# Iseec

## 模拟人类-自然系统相互作用对 21 世纪变暖的影响

大多数关于气候减缓的研究都涉及社会应该做些什么的问题。这些规范性的方法产生了一些理想的解决方案，如到 205 年实现碳中和 1 ，减少短期气候污染物（SLCPs），从空气中提取多达一万亿吨的二氧化碳等。这些规定性的解决方案对指导减缓政策大有帮助。然而，现实情况是，二氧化碳排放量的增长和全球变暖有增无减 2 。问题的部分原因在于，理想的减排目标并不总是充分考虑到人类与自然系统之间的相互作用。这种相互作用阻碍了能源和经济系统的重新设计 3,4 。

许多人认为，现在是采取新方法的时候了 5,6 ，这种方法明确允许自然系统与社会系统之间的相互作用 7-13。已考虑的人与自然相互作用的例子包括：社会对所感知的气候风险的反应，如为减缓政策提供支持和改变个人行为以减少个人温室气体排放 7；金融和其他宏观经济主体在维持经济发展和形成向低碳能源过渡的动力方面的作用 9；社会向稳定气候转变过程中的正临界点的作用 10；气候风险对能源生产经济学的负反馈 12；传统技术对碳减排的限制 10 和气候变化 11。人类与自然系统耦合的另一个最新进展是模拟气候变暖、土地利用和土地覆盖政策之间的反馈 14,15，这为综合地球系统模型铺平了道路 16。

综合评估模型（IAMs）9 的使用日益增多，对政府间气候变化专门委员会（IPCC）第 III 工作组最近的报告 17 中提出的气候减缓行动产生了重大影响。然而，大多数综合评估模型仍处于将自然和人类系统完全结合起来的早期开发阶段。正如 Peng 等人 18 所总结的，"综合可持续管理必须真正以人为本"。例如，最近一篇关于人地综合系统建模的综述 19 指出，"这些研究没有探讨许多潜在的人地系统反馈"，"没有一项研究关注能源与气候之间的相互作用"。本研究试图填补这一空白，重点关注能源与气候之间的相互作用

# 1-MDP--多剂+水

## 评估非稳态环境下适应性灌溉对水资源短缺的影响--一种多代理强化学习方法

自 1955 年哈佛水计划启动以来，水资源管理领域一直致力于为复杂的水系统提供切实可行的解决方案（Reuss，2003 年）。然而，人类（需求）系统和自然（补给）系统的不确定性和非稳定性阻碍了这些努力（Brown 等人，2015 年；Cosgrove & Loucks，2015 年；Herman 等人，2020 年）。为解决这一问题，出现了几种规划方法，可分为三类：动态规划、基于风险的规划和稳健规划。第一类是动态规划，寻求一套能产生最佳结果的决策规则（Castelletti 等人，2010 年；Harrison，2007 年；Higgins 等人，2008 年）。基于风险的规划旨在生成不同风险水平下的解决方案，并探索风险与管理目标之间的权衡（Borgomeo 等人，2018 年；Hall 等人，2020 年；Lund，2002 年），而稳健规划则寻求在各种未来条件下都稳健（可接受）的战略（Lempert & Collins，2007 年）。这些类别并不相互排斥，因为已经提出了几种混合方法。例如，动态适应政策路径将稳健决策纳入动态规划，以确定适应的时机和行动（Haasnoot 等人，2013 年；Kwakkel 等人，2015 年，2016 年）。多目标稳健决策（Many-objective Robust Decision Making）探讨了稳健的规划备选方案、在一系列未来条件下表现良好的计划以及管理目标（包括风险）之间的权衡（Kasprzyk 等人，2013 年；Watson & Kasprzyk，2017 年；Yan 等人，2017 年）。另一个例子是基于风险的随机规划，它通过考虑规避风险以避免不良后果（Piantadosi 等人，2008 年），以及边做边学以适应性地改进决策（Hung & Hobbs，2019 年），从而加强了动态规划。然而，这些研究采用的方法往往没有考虑到人类用水户（利益相关者）与自然供水之间的积极互动。水系统本质上是复杂的人与自然系统，因为利益相关者之间以及与环境之间都会产生相互作用。忽视或简化利益相关者之间的动态关系以及利益相关者的适应行为，可能会导致结论有失偏颇。

