## [1] V. Ramanathan, Y. Xu, and A. Versaci, “Modelling human–natural systems interactions with implications for twenty-first-century warming,” Nat Sustain, vol. 5, no. 3, pp. 263–271, Dec. 2021, doi: 10.1038/s41893-021-00826-z.

主要是包含实验结果的图

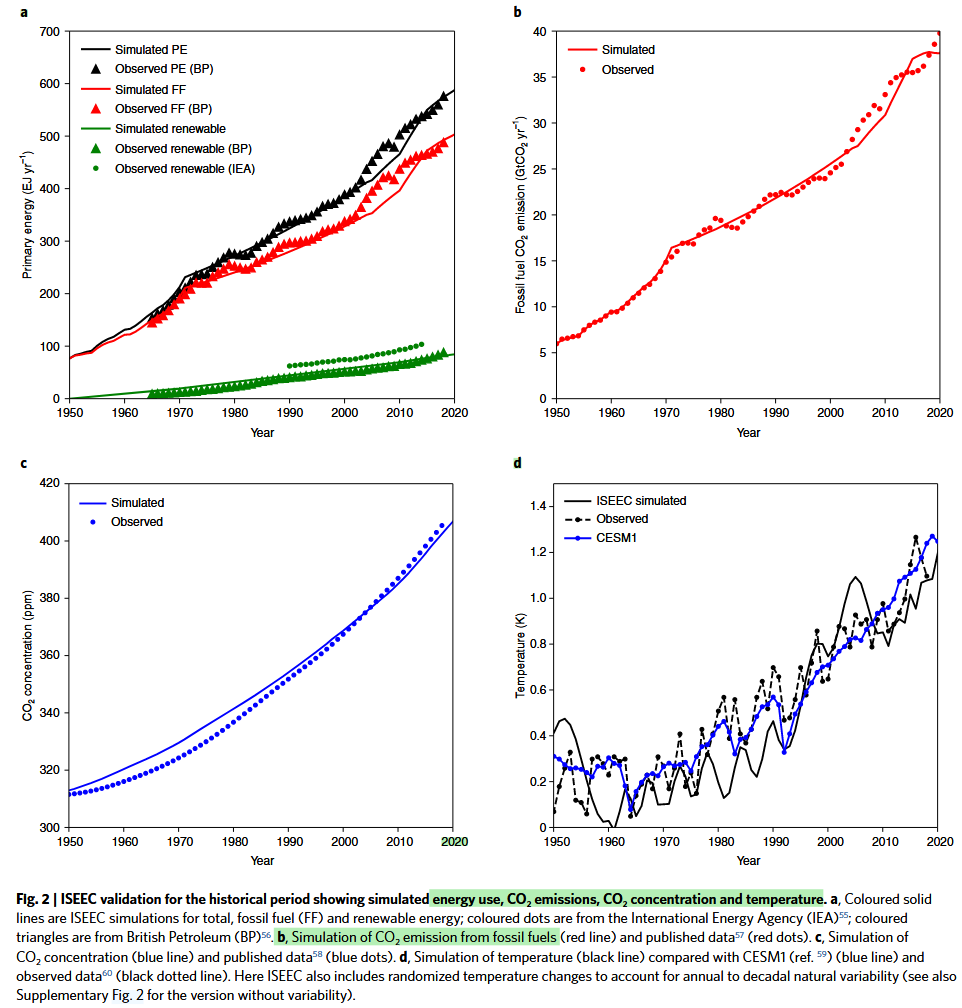


Fig. 2 | ISEEC validation for the historical period showing simulated energy use, CO2 emissions, CO2 concentration and temperature. a, Coloured solid lines are ISEEC simulations for total, fossil fuel (FF) and renewable energy; coloured dots are from the International Energy Agency (IEA)55; coloured triangles are from British Petroleum (BP)56. b, Simulation of CO2 emission from fossil fuels (red line) and published data57 (red dots). c, Simulation of CO2 concentration (blue line) and published data58 (blue dots). d, Simulation of temperature (black line) compared with CESM1 (ref. 59) (blue line) and observed data60 (black dotted line). here ISEEC also includes randomized temperature changes to account for annual to decadal natural variability (see also Supplementary Fig. 2 for the version without variability)

图2 | ISEEC验证历史时期显示模拟的能源使用，CO2排放，CO2浓度和温度。a，有色固体线是ISEEC模拟的总量，化石燃料( FF )和可再生能源；彩色圆点来自国际能源署( IEA ) 55；颜色三角形来自British Petroleum ( BP ) 56。b，模拟化石燃料的CO2排放(红线)和已发表的数据57 (红点)。c，模拟CO2浓度(蓝线)和已发表的数据58 (蓝点)。d，模拟温度(黑线)与CESM1 (参考值59 ) (蓝线)和观测数据60 (黑色点线)。这里ISEEC还包括随机温度变化，以解释年至十年的自然变化(见该版本的补充图2 )

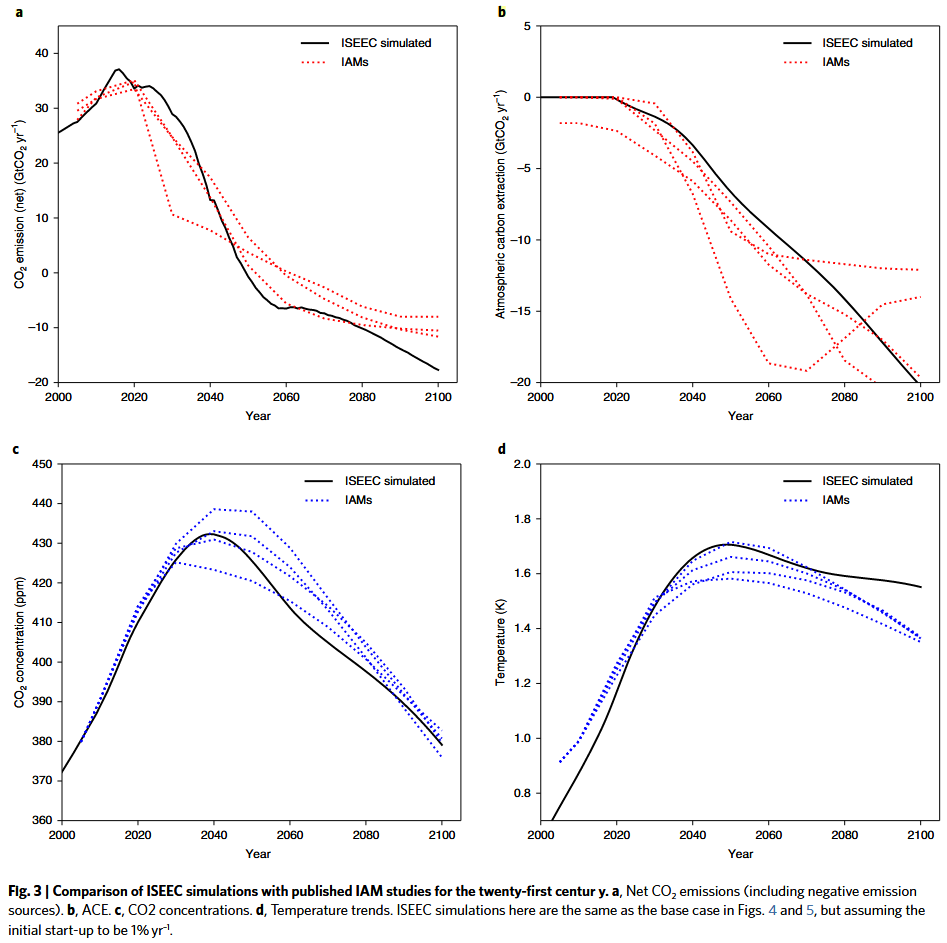


Fig. 3 | Comparison of ISEEC simulations with published IAM studies for the twenty-first centur y. a, Net CO2 emissions (including negative emission sources). b, ACE. c, CO2 concentrations. d, Temperature trends. ISEEC simulations here are the same as the base case in Figs. 4 and 5, but assuming the initial start-up to be 1% yr–1.

