可以“停止”，从而导致轨迹又有急剧的转变。

【结合理论来印证这其中的特点部分】

## 【结合理论来印证这其中的特点部分】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

在TSM中，这样一个安全的区域称为庇护所。但是，如果S在转折点太小，则转弯是向S = 0的“向下”，保持靠近社会基础边界。

【ref引出该概念的特点部分】

## 【ref引出该概念的特点部分】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

在参考21，结果表明，这导致了一个名为“死水”的地区，无法再到庇护所，但人们仍然可以一遍又一遍地保持在边界内。

# 实验小结

【实验小结，】

## 【实验小结，】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

总而言之，代理商了解到，管理层特定变更的时机至关重要。

【从不同尺度上对策略进行分析-国家尺度】

## 【从不同尺度上对策略进行分析-国家尺度】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

对所得途径的一般解释是，例如通过征税化石，对于确保进一步发展非常重要。

【转折，说明策略分析的时间滞问题】

## 【转折，说明策略分析的时间滞问题】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

但是，为了达到安全状态而不会违反可持续性界限，也需要一段时间的降解政策。

# State-action 分析-成功的管理

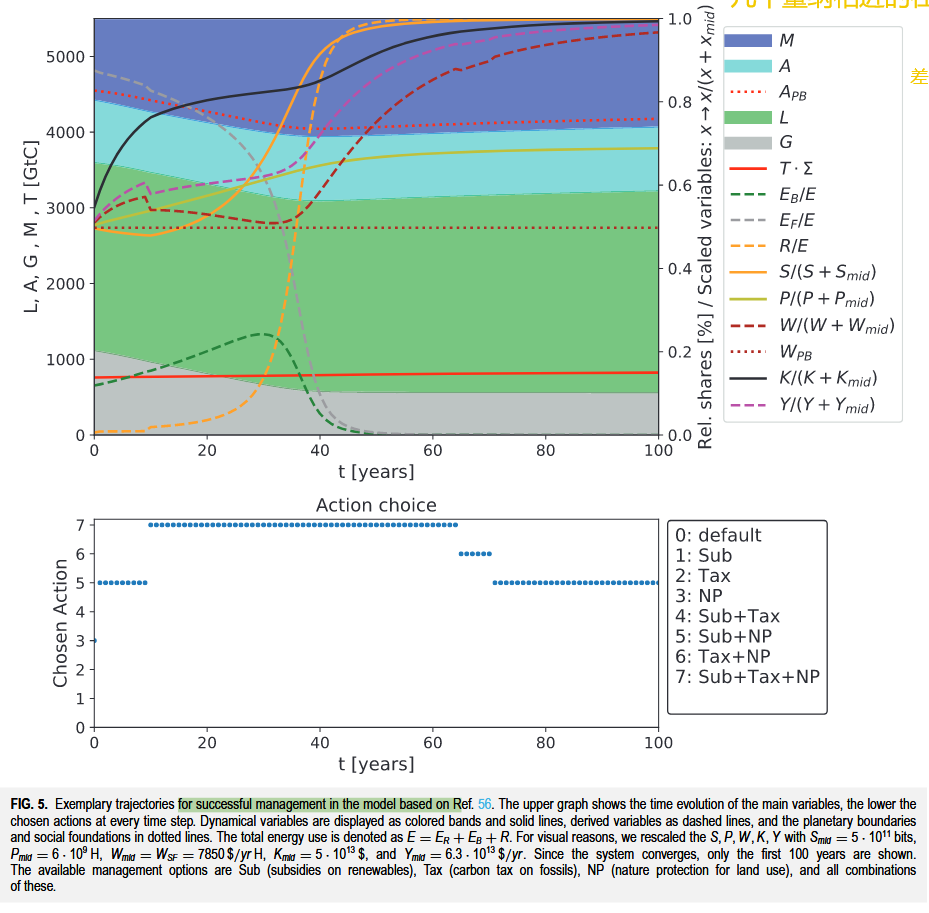


FIG. 5. Exemplary trajectories for successful management in the model based on Ref. 56. The upper graph shows the time evolution of the main variables, the lower the chosen actions at every time step. Dynamical variables are displayed as colored bands and solid lines, derived variables as dashed lines, and the planetary boundaries and social foundations in dotted lines. The total energy use is denoted as E = ER + EB + R. For visual reasons, we rescaled the S, P, W, K, Y with Smid = 5 · 1011 bits, Pmid = 6 · 109 H, Wmid = WSF = 7850 $/yr H, Kmid = 5 · 1013 $, and Ymid = 6.3 · 1013 $/yr. Since the system converges, only the first 100 years are shown. The available management options are Sub (subsidies on renewables), Tax (carbon tax on fossils), NP (nature protection for land use), and all combinations of these.