研究表明，纳入利益相关者的目标（Gold 等人，2019 年；Hadjimichael 等人，2020 年；Quinn 等人，2017 年）以及考虑利益相关者的认知信念和价值观（Glynnet 等人，2018 年；Moallemi 等人，2020 年）对于管理复杂的人类-自然系统至关重要。基于代理的建模（ABM）起源于人工智能领域，在 2000 年代末被引入水资源领域 （Berglund，2015 年）。从那时起，ABM 开始用于评估人类系统动态及其对水系统的影响（例如，Al-Amin 等人，2018 年；

Berglund，2015；Ng 等人，2011；Yang 等人，2009）。ABM 是一种分布式、自下而上的规划方法，用于了解人类对系统性能的影响。ABM 中的代理是与其他代理和相关系统（虚拟环境，如水资源模型）互动的对象。基于代理的建模（ABM）起源于人工智能领域，于 2000 年代末引入水资源领域（Berglund，2015 年）。从那时起，ABM 开始用于评估人类系统动态及其对水系统的影响（例如，Al-Amin 等人，2018 年；Berglund，2015 年；Ng 等人，2011 年；Yang 等人，2009 年）。ABM 是一种分布式、自下而上的规划方法，用于了解人类对系统性能的影响。ABM 中的代理是一个与其他代理和相关系统（虚拟环境，如水资源模型）交互的对象。

的集体行为，但可能并不代表用水户个人的决策。总之，本文的贡献包括(a) 研究人类适应性对非稳态水资源系统影响的建模方法，重点关注人类认知方面；(b) 模拟农业用水户适应性用水决策的 RL 算法，其中包含额外的水供应信息。此外，与 Hyun 等人（2019 年）相比，RL-ABM 参数能更好地描述代理的认知过程，从而提供代理对环境变化的反应信息。本文的其余部分安排如下。第 2 节介绍了农业养殖用水户适应性政策的建模框架以及 ABM-CRSS 耦合。第 3 节介绍了案例研究。第 4 节展示了研究结果，第 5 节进行了讨论。最后，我们在第 6 节中提出了结论和结束语。

Berglund（2015）总结了早期 ABM 的发展，并对水资源规划中两种不同的代理类型进行了案例研究：反应型代理根据行为规则对环境信号做出反应，而主动型代理则追求优化目

标的策略。前者也被称为描述性模型，后者被称为规范性模型（Smith，1991 年）。最近，人们提出了几种 ABM 框架，以解决复杂的人类-自然系统的自然不确定性和非平稳性问题。Giuliani 等人（2016 年）提出了一个规范性 ABM 框架，用于研究气候变化下农业水系统的共同演化。Hyun 等人（2019 年）和 Yang 等人（2020 年）应用心理模型模拟了不确定条件下农民的灌溉决策。Al-Amin 等人（2018 年）将描述性 ABM 与地下水模型相结合，评估了在未来气候情景下有多个城市的流域的限水方案。Castilla-Rho 等人（2015 年）开发了一个类似的 ABM-地下水模型框架，重点是参与式建模。一般来说，规范性人工智能模型假定理性，不考虑人类的认知活动（如学习和风险态度），而描述性人工智能模型只模拟编码行为规则，当环境发生变化时，可能会错误地反映人类的反应。我们相信，强化学习（RL）算法至少可以部分解决这一研究空白。

RL 是机器学习的一个领域，在这个领域中，智能代理可以学习并改进其 决策，以优化长期回报（Sutton & Barto, 1998）。根据问题的具体情况，RL 方法既可以是搜索最优策略（RL 术语中指导代理行动的一组规则）的学习算法，也可以是模拟人类自适应行为的决策模型（Seo & Lee, 2017; Sutton & Barto, 1998）。前者在 RL 界被称为学习方法，后者被称为规划方法（Sutton，1992 年）。RL 有两个基本特征：试错搜索和延迟奖励，这使代理能够通过与环境的交互来调整自己的策略。自 RL 被引入水资源领域以来，它已被广泛应用于水库的优化运行（Castelletti 等人，2010 年；Dariane & Moradi，2016 年；Lee & Labadie，2007 年；Madani & Hooshyar，2014 年；Rieker & Labadie，2012 年），但在大坝选型（Bertoni 等人，2020 年）以及水资源和自然资源分配（Bone & Dragićević，2009 年；Ni 等人，2014 年）方面有少数例外。然而，尽管 RL 越来越受欢迎，但这些应用只关注 RL 在静态环境中寻找最优策略的学习方面。