图3 | ISEEC模拟与已发表的IAM研究对21世纪y . a、净CO2排放(包括负排放源) . b、ACE . c、CO2浓度. d、温度变化趋势的比较。这里的ISEEC模拟与图中的基准情形相同。4和5，但假设初始开机为1 % yr – 1

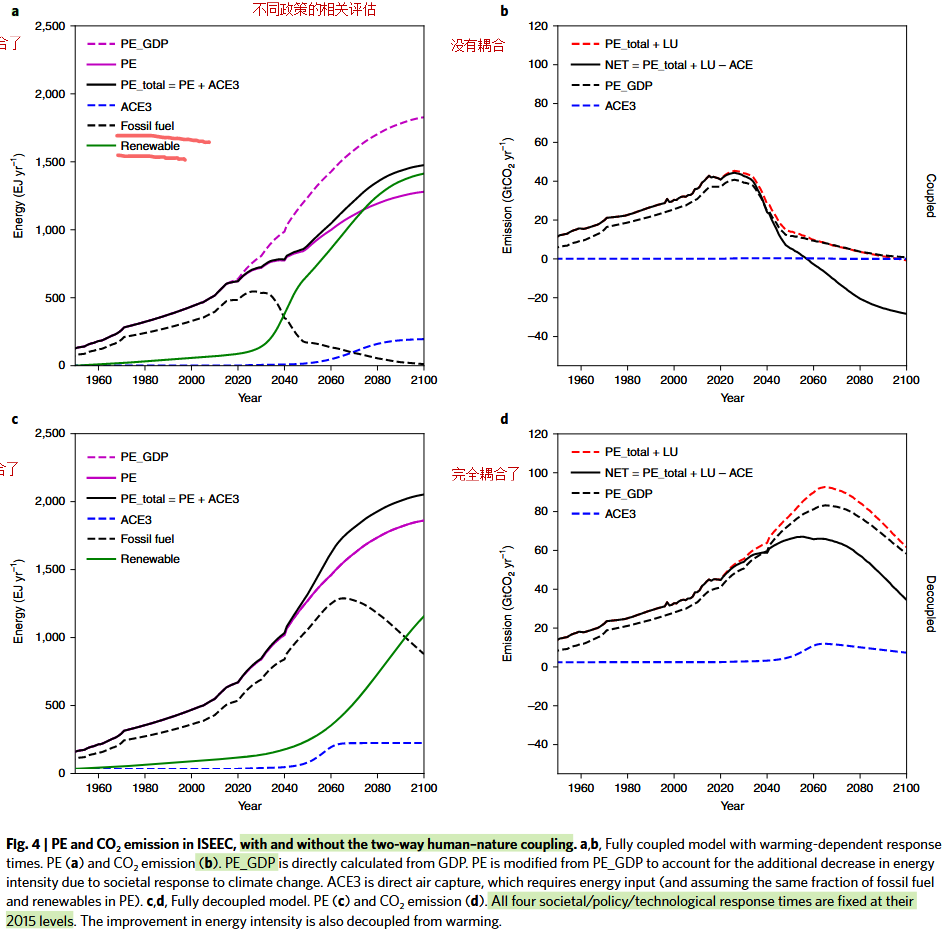


Fig. 4 | PE and CO2 emission in ISEEC, with and without the two-way human–nature coupling. a,b, Fully coupled model with warming-dependent response times. PE (a) and CO2 emission (b). PE\_GDP is directly calculated from GDP. PE is modified from PE\_GDP to account for the additional decrease in energy intensity due to societal response to climate change. ACE3 is direct air capture, which requires energy input (and assuming the same fraction of fossil fuel and renewables in PE). c,d, Fully decoupled model. PE (c) and CO2 emission (d). All four societal/policy/technological response times are fixed at their 2015 levels. The improvement in energy intensity is also decoupled from warming.

图4 | PE和CO2在ISEEC中的排放，有和没有双向的人类-自然耦合。a，b，完全耦合模型，与温度相关的响应时间。PE ( a )和CO2排放量( b )。PE \_ GDP由GDP直接计算得到。PE由PE \_ GDP修正而来，以解释由于社会对气候变化的反应而导致的能源强度的额外降低。ACE3是直接空气捕获，需要能量输入(并假设PE中化石燃料和可再生能源的比例相同)。c，d，全解耦模型。PE ( c )和CO2排放量( d )。所有四个社会/政策/技术响应时间都固定在2015年的水平。能源强度的改善也与变暖脱钩。

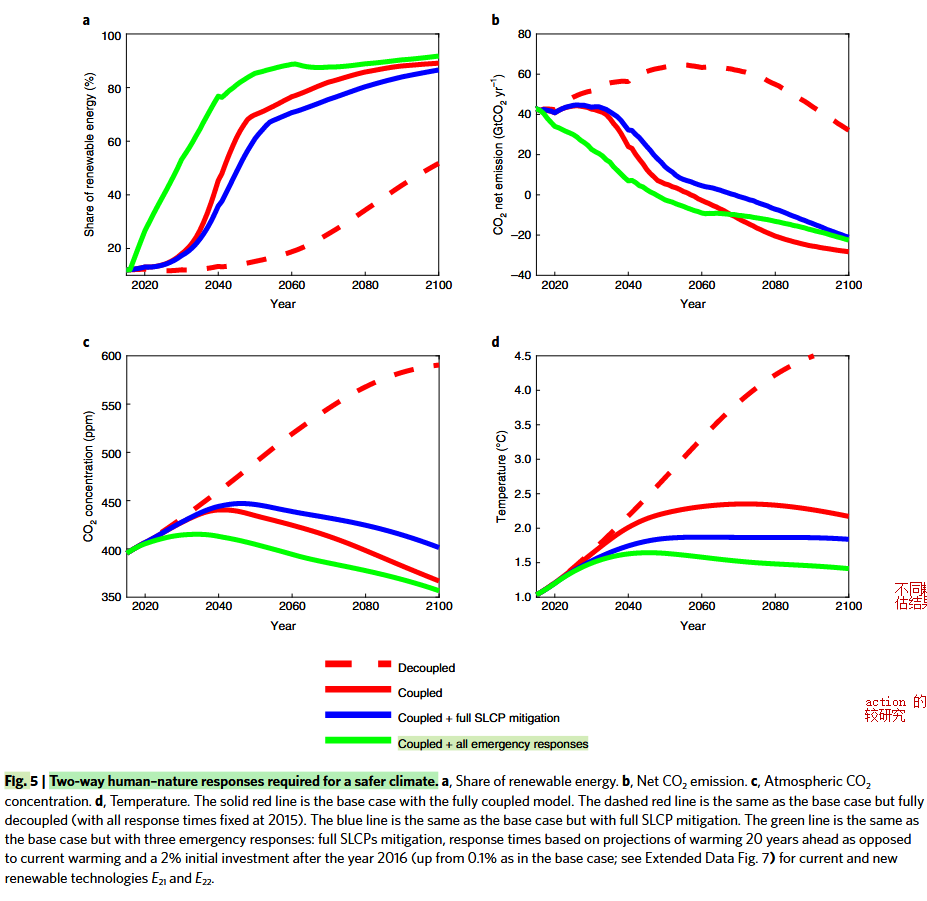


Fig. 5 | Two-way human–nature responses required for a safer climate. a, Share of renewable energy. b, Net CO2 emission. c, Atmospheric CO2 concentration. d, Temperature. The solid red line is the base case with the fully coupled model. The dashed red line is the same as the base case but fully decoupled (with all response times fixed at 2015). The blue line is the same as the base case but with full SLCP mitigation. The green line is the same as the base case but with three emergency responses: full SLCPs mitigation, response times based on projections of warming 20 years ahead as opposed to current warming and a 2% initial investment after the year 2016 (up from 0.1% as in the base case; see Extended Data Fig. 7) for current and new renewable technologies E21 and E22.

图5 |更安全的气候所需的双向人与自然的反应。a，可再生能源的份额。b，净CO2排放。c，大气CO2浓度。d，温度。实心红线为全耦合模型的基本情况。虚线与基准情形相同，但完全解耦(所有响应时间固定在2015年)。蓝线与基准情形相同，但具有完全SLCP缓解。该绿线与基准案例相同，但有三种应急响应：全面的SLCPs缓解，根据预测的变暖20年前的响应时间，而不是目前的变暖，以及在2016年后的2 %的初始投资，目前和新的可再生能源技术E21和E22。

【模型历史部分的验证部分】

Validation of human–natural systems interactions

Equations (1)–(5) were individually validated with solar and wind power data for 2000–2017 and with published projections30 for 2010–2050 (Extended Data Fig. 3). The SSM equations capture the time evolution of PE generated by solar and wind energy across three orders of magnitude of increase in renewable energy shares. The fully coupled ISEEC simulations are validated next. The comparison of simulations and data is restricted to the period from 1950 to 2020, when the historical data have sufficient accuracy (Fig. 2). ISEEC simulates within 10% the time evolution of fossil fuel

人类-自然系统相互作用方程( 1 ) ~ ( 5 )的验证分别采用2000 ~ 2017年太阳能和风能发电数据以及2010 ~ 2050年(扩展数据图3)公布的预测值30进行验证。SSM方程捕捉了太阳能和风能在可再生能源份额增加3个数量级的情况下产生的PE的时间演变。接下来对全耦合的ISEEC模拟进行了验证。模拟和数据的比较仅限于1950年至2020年期间，此时历史数据具有足够的准确性(图2 )。ISEEC对化石燃料的时间演化进行了10 %以内的模拟

and the renewable share of PE (Fig. 2a). The fossil fuel CO2 emission (Fig. 2b) is captured within 5%. The simulated CO2 concentrations (Fig. 2c) agree with observed values within 5 ppm throughout the record, and the trends agree within 2%. The steep surface warming trend (Fig. 2d) from 1970 to 2020 in the data is 0.85 °C compared with the simulated trend of 0.9 °C. The comparison also includes a three-dimensional climate model simulation used in IPCC studies (Community Earth System Model 1 (CESM1) curve). The projected twenty-first-century trends by ISEEC are compared with published projections (by four different IAMs) in Fig. 3. In both the ISEEC and the IAMs projections, CO2 emissions peak around 2020, ACE reaches 20 GtCO2 yr–1 by 2100, CO2 concentration peaks at 430 ppm by 2040 and the warming peaks at 1.7 °C around 2050 and hovers around 1.5 °C by 2100. ISEEC, like many IAMs, employs a highly simplified treatment of complex human–nature interactions. Such interactions in the