如图。 5。基于参考文献的模型中成功管理的示例性轨迹。 56。上图显示了主要变量的时间演变，每个时间步骤所选的动作越低。动态变量显示为彩色带和实线，衍生变量作为虚线，以及虚线中的行星边界和社会基础。总能源使用表示为e = er eb R.出于视觉原因，我们以smid = 5·1011位重新缩放了S，P，W，K，Y，PMID = 6·109 H，WMID = WMID = WSF = WSF = 7850 $/yr H，kmid = 5·1013 $，和Ymid = 6.3 $/YM/YRR。由于系统收敛，因此仅显示最初的100年。可用的管理选项是子（可再生能源的补贴），税收（化石碳税），NP（土地使用的自然保护）以及所有这些组合。

2 . c：GLOBAL模型中的路径

【衔接，说明框架移植的有效性】

## 【衔接，说明框架移植的有效性】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

我们通过将其应用到c：GLOBAL模型中，验证了我们的框架在高维环境中同样有效。

【说明框架移植过程中的缺点，某些理论不适用-生存力理论】

## 【说明框架移植过程中的缺点，某些理论不适用-生存力理论】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

虽然生存力理论等经典方法由于维度的原因不再适用，

【转折，说明drl的优越之处，还是有帮助的】

## 【转折，说明drl的优越之处，还是有帮助的】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

但我们的DRL学习器也能够在该模型中检测到可持续未来的解决方案，

【总图说明-copan模型的state和action图片的意义】

## 【总图说明-copan模型的state和action图片的意义】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

如图5所示

【交代，copan基本实验的特点部分-time-收敛性】

## 【交代，copan基本实验的特点部分-time-收敛性】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

在这里，一个学习片段的最大长度为500年。成功的轨迹通常在大约100年后就已经收敛。然而，为了考虑长期的影响，我们进行了长达500年的模拟，因为我们观察到看似收敛的轨迹有时会在更晚的时间跨越边界，这对学习器提出了额外的挑战。学习者发现的一般策略变成了大部分时间。

【分析-state里面的特点-action转换的原因】

## 【分析-state里面的特点-action转换的原因】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

关键之处在于开征碳税的时机，不能在不违背社会基础边界的前提下立即使用。它只是在以后才被开启，一旦可再生能源通过了他们的大部分学习曲线，就再次被关闭。

An interesting observation regarding the farsightedness of the agent is the following. After some learning episodes, the agent often uses trajectories that do not use any management during the years 20–60, which keeps the system within the boundaries for some time but leads to a violation of ̄A later for some t > 100 yrs. One example trajectory can be found in Fig. 7 in the Appendix. Only after many more episodes, the agent learns to act with foresight and use management options early on that only make a recognizable difference much later and avoid crossing the boundaries. This is indeed a key feature for the success of DRL and shows the potential power of the method.

# 单一算法单一奖励函数的观测信息部分-成功率测试的分析包含

­

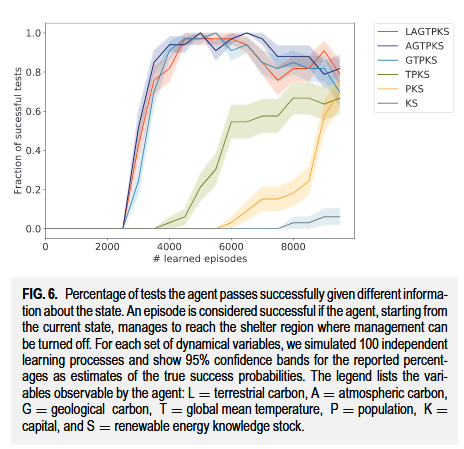


FIG. 6. Percentage of tests the agent passes successfully given different information about the state. An episode is considered successful if the agent, starting from the current state, manages to reach the shelter region where management can be turned off. For each set of dynamical variables, we simulated 100 independent learning processes and show 95% confidence bands for the reported percentages as estimates of the true success probabilities. The legend lists the variables observable by the agent: L = terrestrial carbon, A = atmospheric carbon, G = geological carbon, T = global mean temperature, P = population, K = capital, and S = renewable energy knowledge stock.