本文侧重于 RL 的规划方面，提出了一个 RL-ABM 框架，使代理（即案例研究中的农业用水户）具备适应不断变化的水系统的能力。代理的决策是向水资源管理部门提交的水量请求，而水系统则假定受到气候变化和用水量增长的影响。RL 算法植根于心理学和认知科学，与上述方法相比，它在模拟人类通过与环境互动而不断变化的信念和策略方面具有优势。这一特点对于模拟人类在非稳态系统中的反应至关重要。此外，RL 的参数与人类的认知活动相关，可用于描述代理人的行为特征（例如，根据水系统发出的环境变化信号进行的分水模式）。

RL-ABM 框架适用于美国科罗拉多河流域（CRB），作为一个说明性案例研究。科罗拉多河流域是美国西部和墨西哥最重要的水源地之一，由于近期的干旱和气候变暖，该流域正面临着越来越大的用水压力（美国垦务局，2012 年）。科罗拉多河仿真系统（CRSS）是美国垦务局（USBR）为科罗拉多河流域开发的水资源管理模型，用于水库运行和政策评估（美国垦务局，2007a；Zagona 等人，2001），本案例研究采用该模型作为虚拟环境，与 RL-ABM 相结合，评估水系统对动态农业用水需求的响应。RL-ABM 代理（CRSS 中的农业用水户）可以是单个农场、灌溉沟渠、灌区、美国部落用水户或一组农业实体。因此，代理只代表该群体中用水户的集体行为，但不代表用水户个人的决策。

总之，本文的贡献包括(a) 研究人类适应性对非稳态水资源系统影响的建模方法，重点关注人类认知方面；(b) 用于模拟农业用水户适应性用水决策的 RL 算法，其中纳入了额外的水资源可用性信息。此外，与 Hyun 等人（2019 年）相比，RL-ABM 参数能更好地描述代理人的认知过程，从而提供代理人对环境变化的反应信息。

本文的其余部分安排如下。第 2 节介绍了农业用水户适应性政策建模框架和 ABM-CRSS 耦合。第 3 节介绍了案例研究。第 4 节展示了研究结果，随后在 第 5 节进行了讨论。最后，我们在第 6 节中提出了结论和结束语。

# 基于强化学习的适应气候变化战略：应用于沿海洪水风险管理

要有效地适应气候变化，就必须在复杂、相互关联、具有多重不确定性的系统中制定强有力的政策并进行公共投资。因此，规划者需要能够根据不断变化的现实情况调整计划（1）。动态响应、灵活的适应战略通常优于僵化、固定的战略，因为它纳入了不断发展的风险和不确定性知识。首先，可以规划分阶段的投资，促进成本相对较低的初期行动。其次，动态投资可根据未来的意外状况调整行动或计划，避免灾难性的失败。第三，在当前决策中可以考虑未来可能采取的行动，以避免过高估计终生风险。

灵活适应框架被称为 "适应路径"（2，3）、"动态适应"（4，5）或 "实际选择分析"（6，7）。为模拟这些政策框架，已开发了几种分析方法（8 - 10）并用于评估适应措施的效益和成本，但这些方法并未考虑灵活适应的全部潜力。表 1 将应用于环境政策设计的定量方法按其能力分类：a) 制定动态政策；b) 纳入观测数据；c) 在当前决策中系统地考虑未来观测和战略调整。

(a)类方法可以设计出随着时间推移的动态决策路径。例如，参考文献例如，参考文献 13 使用动态规划（DP）方法（一种经典的连续决策框架），根据目前对未来气候的预测，估算出海岸保护的最佳路径。参考文献 14 参考文献 14 采用启发式算法，随机生成成千上万条潜在的海岸保护路径，并选择更好的路径。这些启发式算法提高了 DP 处理多步决策维度诅咒的能力（14)。然而，这些方法假定信息基础是静态的，不能直接解决灵活政策设计的一个关键优 点：灵活政策设计的一个关键优势：学习能力，以及在收集到外源信息后更新和改进决策的能力。

(b)类方法可以设计出适应新观测结果的动态决策路径。在极端情况下，它们会考虑到新信息可能会使科学信念偏离正确的气候结果，这种现象被称为负学习（21）。具体来说，贝叶斯动态规划（BDP）方法（15，16）将新的观测数据和预测纳入最优路径的估计中，从而扩展了传统的 DP 方法。虽然贝叶斯动态编程 可以将观察和学习纳入当前决策，但却不考虑未来的学习和更新，因此可能无法准确估计需要应对的终生风险。换句话说，未来决策调整的可能性会影响当前决策的最优性。