以及PE的可再生份额(图2a )。化石燃料的CO2排放(图2b )被捕集到5 %以内。模拟的CO2浓度(图2c )在整个记录范围内与观测值在5 ppm以内吻合，变化趋势在2 %以内吻合。资料中1970 ~ 2020年陡峭地表增温趋势(图2d )为0.85℃，与模拟趋势0.9℃相比，也包括IPCC研究中使用的三维气候模式模拟的(社区地球系统模型1 ( CESM1 ))曲线)。图3将ISEEC预测的21世纪趋势与已发表的预测(通过4种不同的IAM)进行了比较。在ISEEC和IAMs的预测中，CO2排放峰值在2020年左右，ACE在2100年达到20 GtCO2 yr-1，CO2浓度峰值在2040年达到430 ppm，升温峰值在2050年左右达到1.7 ° C，在2100年左右徘徊在1.5 ° C左右。与许多IAM一样，ISEEC采用了高度简化的方法来处理复杂的人与自然的相互作用。

在真实结果中的如此反映取决于本地化的社会、经济和文化因素，而这些因素在目前版本的ISEEC中被忽视了。尽管如此，ISEEC能够模拟过去和预测的能源份额、碳循环和气候趋势的总体特征。然而，由ISEEC得出的结论和建议应被认为是建议性的，而不是确定性的。

**The role of two-way energy–climate interactions**

The basic driver of climate changes in ISEEC is the tenfold growth of GDP from about US$95 trillion in 2020 to US$1,000 trillion in 2100 (Extended Data Fig. 1a). Because of the compensating fourfold decrease in the energy intensity from 2016 to 2100 (Extended Data Fig. 1b) and the additional reduction of 30% (through the rEI term in equation (1)) in energy intensity due to societal response to the warming, PE increases only twofold from 600 EJ yr–1 in 2020 to

ISEEC气候变化的基本驱动力是GDP从2020年的约95万亿美元增长到2100 (扩展数据图1a)年的1000万亿美元，增长了10倍。由于2016年至2100年(扩展数据图1b)能源强度的补偿四倍下降和社会对变暖的响应导致能源强度的额外减少30 % (通过方程( 1 )中的rEI项) )，PE仅从2020年的600 EJ yr - 1增加了两倍

1,300 EJ yr–1 in 2100 (PE in Fig. 4a). In addition, ACE increases PE by 15% to bring PE\_total to 1,500 EJ yr–1 (Fig. 4a). The gap between PE\_GDP and PE in Fig. 4a shows that decreasing the energy intensity through further improvements in energy efficiency is essential for avoiding catastrophic warming of more than 4 °C. The next energy–climate interaction initiated by the warming is a reduction in the carbon intensity of PE through societal actions to increase the share of renewables in PE. The inertia in societal actions is the basic limiting factor for how fast this transition to renewables can take place. Response times for current (E21) and new (E22) renewable technologies decrease (Extended Data Fig. 4) from 15 (for E21) to 30 (for E22) years in 2015 (at a warming of 1 °C) to 5–15 years in 2030, when the warming reaches 1.5 °C.

The faster response times accelerate the growth of renewables. Fossil fuel energy and CO2 emissions peak by 2025 (Fig. 4a,b), but CO2 concentration does not reach its peak until 2047. This 22-year lag between emission and concentration peaks is largely due to the inertia in the diffusion of renewable energy technologies (Extended Data Figs. 5 and 6) and the inertia in the carbon cycle. The warming curve takes another 28 years to bend and peaks at 2.3 °C in 2075 (Fig. 5d). This additional 28-year lag is due to a combination of two factors: the inertia in the coupled ocean– (upper 300 m) atmosphere climate system (about 15 years) and the additional primary energy (15%) used for ACE, some of which relies on fossil fuels. The total time lag between the emission curve peak and the warming curve peak is thus almost 50 years in the base case of the model (Fig. 5c,d and Supplementary Table 3).

To quantify the role of the two-way human–natural system interactions, the response times and the energy intensity ratio were made independent of the warming and held fixed at their 2016 values (Fig. 4c,d). The simulation in the decoupled case is very similar to the business-as-usual scenario of IPCC as well as those by other integrated assessment model studies. By 2100, cumulative CO2 emission since the pre-industrial era grows to 5,000 Gt and the warming exceeds 4 °C (Fig. 5d).

2100 (图4a中PE)年为1300 EJ Yr - 1。此外，ACE使PE增加15 %，使PE \_ total达到1500 EJ yr-1 (图4a )。图4a中PE \_ GDP和PE之间的差距表明，通过进一步提高能源效率来降低能源强度对于避免超过4℃的灾难性变暖至关重要。由变暖引发的下一个能源-气候相互作用是通过社会行动来减少PE的碳强度，以增加PE中可再生能源的份额。社会行动中的惯性是这种向可再生能源转变的速度的基本限制因素。当前( E21 )和新( E22 )可再生能源技术的响应时间从2015年的15年( E21 )到30年( E22 )，(扩展数据图4)从2015年的(在1℃的增温条件下)下降到2030年的5 ~ 15年，增温达到1.5℃，更快的响应时间加速了可再生能源的增长。化石燃料能源和CO2排放在2025年(图4a , b)达到峰值，但CO2浓度直到2047年才达到峰值。排放峰值和浓度峰值之间的这22年的滞后很大程度上是由于扩散的惯性可再生能源技术(扩展数据图。5和6)和碳循环中的惯性。增温曲线需要28年才能弯曲，在2075年达到峰值2.3 ° C (图5d )。这种额外的28年滞后是由于两个因素的组合：海洋- (上300 m)大气气候系统(约15年)中的惯性和用于ACE的额外一次能源( 15 % )，其中一些依赖化石燃料。因此，在模式(图5c、d及附表3)的基准情形下，排放曲线峰值与升温曲线峰值之间的总时滞约为50年。

为了量化人类-自然系统双向相互作用的作用，响应时间和能量强度比独立于增温，并固定为2016年的值(图4c、d)。解耦情况下的模拟与IPCC的"一切照旧"情景以及其他综合评估模型研究的情景非常相似。到2100年，前工业化时代以来的CO2累积排放量增长到5000 Gt，增温超过4 ° C (图5d )。

Numerous sensitivity studies were done (Supplementary Information and Supplementary Table 3) to evaluate the robustness of the results (Fig. 5) to the various model parameters and assumptions, especially warming-dependent response times and carbon intensity. The 2100 warming ranges from 1.9 to 2.7 °C (compared with the 2.2 °C of the coupled case in Fig. 5d), and the delay between peak emission and peak warming ranges from 50 to 70 years.

为了评估结果(图5 )对各种模型参数和假设的稳健性，特别是与增温相关的响应时间和碳强度，进行了大量的敏感性研究。2100年增温范围为1.9 - 2.7 ° C (与图5d中耦合情况的2.2 ° C进行了对比)，峰值排放与峰值增温之间的时滞范围为50 - 70年。

【讨论部分】

The ISEEC framework, with just GDP as the main external input, is able to reproduce historical data as well as published projections for the evolution of fossil/renewable sources of energy, CO2 emission, CO2 concentration and global warming. The simulations identify a five-decades-long delay between societal response (the emissions peak) to the warming and nature’s response (the warming peak) to the societal response, which has to be reduced to about two decades or less for limiting the warming to safer levels (<1.5 °C).

Emergency responses by society are required for limiting peak warming to 1.5 °C or less by 2100. The more than ten sensitivity studies shown in Supplementary Table 3 identify the ratio of renewable energy to the total energy (R\_1.5 C) at the time when the warming reaches 1.5 °C (around 2030 for most cases in Fig. 5) as one critical factor that distinguishes stable (<1.5 °C) from unstable (>2.5 °C) warming pathways. If R\_1.5 C is less than 0.2, the 2100 warming exceeds 4 °C, and when R\_1.5 C is larger than 0.5, the warming can peak below 2 °C.

ISEEC框架仅以GDP作为主要外部输入，能够再现化石/可再生能源、CO2排放、CO2浓度和全球变暖的历史数据和已发表的预测。模拟发现，社会对变暖的响应(排放峰值)与自然对社会的响应(增温峰值)之间存在长达5年的延迟，为了将变暖限制在更安全的水平( < 1.5℃)，必须将其减少到大约20年或更少。

社会需要采取应急措施，将升温峰值限制在2100年之前的1.5℃或更低。补充表3所示的十余项敏感性研究确定了增温达到1.5 ° C (在图5中的大多数情况下, 2030年左右)时可再生能源与总能量的比值( R \_ 1.5C )是区分稳定( 2.5 ° C )增温途径的关键因素。当R \_ ( 1.5 ) C小于0.2时，2100℃增温超过4℃；当R \_ ( 1.5 ) C大于0.5时，增温在2℃以下达到峰值

As shown by the green curve in Fig. 5 (Case 5 in Supplementary Table 3), three additional societal actions are required for limiting the warming below 1.5 °C:

1. Trusting scientific projections: ISEEC thus far assumes societal response is determined by the warming society is currently experiencing; if society has more trust in science, it could respond to projected warming 20 years into the future (Case 3 in Supplementary Table 3). (2) Boosting early investment: the second emergency response is to increase the start-up investment term, η0 (equation (3)) from 0.1% yr–1 to 2% yr–1 in 2016, when the warming reaches 1 °C (Case 4 in Supplementary Table 3). (3) Reducing SLCPs: for the portion of the SLCPs (hydrofluorocarbons (HFCs), methane, ozone and black carbon) that are not co-emitted with fossil fuels, deploy the maximum available mitigation actions (blue line in Fig. 5).