如图。 6。测试百分比成功地通过了有关状态的不同信息。如果代理商从当前状态开始，则将一集被认为是成功的，设法到达了可以关闭管理层的庇护所地区。对于每组动态变量，我们模拟了100个独立的学习过程，并显示了报告百分比的95％置信频带，作为对真实成功概率的估计。该传说列出了代理可观察到的变量：L =地面碳，A =大气碳，G =地质碳，T =全球平均温度，P =人口，K =资本和S =可再生能源知识库存。

However, taking a look at the stability of the learning (see Fig. 6), we observe that the learning success in the copan:GLOBAL model also decreases again after a still larger number of episodes. As a possible explanation, we suggest that this is connected to the replay buffer. To avoid this phenomenon, the replay buffer needs to contain experiences, especially about the time steps where the dynamics of the system changes significantly.58 After many successful runs, we still continue collecting observations in the memory buffer at every time step. Therefore, it mostly contains experiences for time points t > 50 yrs. However, especially the first time steps are crucial to avoid transgressing boundaries at later times as outlined above. These are, therefore, essential for the learning success. It seems that the agent tends to forget about experiences from early time steps and the learning success decreases. Further investigation considering the question of which experiences should be stored in the replay buffer could be the first step to overcome this issue.

【转折引子，单一算法单一奖励函数的观测信息部分-成功率测试的分析包含-图6】

## 【转折引子，单一算法单一奖励函数的观测信息部分-成功率测试的分析包含-图6】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

但是，从学习的稳定性来看（见图6），我们观察到Copan中的学习成功：全球模型在发作数量较大后也会再次降低。作为可能的解释，我们建议将其连接到重播缓冲区。

【对于算法细节中的问题进行了拓展分析，】

## 【对于算法细节中的问题进行了拓展分析，】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

为了避免这种现象，重播缓冲区需要包含体验，尤其是在系统动力学发生显着变化的时间步骤中。58经过许多成功的运行，我们仍然继续在每个时间步骤中收集内存缓冲区中的观察值。因此，它主要包含时间点T＆GT的经验；t> 50年。

【实践具体过程的分析】

## 【实践具体过程的分析】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

但是，尤其是第一次步骤对于避免以上概述的以后的时间避免违反边界至关重要。

【小结实验，agent做出策略塞饭上的原因】

## 【小结实验，agent做出策略塞饭上的原因】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

因此，这些对于学习成功至关重要。看来，代理商倾向于忘记早期步骤和学习成功的经验。

【后续对该实验结果的建议】

## 【后续对该实验结果的建议】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

进一步的调查考虑了应该在重播缓冲区中存储哪些经验的问题，这可能是克服此问题的第一步。

However, taking a look at the stability of the learning (see Fig. 6), we observe that the learning success in the copan:GLOBAL model also decreases again after a still larger number of episodes. As a possible explanation, we suggest that this is connected to the replay buffer. To avoid this phenomenon, the replay buffer needs to contain experiences, especially about the time steps where the dynamics of the system changes significantly.58 After many successful runs, we still continue collecting observations in the memory buffer at every time step. Therefore, it mostly contains experiences for time points t > 50 yrs. However, especially the first time steps are crucial to avoid transgressing boundaries at later times as outlined above. These are, therefore, essential for the learning success. It seems that the agent tends to forget about experiences from early time steps and the learning success decreases. Further investigation considering the question of which experiences should be stored in the replay buffer could be the first step to overcome this issue.

【转折引子，单一算法单一奖励函数的观测信息部分-成功率测试的分析包含-图6】

## 【转折引子，单一算法单一奖励函数的观测信息部分-成功率测试的分析包含-图6】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

但是，从学习的稳定性来看（见图6），我们观察到Copan中的学习成功：全球模型在发作数量较大后也会再次降低。作为可能的解释，我们建议将其连接到重播缓冲区。

【对于算法细节中的问题进行了拓展分析，】

## 【对于算法细节中的问题进行了拓展分析，】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

为了避免这种现象，重播缓冲区需要包含体验，尤其是在系统动力学发生显着变化的时间步骤中。58经过许多成功的运行，我们仍然继续在每个时间步骤中收集内存缓冲区中的观察值。因此，它主要包含时间点T＆GT的经验；t> 50年。

【实践具体过程的分析】

## 【实践具体过程的分析】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

但是，尤其是第一次步骤对于避免以上概述的以后的时间避免违反边界至关重要。

【小结实验，agent做出策略塞饭上的原因】

## 【小结实验，agent做出策略塞饭上的原因】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

因此，这些对于学习成功至关重要。看来，代理商倾向于忘记早期步骤和学习成功的经验。

【后续对该实验结果的建议】

## 【后续对该实验结果的建议】

[7] F. M. Strnad, W. Barfuss, J. F. Donges, and J. Heitzig, “Deep reinforcement learning in world-earth system models to discover sustainable management strategies,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, no. 12, p. 123122, Dec. 2019, doi: .

进一步的调查考虑了应该在重播缓冲区中存储哪些经验的问题，这可能是克服此问题的第一步。