决策树或实际选择方法通过搜索情景树来生成灵活的计划，可以克服这一局限性（7，17，18）。然而，实际方案方法涉及到一棵事件树，其中的方案随着政策路径中时间步骤总数的增加而呈指数增长。只有在潜在解决方案和情景数量有限的情况下，真实选项分析才是可行的。直接政策搜索（DPS）方法（19，20）通过将每个时间步骤的决策建模为观察结果的特定函数，并通过模拟优化函数参数，从而降低了计算成本。因此，错综复杂的随机顺序决策被视为一个参数优化问题。尽管 DPS 方法计算效率高，但在适应性气候决策中仍可能无法实现真正的最优化。

RL 是机器学习的一个领域，涉及代理在不断变化的环境状态下应如何行动，以最大化其累积回报（22）。RL 方法系统地纳入观察结果，考虑未来的结果和反应，并在连续的未来环境心理状态范围内提供政策设计。此外，为了提高数值效率，RL 还可以进行各种近似（例如，在描述状态和奖励时）。RL 已在国际象棋（23，24）、自动驾驶（25）和机器人控制（26）等领域取得成功，并已被用于处理具有较大决策空间的连续环境决策，包括电力存储

( 27 ) 和水资源管理 ( 28 ) 。然而，它还未被用于解决气候变化风险管理中的不确定性问题。我们研究了 RL 应用于适应性气候决策的潜力和性能。从广义上讲，我们研究了系统学习和更新在气候适应中的价值。

为了说明这一点，我们采用 RL 方法来模拟应对沿海洪水风险的适应战略。潜在的沿海适应策略包括："保护"，如在沿海岸线修建海堤或在高地修建堤坝；"适应"，如改造建筑物（通过激励措施和保险法规加以鼓励）；以及 "撤退"，或从危害中撤出（可通过补贴 "买断"加以鼓励） ( 29)。我们考虑了热带气旋（TC）和海平面上升（SLR）的预测变化，热带气旋可能会在气候变 化的情况下引起更高的风暴潮（29-33），而海平面上升一直是并将继续是沿海洪水增加 的主要原因。然而，未来海平面上升的预测具有巨大和深刻的不确定性，特别是与人类排放和冰盖物理相关的不确定性，目前阻碍了最佳风险缓解战略的建模 ( 29 , 34 - 38 ) 。我们开发了 RL 方法，以确定美国纽约市曼哈顿的最优沿海风险减缓策略（包括适应性海堤建设， 以及涉及撤出、改造和筑堤的组合策略），并将 21 世纪持续的 SLR 观测纳入其中（材料与方法）。RL 通过状态和奖励近似方法有效地处理了计算成本（随着 SLR 情景数量和决策更新时间分辨率的增加，传统算法的计算成本呈指数级增长）[材料与方法；( 39 )]。

我们以沿海洪水风险管理为重点，评估了 RL 在更广泛的 气候适应战略优化框架中的有效性。通过与 DP、BDP 和 DPS 的比较，我们发现 RL 在制定灵活的战略以最大限度地降低成本和尾端风险方面具有优势。当实际气候条件与最初设想相差甚远时，RL 框架还能最大限度地减少 "遗憾"。这些结果凸显了持续学习和系统适应在应对气候项目中的巨大不确定性方面的重要性，以及 RL 在模拟最佳气候适应策略方面的潜力。

# Ste

# 到 2050 年稳定地球气候的社会临界动力

防止危险的气候变化及其破坏性后果是人类的一项决定性任务(1, 2)。它也是实现可持续发展不可或缺的先决条件（3, 4）。根据《巴黎气候协定》（5）的规定，将全球升温控制在 1.5 °C，这在科学上意味着到本世纪中叶，世界能源和运输系统、工业生产和土地使用将完全实现净去碳化。Rockström 等人（6）在他们的 "快速去碳化路线图 "中强调，要实现这一目标，需要快速增加零碳能源在全球能源系统中的比例，同时可能还要大力加强陆地碳汇试验。在一种设想方案中，能源系统中的零碳比例在未来几十年中每 5 到 7 年翻一番（6）。尽管经过数十年的国际谈判努力，目前碳排放量仍以每年 0% 到 2% 的速度上升，因此，碳排放量需要转向每年快速下降 7% 甚至更多。这些减排速度甚至将远远超过 20 世纪大规模社会经济危机时期的减排速度，如第二次世界大战和共产主义崩溃时期（图 1）。