如图5 (附表3中的情形5)中的绿色曲线所示，为了将变暖限制在1.5℃以下，还需要三个额外的社会行动：

( 1 )相信科学预测：迄今为止，ISEEC假设社会反应是由目前正在经历的变暖社会决定的；如果社会对科学有更多的信任，那么它就可以对未来(附表3中的情形3)预测的20年变暖做出反应。

( 2 )增加前期投资：第二个应急响应是在2016年增温达到1 ° C (附表3中的情形4)时，将启动投资项η 0 (式( 3 ) )从0.1 % yr - 1增加到2 % yr - 1。

( 3 )减少SLCPs：对于未与化石燃料共同排放的部分SLCPs (氢氟碳化合物( HFCs )、甲烷、臭氧和黑碳)，部署最大可用缓解行动(图5中蓝色线条)。

If society adopts all of the preceding three emergency responses, the 50-year lag between emissions peak and warming peak is reduced to just 20 years (Case 5 in Supplementary Table 3), with SLCPs reduction playing a major role, and the 2100 warming can be limited to 1.4 °C (solid green line in Fig. 5c). We provide examples of the required practical responses, as also articulated in previous studies31–33.

如果社会采取上述三种应急响应措施，排放峰值与变暖峰值之间的50年滞后时间缩短至20年(附表3情形5)，其中SLCPs的减少起主要作用，2100年的变暖可以限制在1.4 ° C (图5c中实线绿色)。我们提供了所需的实际反应的例子，正如以前的研究中所阐述的那样31 - 33。

(1) A step-change in investment in clean energy infrastructure (for example, electric grids, electrifying end use and network of charging stations) and technologies (for example, hydrogen electrolysers and small modular nuclear) through national and subnational actions; massive gains in energy efficiency through clean technologies and behavioural changes; retrofitting legacy fossil fuel plants with carbon capture, utilization and storage; a step-change in investment in ACE, repurposing and sequestration

(2) Carbon pricing as a financing/investment instrument, the lack of which is a major barrier, and potential alternatives that can facilitate a faster transition

(3) Giving clean energy access to the poorest three billion who still rely on primitive solid fuels such as coal, firewood and organic waste

(4) Reducing methane emissions from the transmission of natural gas from food and other organic waste in landfills and farms; eliminating black carbon emissions from diesel vehicles and cooking with solid fuels; phasing out by 2030 the use of HFCs as refrigerants and coolants

（1）清洁能源基础设施投资的逐步改变（例如，电网，充电站的电气终端使用和网络）和技术（例如，氢电解器和小型模块化核能）通过国家和统一行动；通过清洁技术和行为改变，能效率巨大的提高；用碳捕获，利用和储存的遗产化石燃料厂改造；在ACE，重新利用和隔离的投资中的逐步变化

（2）碳定价作为融资/投资工具，缺乏主要障碍，并且可以促进更快过渡的潜在替代方案

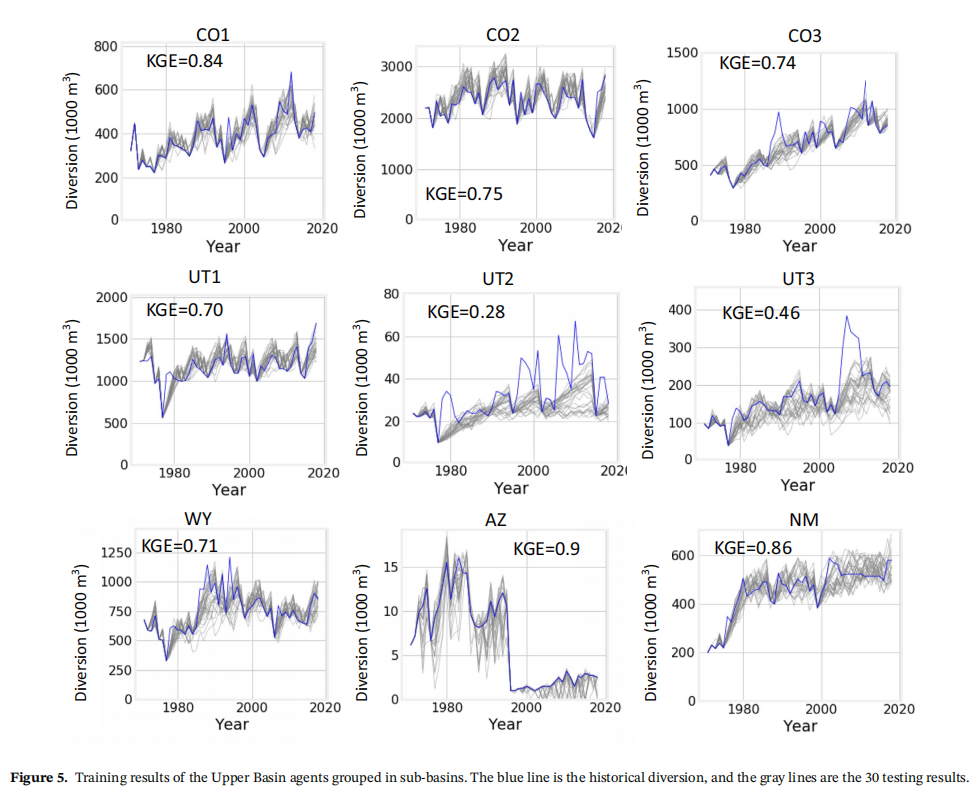
（3）使能源能提供清洁能源，从而使最贫穷的30亿仍然依赖于诸如煤炭，有机烟草，烟草，有机烟草，烟草，有机燃料，有机燃料，

（4）从垃圾填埋场和农场中的食物和其他有机废物传播的天然气从传播天然气中减少甲烷排放；消除柴油车的黑色碳排放，并用固体燃料烹饪；到2030年逐步淘汰HFC作为制冷剂和冷却剂

The simplified ISEEC architecture is intended to provide a proof of concept of energy–climate interactions triggered by societal intervention. The explicitly treated two-way coupling between human and natural systems provided insights into what is urgently needed in the design of climate solutions. Societal responses to climate risks are regionally and locally dependent, which is ignored in this first version of ISEEC. The next step is to implement ISEEC in a three-dimensional Earth system model with sectoral and national granularity for societal responses.

简化的ISEEC架构旨在为社会干预引发的能源-气候相互作用的概念提供证明。明确处理的人类和自然系统之间的双向耦合为气候解决方案的设计提供了迫切需要的见解。社会对气候风险的反应具有区域和地方依赖性，这在ISEEC的第一个版本中被忽略。下一步将在具有部门和国家粒度的三维地球系统模型中实现ISEEC，以用于社会响应。

## [2] F. Hung and Y. C. E. Yang, “Assessing Adaptive Irrigation Impacts on Water Scarcity in Nonstationary Environments—A Multi‐Agent Reinforcement Learning Approach,” Water Resources Research, vol. 57, no. 9, p. e2020WR029262, Sep. 2021, doi: 10.1029/2020WR029262.



**Figure 5.** Training results of the Upper Basin agents grouped in sub-basins. The blue line is the historical diversion, and the gray lines are the 30 testing results.

图 5.上流域代理的训练结果分组到子流域中。蓝线是历史转移，灰线是 30 个测试结果。

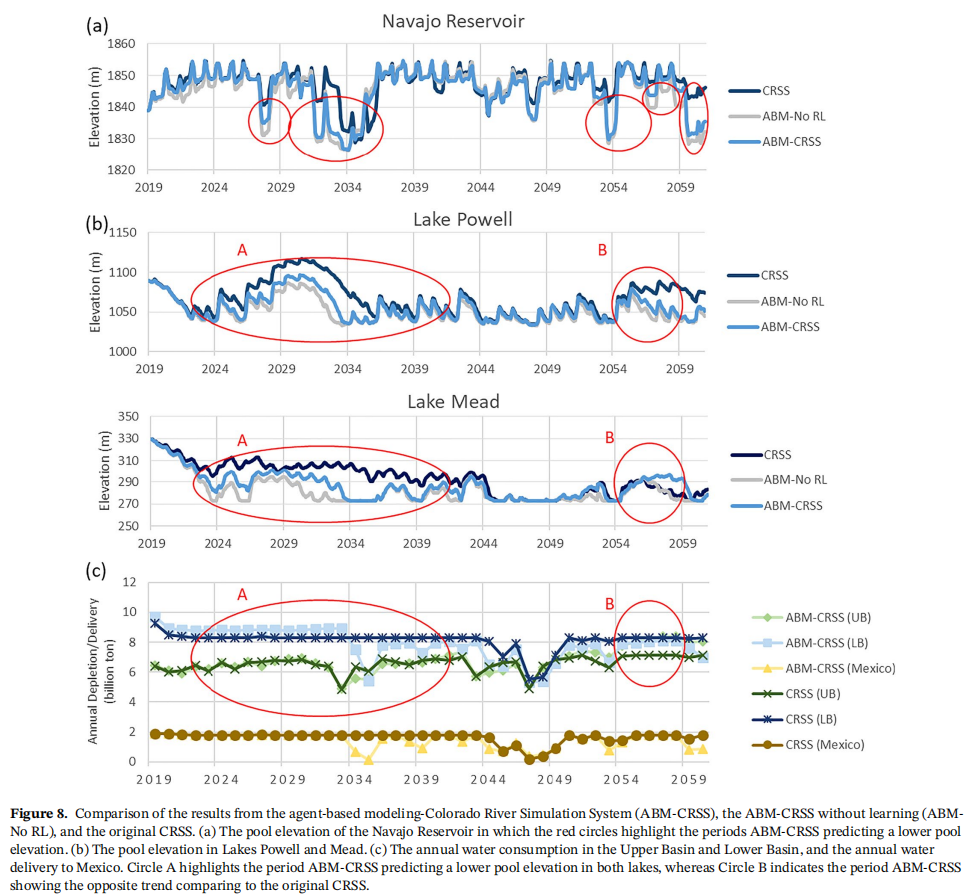


Figure 8. Comparison of the results from the agent-based modeling-Colorado River Simulation System (ABM-CRSS), the ABM-CRSS without learning (ABMNo RL), and the original CRSS. (a) The pool elevation of the Navajo Reservoir in which the red circles highlight the periods ABM-CRSS predicting a lower pool