在这里，具有历史决定意义的问题是，能否以及如何共同实现如此快的部署速度。目前的低碳能源部署速度与所需的转变相适应，但由于政治和经济决策固有的僵化性（7、8）以及新的技术需求（9、10），预计在扩大规模时会遇到相当大的阻力。尽管越来越多的国家已经引入或致力于引入碳定价，但碳定价所涵盖的举措仅占 2017 年全球温室气体排放量的 15%（11），迄今为止仅推动了微弱的减排（12）。越来越多的人认识到，仅靠一切照旧的技术进步和碳定价不可能实现温室气体排放量的快速和大幅减少（13）。

与此同时，来自不同科学领域的证据表明，在自然（14-16）、社会经济（17-20）和社

会生态系统（SES）（21，22）的某些临界条件下，可以观察到快速的变化率。人们越来越关注临界动态的概念，将其视为此类破坏性 系统变化背后的非线性机制。Milkoreit 等人（23）根据对社会-生态临界点研究的综述，提出了社会临界点（STPs）的通用定义，即 "在一个社会-生态系统中，一个微小的量变不可避免地引发社会-生态系统中社会部分的非线性变化，这种变化由自我强化的正反馈机制驱动，不可避免地、往往不可逆转地导致社会系统的质变"。历史上有一些例子表明，动态的社会传播效应会导致小干预的大规模自我放大：例如，马丁-路德（Martin Luther）一个人的著作，通过新出现的印刷技术注入到准备好进行这种变革的公众中，引发了新教教会在世界范围内的建立（24）。气候政策领域的一个例子是，为激励可再生能源生产的增长，引入了关税、补贴和强制措施。这导致了以相互促进的市场增长和预期技术成本改善为形式的实质性系统反应（25, 26）。

在本文中，我们研究了一些潜在的去碳化 "社会临界要素"（STEs）（27, 28），它们代表了地球社会经济系统的特定子领域。这些子系统的临界点可由 "社会临界点干预措施"（STIs）触发，从而促使世界系统迅速过渡到人为温室气体净零排放状态。本研究报告的结果基于在线专家调查、专家研讨会和广泛的文献综述（SI 附录）。

在本文中，我们研究了一些潜在的去碳化 "社会临界要素"（STEs）（27, 28），它们代表了地球社会经济系统的特定子领域。这些子系统的临界点可由 "社会临界点干预措施"（STIs）触发，从而促使世界系统迅速过渡到人为温室气体净零排放状态。本研究报告的结果基于在线专家调查、专家研讨会和广泛的文献综述（SI 附录）。

# 1+RL+ABM

# Stand2019-ays

# 在 WorldEarth 系统模型中进行深度强化学习，发现可持续管理策略

I.引言

为确定实现全球可持续性的途径而投入的努力需要考虑到社会经济和社会文化世界与地球生物物理之间的重要反馈回路

1,2 这些途径可能需要新颖的、尚未发现的、多层次的政策，从地方到全球范围，对这一耦合的世界-地球系统进行治理，以实现安全和公正的运行空间。3,4 为努力实现安全和公正的运行空间，联合国的政策制定者在 17 个可持续发展目标（SDG）5 的决议和《气候变化

巴黎协定》6 的通过过程中，就全球政治合作以实现可持续的未来达成了一致。安全和公正的操作空间基于一套生物物理地球边界（以气候变化或生物圈完整性丧失 等维度为基础），Rockström 等人在参考文献 3、4、7 和 8 中对其进行了阐述，Raworth 对其进行了社会基础（如减贫）的扩展。地球系统模型领域开发计算机模型，以展示实现可持续未来的可能途径。然而，如何确定和描述地球边界内和社会基础之上的具体轨迹，仍然是一个需要不断努力研究的问题。

在本文中，我们假定该问题具有以下基本结构，并在全球范围内加以考虑：一个抽象的单一决策者与一个动态的（多数情况下是非线性的）环境相互作用，在一定范围内找出可持续的轨迹。综合评估建模（IAM）通过优化社会福利函数来解决这一问题，以估算可持续管理战略的设计 12。

13,14 为了确定 IAM 的路径，经常使用 GAMS15 等数值求解器。然而，这些 IAM 模型高度依赖于优化目标函数的选择。在许多情况下

16 作为另一种方法，最优控制理论(OCT)可用于解决动态系统应保持在某些约束条件下的问题。在这些系统中，最优控制理论试图通过优化特定的目标函数来确定某个控制变量的最优选择 17。在地球系统模型中，最优控制理论的重点是气候调节器的设计及其对气候调节的影响 18,19。在这一领域，识别轨迹的问题通常是通过依赖于"..... "的方法来解决的。

20 然而，由于维度诅咒的存在， ，它并不能很好地适用于变量数量较少的系统。

强化学习（RL）22 的使用也可被视为世界-地球系统模型中智能决策的一种可行方法 23。然而，与前述方法不同的是，RL 并不是通过数值求解优化问题来发现解 决方案，而是通过探索和利用过去的经验，在奖励的引导下进行动态搜索。

函数。然而，主要用于经典 RL 解法的表格法无法直接应用于本文所关注的系统，因为在世界-地球系统模型中，我们大多使用连续的状态空间。

上述所有方法的共同点是，随着环境复杂性的增加，这些方法都会达到极限。然而，深度强化学习（DRL）24 算法已被证明能在其他高度复杂的环境中令人惊讶地检测出解决方案。尽管早在 1995 年就有报道称利用神经网络作为非线性函数近似器进行了首次成功的强化学习实验 26 ，但 DRL 算法直到 2013 年才取得突破性进展 24,25。27,28 这种方法成功的关键在于将 Q 学习 29 、神经网络 30 和经验重放 31 结合在一起。34 由于 DRL 普遍适用于各种环境，因此被广泛应用于各种领域，如计算机集群的资源管理、35 化学反应优化、36 象棋和围棋等抽象策略游戏、32,33 自动驾驶、37 尤其是机器人技术。

由于 DRL 具有广泛的适用性，我们提出了一个框架，将 DRL 作为一种既强大又易于使用的工具，在地球系统模型中对轨迹进行有效的识别和分类。作为概念验证，我们在各种风格化的世界-地球系统模型中使用了我们的 DRL 框架。2,42 这些模型旨在研究人类世中人类与自然的共同演化动态。强化

43-45 但据我们所知，目前还没有将 DRL 应用于地球系统模型的方法。我们相信，这种方法将开启迄今为止尚未使用的可能性，发现迄今为止未知的管理问题。

在尊重世界社会基础的同时，将地球系统保持在地球范围内的战略。最近，人们概述了如何利用机器学习技术解决人为气候变化相关问题的各种方法 46。

II.方法

# 强化学习能否为决策者提供支持？综合评估模型初步研究

**1 引言**

气候是一个高维度的动态系统，具有很强的相互依存性。

所有这些因素相互作用，产生了高度非线性的反应和行为。气候在很大程度上也受制于人类行为--另一个极其复杂的系统--因此，现在有必要从社会-气候的角度来推理气候变化[Moore 等人，2022 年]。为了在面对极端后果时找到某种解决方案，政策制定者和咨询小组采用了*综合评估模型*（IAMs），这是目前最先进的气候变化模型，它将人类发展知识（如经济理论）与生态学和地球物理学等行星科学结合在一起[Parson 和 Fisher-Vanden，1997]。从计算的角度来看，探索和分析用于大规模评估的 IAMs 的特性（例如，衡量与真实世界的保真度）[Pörtner 等人，2022 年]通常是难以实现的，这导致研究人员实施拙劣的简化假设并降低其有效性[Asefi-Najafabady 等人，2021 年]。较小的 IAM 模型旨在通过采用较少的状态变量和较简单的动力学集合提供一种替代方案，使其易于数学探测和分析[Kittel 等人，2017 年；Nitzbon 等人，2017 年]。文献通常使用 ODE 求解器来探索这些模型；但最近[Strnad 等人，2019]的研究表明，可以将它们作为标准强化学习（RL）[Sutton 和 Barto，2020]的环境，并使用训练好的策略来探索模型。这些模型可用于理解系统，并为改进更复杂的 IAMs 或甚至我们对气候变化政策的理解提供更高的洞察力。

在此基础上，我们测试了更多的 RL 算法、奖励函数以及不同的实验设置。除其他外，我们还表明：(a) 现代 RL 可以在这种环境下利用各种奖励函数学习有效的政策；(b) 不同的代理和奖励函数会产生大量不同的解决方案，从而以不同的方式探索 IAM；(c) 在设计奖励函数时必须小心谨慎，因为对于不同的初始化点，奖励函数在达到理想状态方面显示出不同的成功率；最后--(d) RL 可以帮助我们更深入地了解所应用模型的特性和局限性。