elevation. (b) The pool elevation in Lakes Powell and Mead. (c) The annual water consumption in the Upper Basin and Lower Basin, and the annual water

delivery to Mexico. Circle A highlights the period ABM-CRSS predicting a lower pool elevation in both lakes, whereas Circle B indicates the period ABM-CRSS

showing the opposite trend comparing to the original CRSS

图8展示了基于代理的建模系统（ABM-CRSS）、无学习的ABM-CRSS（ABMNo RL)以及原始CRSS的结果对比。(a)纳瓦霍水库的水位，红色圆圈标出了ABM-CRSS预测水位较低的时期。(b)波威尔湖和米德湖的水位。(c)上盆地和下盆地的年用水量，以及每年向墨西哥输送的水量。圆A显示了ABM-CRSS预测两个湖泊水位较低的时期，而圆B则显示了与原始CRSS相反的趋势。

## [3] K. Feng, N. Lin, R. E. Kopp, S. Xian, and M. Oppenheimer, “Reinforcement learning–based adaptive strategies for climate change adaptation: An application for coastal flood risk management,” Proc. Natl. Acad. Sci. U.S.A., vol. 122, no. 12, p. e2402826122, Mar. 2025, doi: 10.1073/pnas.2402826122.

## [4] T. Wolf, N. Nardelli, J. Shawe-Taylor, and M. Perez-Ortiz, “Can reinforcement learning support policy makers? A preliminary study with integrated assessment models,” Dec. 11, 2023, arXiv: arXiv:2312.06527. doi: 10.48550/arXiv.2312.06527.

## [5] T. Wolf, “Climate Change Policy Exploration using Reinforcement Learning,” Oct. 23, 2022, arXiv: arXiv:2211.17013. Accessed: Mar. 27, 2024. [Online]. Available: <http://arxiv.org/abs/2211.17013>

## [6] S. S. Shuvo, Y. Yilmaz, A. Bush, and M. Hafen, “Modeling and Simulating Adaptation Strategies Against Sea-Level Rise Using Multiagent Deep Reinforcement Learning,” IEEE Trans. Comput. Soc. Syst., vol. 9, no. 4, pp. 1185–1196, Aug. 2022, doi: 10.1109/TCSS.2021.3122282.

## [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: [10.1063/1.5124673](https://doi.org/10.1063/1.5124673).

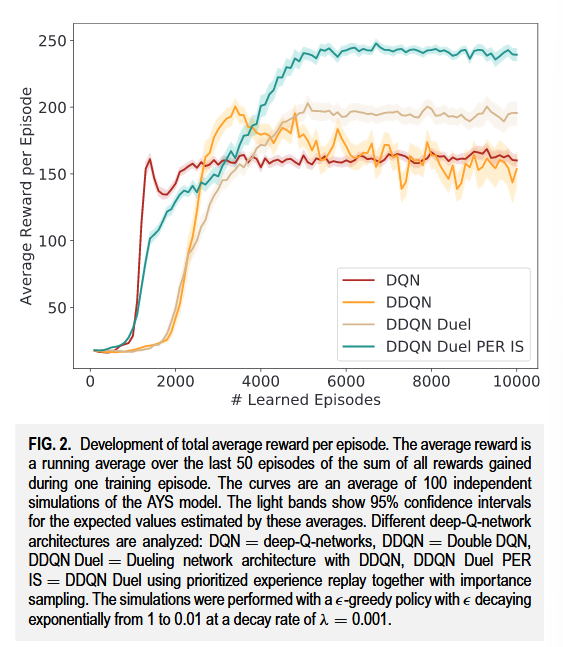


FIG. 2. Development of total average reward per episode. The average reward is a running average over the last 50 episodes of the sum of all rewards gained during one training episode. The curves are an average of 100 independent simulations of the AYS model. The light bands show 95% confidence intervals 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

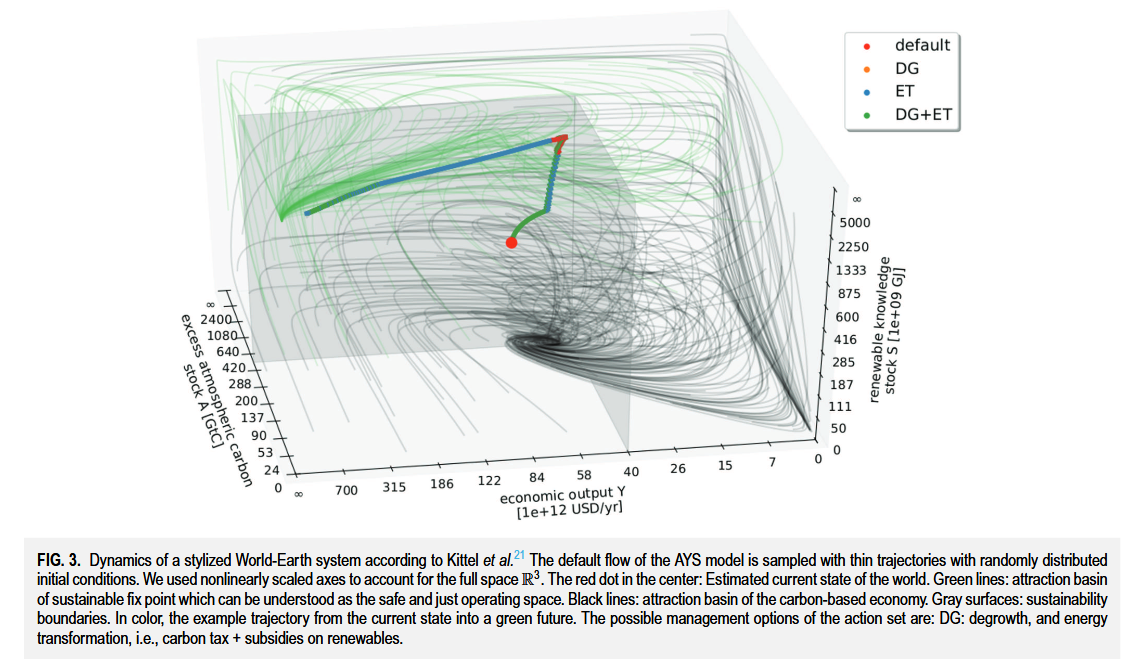


FIG. 3. Dynamics of a stylized World-Earth system according to Kittel et al.21 The default flow of the AYS model is sampled with thin trajectories with randomly distributed initial conditions. We used nonlinearly scaled axes to account for the full space R3. The red dot in the center: Estimated current 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。

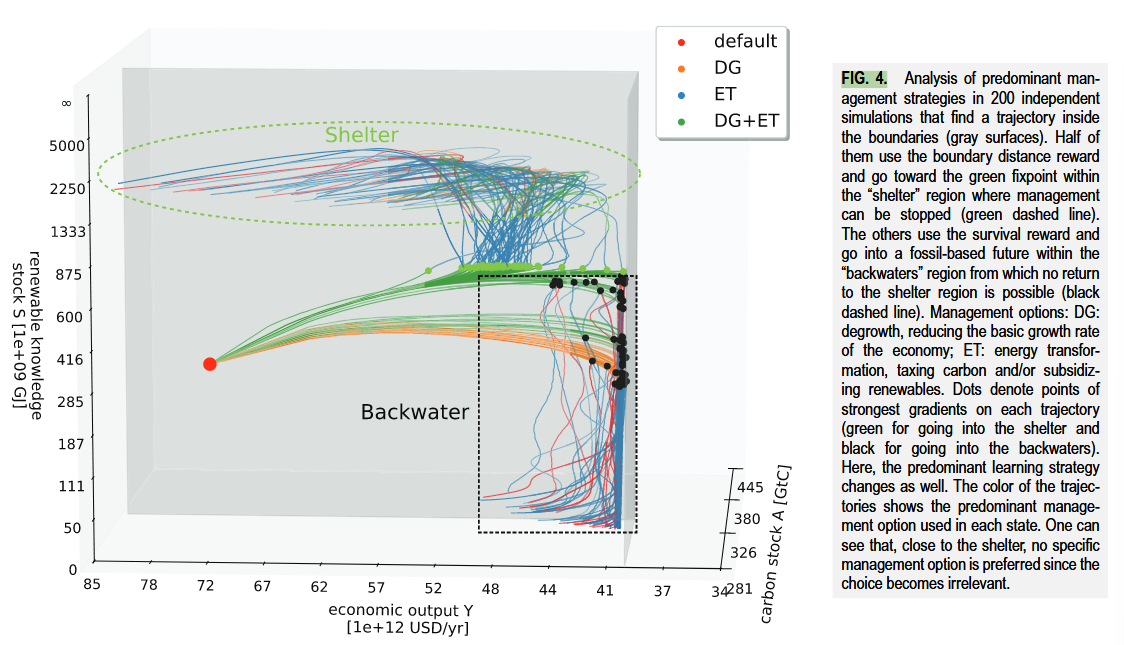


FIG. 4. Analysis of predominant management strategies in 200 independent simulations that find a trajectory inside the boundaries (gray surfaces). Half of them use the boundary distance reward and go toward the green fixpoint within the “shelter” region where management can be stopped (green dashed line). 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。

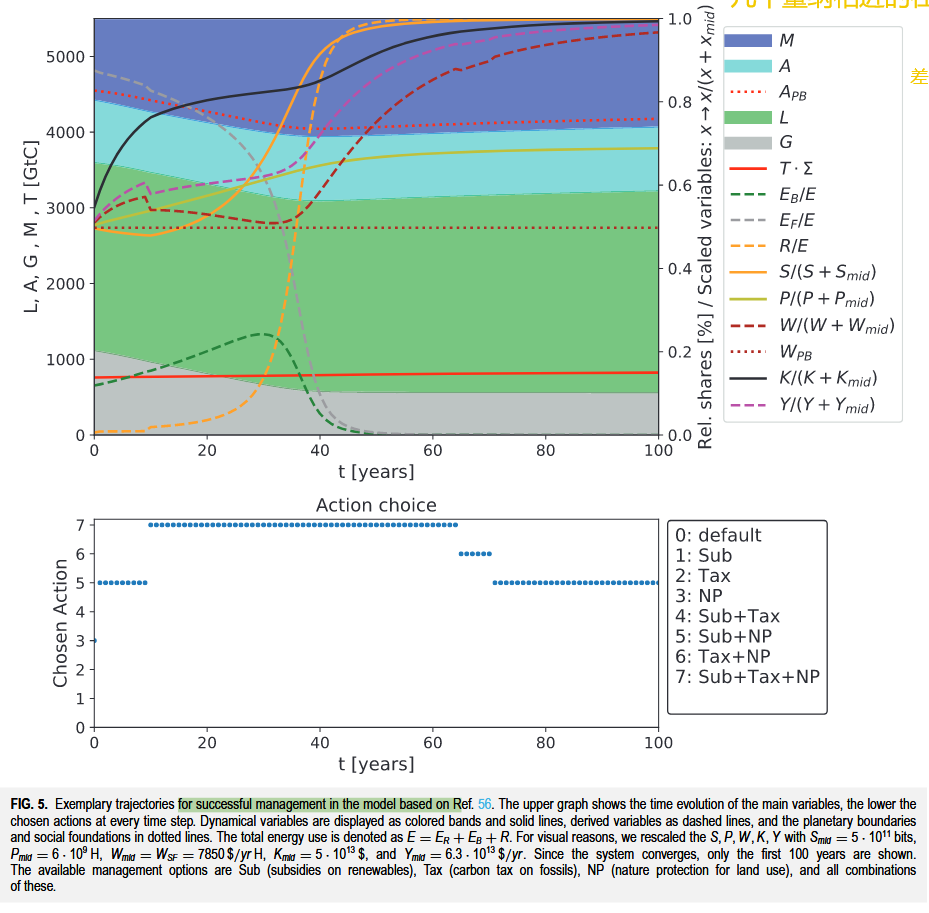


FIG. 5. Exemplary trajectories for successful management in the model based on Ref. 56. The upper graph shows the time evolution of the main variables, the lower the chosen actions at every time step. Dynamical variables are displayed as colored bands and solid lines, derived variables as dashed lines, 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。

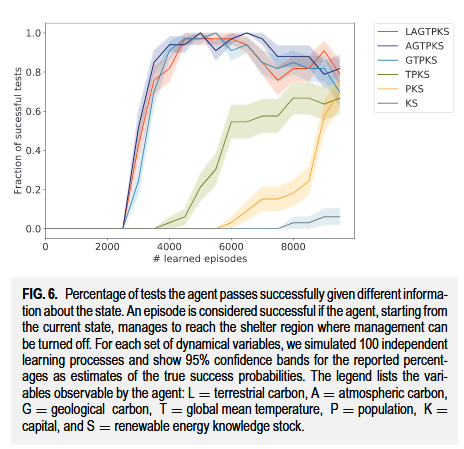


FIG. 6. Percentage of tests the agent passes successfully given different information about the state. An episode is considered successful if the agent, starting from the current state, manages to reach the shelter region where management can be turned off. For each set of dynamical variables, we simulated 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。

【实验分析】

III. APPLICATION TO WORLD-EARTH MODELS

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.

Ⅱ.应用TO WORLD-EARTH MODELS

基于上述方案，我们实现了一个Agent，该Agent通过使用DRL (见Sec .ⅡB)来学习，以管理Sec中描述的环境。II D .对智能体的训练次数最多为104次，每50次评估一次学习成功。在合理的计算时间(单机一到两个小时)内，可以在标准笔记本电脑上进行单次仿真实验。利用一个调整后的超参数集(详情见附录表一)，我们可以总结出这项工作的三个关键发现，如下所述。首先，我们发现在环境中增加奖励的学习确实是可能的。第二，我们调查了学习者发现的具体路径，并观察到了

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. 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
2. 训练和稳定性为了验证我们算法的整体适用性，我们首先从总体上分析了学习行为。与监督学习中通过在一组测试数据上评估算法的性能不同，在RL问题中如何准确地评估智能体的训练过程并不明显。

在此，我们沿用Mnih等( 25 )的方法对训练进行适当的可视化。我们将代理人在一次跑步过程中所获得的总报酬绘制在学习片段的数量上。每个值被计算为超过50个情节的运行平均值。每条曲线为100次独立模拟的平均值。因此，我们看到，在经历了一定数量的情节之后，在我们的环境中，每一情节的平均报酬显著增加

reveal a high reward. In other words, it learns to manage the environment. We conclude that management can indeed be learned by the agent. 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

显示出高额的报酬。换句话说，它学会了管理环境。我们得出结论，管理确实可以被代理人学习。更进一步，我们观察到智能体的学习是通过使用DQN - Learning的扩展来稳定的，如Sec所述。ⅡC.情节表明，使用结合双Q学习( DDQN + Duel)和优先经验重放与重要性采样( PER IS )的竞争网络体系结构显著提高了DQN - Agent的性能。PER IS的积极影响可以通过观测来解释，在两种环境中，我们在结果的轨迹中发现了状态，其中动力学显著地改变了(由于将概述如下)。包含这些状态的经验将在PER IS的学习过程中享有特权。这与文献中的结果完全一致。52。因此，下面概述的所有结果都是通过使用我们的最佳执行剂( DDQN + Duel + PER IS)来实现的，如果没有其他说明。此外，这与其他不同学习架构的比较在定性上也是一致的，如文献所述。34，学习曲线呈现出与其他DRL应用中所见类似的形状。24，25，34

1. Management pathways in World-Earth system models In the following, we discuss the pathways in the two environments described in Sec. II D that were found by using the outlined framework of DRL. In Sec. III B 1, we explore the AYS model and, in Sec. III B 2, the copan:GLOBAL model. Specifically, in both environments, we are interested in whether the learner finds trajectories toward regions, which we can associate with a safe and just operating space without violating sustainability boundaries. First, we present some successful examples. As a next step, we look at the specific trajectories in more detail, hoping to understand the general strategy the agent found to reach its aim (i.e., maximize the total reward).

世界-地球系统模型中的管理路径在下面，我们讨论Sec中描述的两种环境中的管理路径。【实验环境描述对应，有哪些env】II D是利用DRL的框架发现的。在Sec .三

B 1，我们研究了AYS模型，并在Sec .三

B2，Copan：全球模型。【说明实验结果中关注的重点】具体而言，在两者中，在环境方面，我们感兴趣的是学习者是否找到了通往区域的轨迹，在不违反可持续性边界的情况下，我们可以与安全和公正的操作空间相联系。

首先，我们给出了一些成功的例子。【对重点中的轨迹进行策略分析】作为下一步，我们将更详细地观察具体的轨迹，希望能够理解代理人为达到其目标(也就是说,最大化总报酬)而发现的一般策略。

**1. Pathways in the AYS model**

In the AYS model, the agent can choose between the following actions: “energy transformation” (taxing carbon emissions and/or subsidizing renewables) or “degrowth management” (reducing the basic economic growth rate) or neither or both of them. As a first analysis step, we look at the pathways the agent takes after it was trained for a sufficiently long time (i.e., the convergence of the learning is reached, see Fig. 2). We find that, even though the dynamics of the environment is unknown to the agent in advance, it is able to find trajectories within sustainability boundaries (see Fig. 3) that were deemed impossible in another study based on a viability theory algorithm that used state space discretization.21

Due to the setup of our framework, each of the two management options can only be switched on and off. In Fig. 3, in the region near to the boundaries, the energy transformation (ET) option (representing an energy tax or subsidy) is switched on and off in short alternations, achieving essentially the effect that a continuous application of a smaller tax/subsidy would have. Hence, offering different tax levels as individual options might improve the learning success further.