**2 AYS 环境、RL 奖励函数和代理**

# 利用强化学习探索气候变化政策

导言

1.1 动机

1.1.1 人类的安全空间

根据政府间气候变化专门委员会（IPCC）的最新报告[2]，"人类的影响已经使大气、海洋和陆地变暖，这一点是毋庸置疑的。根据政府间气候变化专门委员会（IPCC）的最新报告[2]，"大气层、海洋、冰冻圈和生物圈已经发生了广泛而迅速的变化。

过去 10-12 千年的全新世气候相对稳定，见证了人类的大部分进步：从最初的农业社会到工业革命的开始。全新世之前的地质时期被称为更新世，俗称冰河时期[3]。最近，由于人类对气候的影响，科学家开始将人类工业化时代命名为 "人类世"[4, 5]。这个拟议的地

质时代的特点是人类对环境和气候的影响。

临界点指的是某些阈值，在这些阈值上，微小的扰动就能使系统发生剧烈变化[6]。对于气候而言，不同的临界点会导致整体气候平衡发生巨大变化。科学家认为，全新世的气候具有微妙的平衡[6, 7]，不应受到干扰，以免给地球上的生命带来难以抵御的灾难性后果。据了解，气候是复杂的社会-生态系统耦合的一部分，也具有适应性[8]。因此，人类采取的任何行动都会产生许多不可预测的连锁效应[7]。因此，在制定与气候有关的政策时，必须考虑到这一耦合系统[9]。

为减轻未来气候变化的影响，科学家建议保持在 "行星边界"[10, 11]之内。这些界限定义了与重要临界点相关的可测量数量的极限，例如将升温幅度保持在高于工业化前水平+2◦C以下。跨越行星边界会在气候系统中引发反馈回路（或 "滚雪球 "效应），导致气候迅速发生不可控制的不稳定[7]。

1.1.2 作为控制问题的气候变化

IPCC 使用 SSP（共享社会经济路径）来预测人类不同行为对气候可能产生的影响。描述了从大量减少温室气体排放到大量增加温室气体排放的五种途径。每种可能的行为都会对地球和地球上的生命造成影响。要预测这些途径，需要建立具有极高不确定性的模型，因为模拟地球的所有复杂性是不可行的。IPCC 并不给出每种情景的确切概率，而是使用 "极高概率"、"低概率"、"中等概率 "等术语进行预测。这些 SSP 由综合评估模型 (IAM) 生成。这些模型涵盖了气候、生物和经济等不同领域的广泛科学知识[12]。

我们知道，天气系统以及气候是一个动态系统 ，初始条件的微小变化都可能导致随着时间的推移出现惊人的不同结果[13, 14]。动力系统是确定性的。如果初始条件完全已知，我们就能准确预测结果。从技术上讲，拥有完全准确的数据和模型是不可能的，因此气象学家会在不同的初始条件下进行多次模拟，然后求取平均值，这也被称为集合模型。这也是 IPCC 生成 SSP 的方法，他们使用来自综合评估模型联合会的多个 IAMs 的信息，然后求取平均值。在动力系统中，只要时间足够，这种方法可以产生非常广泛的答案，这就是为什么对未来更远的天气预测往往不准确的原因。

气候的定义是 "某一特定地方通常出现的一般天气条件 "1。

.全球气候就是地球的总体天气模式。全球气候比长期天气更容易预测，因为它是长期天气的平均值。然而，由于人类的影响，这一系统正迅速受到扰动[15]。在动态系统中添加扰动会使系统更加难以预测。扰动可能包括人类碳排放的变化，这会产生连锁和不可预测的影响。理想情况下，如果我们能准确预测，就能准确控制。我们希望控制气候，使其不超出地球范围，从而避免对人类造成最严重的灾难性影响。政策制定者和世界性组织可以通过改变人类的行为来控制气候，从而实现可持续发展的未来。

将动态系统控制在一定范围内的任务称为最优控制理论（OCT），通常简称为控制[16]。这一数学分支领域研究的是通过控制系统的变量来优化系统中某些目标函数的方法。就气候而言，我们希望通过保持在地球边界内，最大限度地减少气候的不稳定性