1 . AYS模型中的路径

【基本交代agent的action代表意义】

在AYS模型中，代理人可以在以下行动之间进行选择："能源转型" (征收碳排放税和/或补贴可再生能源)或"去增长管理" (降低基本经济增长率)，或者两者都有或都没有。【算法收敛的结果交代-图2分析的交代】作为第一个分析步骤，我们研究了该代理经过足够长的(也就是说,学习的收敛性已经达到,见图2)训练后所采取的路径。【具体单个算法单个reward对应的3维轨迹，结果分析】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】

由于我们框架的设置，两个管理选项中的每一个只能被打开和关闭。【对于高维轨迹中的细节交代，边界情况的特点】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】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。

To get a deeper understanding of the found solutions, we take a closer look at the different trajectories that were detected by using the DRL framework. Depending on the chosen reward function, the paths found by the agent differ. If the boundary distance reward is chosen, after sufficiently long learning, the agents always find a path toward the “green fixpoint” at (A, Y, S) = (0, ∞, ∞), where the distance to the boundaries is maximized. For the survival reward, the agent is only interested in staying within the boundaries. Therefore, it finds pathways leading to the green fixpoint as well as pathways toward a region close to the boundaries with S = 0 where it then manages to stay. Although many viable paths are found by the learner, we find that the learning strategies that were found can be generalized. We analyzed the management options the agent uses most on different parts of the trajectories. They are depicted in Fig. 4. These different regions of predominant management options are now used for the following discussion. The different regions colored in Fig. 4 may be analyzed with respect to a mathematical theory of the qualitative topology of the state space of a dynamical system with management options and desirable states, called topology of sustainable management (TSM).60 Interestingly, these regions can be seen to correspond roughly to some concepts from the TSM framework, in particular, the concept of “shelter” and “backwaters.” The approximate locations of these regions are depicted by dashed lines in Fig. 4.

【过渡句，引出下一个实验的目标和结果 】为了更深入地了解所发现的解决方案，我们仔细研究了使用DRL框架检测到的不同轨迹。根据选择的奖励函数，代理人找到的路径不同。【分析不同奖励函数下， 单一算法下结果不同】**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】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】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】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】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。

We identify a general strategy the agent uses. Starting from the current state, we found that in order to stay within the boundaries forever, it is not sufficient to use only one single management option of energy transformation (ET) or degrowth (DG) in the beginning. Rather, both ET and DG have to be applied to ensure keeping the system within the sustainability boundaries in future times. To understand this behavior, one has to recall the effect of the two possible management options DG and ET (for details, we refer to the Appendix). Both boundaries A ̄ and Y ̄ are potentially dangerous for the learner. Using only option ET will lead to an increase of renewable knowledge but violate the A ̄ boundary. Applying option DG will on the one hand respect the ̄A boundary but on the other hand hit the Y ̄ boundary. The strategy found by the learner is a mix of both options: First, it uses option ET + DG to reach a certain distance from the A ̄ boundary. However, the Y ̄ limit comes critically near. At a specific time point, the agent has to change its predominant management strategy to ET such that the renewable knowledge stock increases faster and the agents avoid transgressing the boundary value of Y ̄ . At this specific point where DG + ET changes to ET, a sharp turn in the trajectory happens (see Fig. 4). If S is large enough at this point, the turn is “upwards” and after some time, a region is reached where every trajectory is now leading toward unlimited growth of economic output and renewable knowledge regardless of the chosen management option, so that management can be “stopped,” leading to another sharp turn in the trajectories. In TSM, such a secure region is called a shelter. However, if S is too small at the turning point, the turn is “downward” toward S = 0, staying close to the social foundation boundary. In Ref. 21, it was shown that this leads to a region called the backwaters, from which the shelter could not be reached any longer, but one can still stay within the boundaries by managing over and over again.

Summarizing, the agent learns that the timing of the particular change of management is of crucial relevance. A general interpretation of the resulting pathways would be that ET, e.g., via taxing fossils, is highly important to ensure further development. However, to reach a secure state without violating the sustainability boundaries, a degrowth policy is needed for some time as well.

【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】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】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】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】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】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】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】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】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】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，】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】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】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。

2. Pathways in the c:GLOBAL model

We verify that our framework works as well in higherdimensional environments by applying it to the c:GLOBAL model. While classical approaches like viability theory are no longer well applicable because of the dimension, our DRL learner is also capable of detecting solutions toward a sustainable future in this model, see Fig. 5. Here, one learning episode has a maximum length of 500 yrs. Successful trajectories often converge already after around 100 yrs. However, to account for long-term effects, simulations were executed for times up to 500 yrs since we observed that seemingly converged trajectories sometimes transgressed boundaries at much later times, posing an additional challenge for the learner. The general strategy found by the learner turns out to be this. The NP option is used throughout and renewables are subsidized during most of the time. The crucial point is the timing of the carbon tax, which cannot be used immediately without violating the social foundation boundary. It is switched on only later and switched off again once renewables have passed through most of their learning curve.

2 . c：GLOBAL模型中的路径

【衔接，说明框架移植的有效性】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】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】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】如图5所示。【交代，copan基本实验的特点部分-time-收敛性】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】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.

【引子，另一个轨迹变化的分析 】关于代理的远视性的一个有趣的观察是以下内容。经过一些学习的情节，该代理商经常使用在20 - 60年内不使用任何管理的轨迹，这使系统在边界内保持一段时间，但导致稍后违反了t＆gt的行为。 100年。【不成功的实例分析-轨迹特点分析】一个示例轨迹可以在附录中的图7中找到。只有经过更多的情节，代理商才学会采取远见和使用管理选项的行动，这只会在很久以后才能识别差异，并避免越过边界。【小总结】这确实是DRL成功的关键特征，并显示了该方法的潜在功能。

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】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，】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】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】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】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。

1. Partial observability and noise As a generalization of Markov decision processes, partially observable Markov decision processes (POMDPs) are of great research interest. Here, the agent is only able to observe only part of the actual system state.61 We are interested in the performance of our DRL agent under such observational constraints since a real-world manager will only have access to vastly restricted information about the Earth system’s current state. Moreover, we added noise to the observations of the agent. Our experiments show (see Fig. 6) that, even under partial observability of the state, the agent is still capable of detecting sustainable solutions.

【不同实验维度不同-虽然是类似框架】3 .部分可观测性和噪声作为Markov决策过程的推广，部分可观测的Markov决策过程( POMDPs )引起了广泛的研究兴趣。在这里，代理只能观测到实际系统状态的一部分。我们对DRL代理在这种观测约束下的性能感兴趣，因为真实世界的管理者只能获得关于地球系统当前状态的非常有限的信息。此外，我们对智能体的观测值添加了噪声。我们的实验表明(见图6)，即使在状态部分可观的情况下，智能体仍然能够检测出可持续解。

We observe that the learning curves for observing either the full state(L, A, G, T, P, K, S) or only the variable combinations (A, G, T, P, K, S) or (G, T, P, K, S) have very similar shape. So, it seems that there is little added value in observing the carbon stocks L and A when already observing the geological stock G whose decline is essential for the timing of the carbon tax (but which is also the hardest to observe in reality). However, even if we limit the agent’s observation capabilities to the socioeconomic variables (P, K, S), the agent achieves a similar performance after a certain number of episodes, only considerably later. This can be explained by the dominant force humans exert on the Earth system.

我们观察到，用于观察完整状态的学习曲线（L，A，G，T，P，K，S）或仅可变组合（A，G，T，P，K，S）或（G，T，T，P，K，S）具有非常相似的形状。因此，在观察碳库存L和A时，似乎几乎没有附加价值，而在观察地质股票G时，其下降对于碳税的时机至关重要（但实际上也很难观察到）。但是，即使我们将代理商的观察能力限制在社会经济变量（P，K，S）中，该代理在一定数量的发作之后就达到了相似的性能，只有在大量之后。这可以通过人类在地球系统上施加的主要力量来解释。

To test the robustness of the DRL algorithm for a noisy state input, we added white observational noise on the input state st the agent receives from the environment. Not surprisingly, noise can disturb the agent’s learning and lead to a massive decrease in performance if the environment gets more complex (see Fig. 8 in the Appendix for details). Neural networks are known to be vulnerable by perturbed input62,63 and the harmful effect of noise has already been observed and discussed as well in DRL applications.64–66 Still, for further experiments with more realistic scenarios, the influence of noise has to be investigated more systematically.

For the analysis of trajectories in the Earth system, we can deduce the following. Even if the full state will not be observable to the agent, it is just based on the distance boundary reward signal still able to sufficiently “understand” the system’s dynamics in order to find appropriate management pathways. Furthermore, in our experiments, we see that noise will be a limiting factor for some DRL algorithms. In simulations with very noisy environments, some preprocessing of the input state might be necessary to use DRL successfully.

为了测试DRL算法对带噪状态输入的鲁棒性，我们在智能体从环境中接收到的输入状态st上添加了白色观测噪声。毫不奇怪，如果环境变得更加复杂(详情见附录图8)，噪声会干扰智能体的学习并导致性能的大幅下降。众所周知，神经网络容易受到扰动输入62，63的影响，噪声的有害影响已经在DRL应用中得到了观察和讨论。然而，为了更真实地进行更多的实验，噪声的影响必须进行更系统的研究

为了分析地球系统中的轨迹，我们可以推断以下内容。即使代理无法观察到的全状态，也仅基于距离边界奖励信号仍然能够充分“理解”系统的动态以找到适当的管理途径。此外，在我们的实验中，我们看到噪声将是某些DRL算法的限制因素。在具有非常嘈杂的环境的模拟中，成功使用DRL可能需要对输入状态进行一些预处理。

IV. CONCLUSION The main contribution of this work is the development of a framework for using DRL in Earth system models, mathematically formalized in a Markov decision process. Throughout this paper, we have combined the technique of deep reinforcement learning with Earth system modeling in order to detect global sustainable management strategies. We have presented a prototype for which we hope extensions based on our work will become a helpful tool to discover and analyze management pathways and to get a deeper understanding of the impact of global governance policies.

Ⅳ.结论这项工作的主要贡献是开发了一个框架，用于在地球系统模型中使用DRL，以马尔可夫决策过程进行数学形式化。在整个论文中，我们将深度强化学习技术与地球系统建模相结合，以检测全球可持续管理策略。我们提出了一个原型，希望基于我们的工作的扩展将成为发现和分析管理途径的有用工具，并更深入地理解全球治理政策的

As a proof of concept, we have applied it to two exemplary models from Earth system science, taken from the World-Earth modeling literature. They include components of Earth system modeling as well as constraints of planetary boundaries and social foundations. We have shown that our algorithm successfully identified trajectories toward a secure region for the Earth system which a competing approach using viability theory and a discretization of the state space were not able to find.21 Even very simple reward functions were sufficient, and only partial observations of the system state were necessary for the learner to understand the complex, nonlinear system’s dynamics. However, noisy observations have presented a challenge. We have found significant learning improvements by using the combination of DQN with dueling network architectures and prioritized experience replay and importance sampling

作为概念的证明，我们将其应用于来自世界地球模型文献的地球系统科学的两个模型模型。它们包括地球系统建模的组成部分以及行星边界和社会基础的约束。我们已经表明，我们的算法成功地识别了针对地球系统的安全区域的轨迹，使用可行性理论的竞争方法和状态空间的离散化方法无法找到。21即使是非常简单的奖励功能也足够了，只有对系统状态的部分观察才能使学习者了解复杂的，非线性的非线性系统的动力学。但是，嘈杂的观察结果提出了挑战。我们发现通过使用DQN与决斗网络架构的组合以及优先的经验重播和重要性抽样，发现了重大的学习改进

With respect to management strategies that the learner found in the AYS and the c:GLOBAL model, we can support the intuition that there is not one single way for staying within the boundaries nor can the impact of global management be observed immediately. Rather, we conclude from our models that only an intelligent combination and timing of global policies may lead to a sustainable future. We found that besides making renewables more attractive, also a temporary slowing down of economic growth might be necessary for staying within planetary boundaries.

关于学习者在 AYS 和 c：GLOBAL 模型中找到的管理策略，我们可以支持这样一种直觉，即没有单一的方法可以保持在界限内，也不能立即观察到全局管理的影响。相反，我们从模型中得出结论，只有全球政策的明智组合和时机才能带来可持续的未来。我们发现，除了使可再生能源更具吸引力外，为了保持在地球边界内，可能还需要暂时放缓经济增长。

Moreover, we have shown that our method is applicable as well in environments with only partially observable state spaces. Due to its connection to real-world problems,61 for example, in 3D navigation,67 partial observability of state spaces is widely discussed in the reinforcement learning community. Hence, in future work, the effects of reducing the dimensionality of the state space in our World-Earth system models need to be studied in more detail.

此外，我们已经证明，我们的方法也适用于只有部分可观察状态空间的环境。由于它与现实世界的问题有关，61 例如，在 3D 导航中，67 状态空间的部分可观察性在强化学习社区中被广泛讨论。因此，在未来的工作中，需要更详细地研究在我们的世界-地球系统模型中降低状态空间维度的影响。

We used DRL to identify trajectories under certain constraints. Formally, this can be regarded as an optimization problem, which could be approached with other methods as well. For example, the IAM community typically uses commercial solvers for the optimization of long-term social welfare functions, which are influenced by nonlinear underlying dynamics. However, the choice of the welfare function is not directly intuitive and hard to justify straightforwardly.16 As an example, Pindyck16 puts forward the significant differences in the outcome of two established models in IAM. The results in Refs. 68 and 69 differ widely, mainly based on the different values of the discount rates for the choice of which no uniform theory exists. However, in our models, the constraints imposed by sustainability boundaries, as well as the choice of simple reward functions, could be argued to be easier to justify and to understand intuitively in some contexts.

我们使用 DRL 来识别某些约束下的轨迹。从形式上讲，这可以被视为一个优化问题，也可以使用其他方法来解决。例如，IAM 社区通常使用商业求解器来优化长期社会福利函数，这些函数受非线性底层动力学的影响。然而，福利函数的选择并不是直接直观的，也很难直接证明.16 例如，Pindyck16 提出了 IAM 中两个已创建模型的结果的显着差异。The results in Refs.68 和 69 差异很大，主要是基于选择的折扣率值不同，不存在统一的理论。然而，在我们的模型中，可持续性边界施加的约束，以及简单奖励函数的选择，可以说在某些情况下更容易证明和直观地理解。

We encourage the reader to apply our framework to his or her preferred models. Since we formulated our problem as an MDP, our approach is not restricted to deterministic environments but can be generalized to environments that include stochastic dynamics and agent-based components. One could think about replacing the global society used in the models above by agent-based models of regionally distributed interacting societies. Following the model developed by Wiedermann et al.,70,71 which is a stochastic environment based on an adaptive network model, could be a first step in this direction. On the other hand, the biophysical dynamics could be incorporated in more detail as well by using more complex global vegetation models such as LPJ.72

我们鼓励读者将我们的框架应用于他或她喜欢的模型。由于我们将问题表述为 MDP，因此我们的方法不仅限于确定性环境，而且可以推广到包括随机动力学和基于代理的组件的环境。人们可以考虑用基于代理的区域分布互动社会模型来取代上述模型中使用的全球社会。遵循 Wiedermann 等人开发的模型70,71，这是一个基于自适应网络模型的随机环境，可能是朝着这个方向迈出的第一步。另一方面，通过使用更复杂的全球植被模型（如 LPJ.72），也可以更详细地纳入生物物理动力学

Further, an interesting next step could be to use DQN agents to represent major real-world agents such as governments in a multiagent environment setting. Here, first experiments in simple grid worlds have already been performed to investigate sequential social dilemmas73 and common-pool resource appropriation.74 Connections to game theory in the climate context are conceivable as well.47,75

此外，一个有趣的下一步可能是使用 DQN 代理来表示主要的真实代理，例如多代理环境中的政府。在这里，已经在简单的网格世界中进行了第一个实验，以研究顺序社会困境73和公共池资源挪用74。在气候背景下，与博弈论的联系也是可以想象的47,75。

Another approach that might be promising is to include modelbased RL in our framework. Regarding computation time, modelbased RL tends to be much more efficient.76 The key difference is that model-free methods act in the real environment in order to collect rewards and update the action value functions accordingly. In contrast, the agent in model-based methods uses RL to learn a model of the environment and then predicts the system dynamics in a second step. Once the model is learned, actions can be chosen by using optimal control theory. Specifically, as environments in World-Earth models are often based on a set of biophysical and socioeconomic differential equations, this approach might be promising. However, highly complex environments often cannot be learned perfectly, such that solutions of this method involve the risk of being suboptimal. A possible approach to overcome this issue is recently developed algorithms that aim to combine advantages of both methods in one algorithm.77

另一种可能很有前途的方法是在我们的框架中包含基于模型的 RL。关于计算时间，基于模型的 RL 往往效率更高。76 关键区别在于，无模型方法在真实环境中起作用，以收集奖励并相应地更新动作值函数。相比之下，基于模型的方法中的代理使用 RL 来学习环境模型，然后在第二步中预测系统动力学。一旦学会了模型，就可以通过使用最优控制理论来选择行动。具体来说，由于世界-地球模型中的环境通常基于一组生物物理和社会经济微分方程，因此这种方法可能是有前途的。然而，高度复杂的环境往往无法完美地学习，因此这种方法的解决方案存在次优的风险。 克服这个问题的一种可能方法是最近开发的算法，旨在将两种方法的优点结合到一种算法中77。

Another fruitful exchange could emerge between the field of Earth system analysis and the field of safe and beneficial AI.78 For example, the important question of the latter field of how selflearning agents can safely explore an environment without pursuing catastrophic action directly translates to finding sustainable policies in Earth system analysis. Here as well, management strategies need to navigate uncertain environments without activating tipping elements in the Earth system with potentially catastrophic impacts on human societies.79,80

另一种可能很有前途的方法是在我们的框架中包含基于模型的 RL。关于计算时间，基于模型的 RL 往往效率更高。76 关键区别在于，无模型方法在真实环境中起作用，以收集奖励并相应地更新动作值函数。相比之下，基于模型的方法中的代理使用 RL 来学习环境模型，然后在第二步中预测系统动力学。一旦学会了模型，就可以通过使用最优控制理论来选择行动。具体来说，由于世界-地球模型中的环境通常基于一组生物物理和社会经济微分方程，因此这种方法可能是有前途的。然而，高度复杂的环境往往无法完美地学习，因此这种方法的解决方案存在次优的风险。 克服这个问题的一种可能方法是最近开发的算法，旨在将两种方法的优点结合到一种算法中77。