

Climate Surrogates for Scalable Multi-Agent Reinforcement Learning: A Case Study with CICERO-SCM

Oskar Bohn Lassen
Technical University of Denmark
Kongens Lyngby, Denmark
obola@dtu.dk

Filipe Rodrigues
Technical University of Denmark
Kongens Lyngby, Denmark
rodr@dtu.dk

Serio Angelo Maria Agriesti
Technical University of Denmark
Kongens Lyngby, Denmark
samaa@dtu.dk

Francisco Camara Pereira
Technical University of Denmark
Kongens Lyngby, Denmark
camara@dtu.dk

ABSTRACT

Climate policy studies require models that capture the combined effects of multiple greenhouse gases on global temperature, but these models are computationally expensive and difficult to embed in reinforcement learning. We present a multi-agent reinforcement learning (MARL) framework that integrates a high-fidelity, highly efficient climate surrogate directly in the environment loop, enabling regional agents to learn climate policies under multi-gas dynamics. As a proof of concept, we introduce a recurrent neural network architecture pretrained on (20,000) multi-gas emission pathways to surrogate the climate model CICERO-SCM. The surrogate model attains near-simulator accuracy with global-mean temperature RMSE $\approx 0.0004\text{K}$ and approximately $1000\times$ faster one-step inference. When substituted for the original simulator in a climate-policy MARL setting, it accelerates end-to-end training by $> 100\times$. We show that the surrogate and simulator converge to the same optimal policies and propose a methodology to assess this property in cases where using the simulator is intractable. Our work allows to bypass the core computational bottleneck without sacrificing policy fidelity, enabling large-scale multi-agent experiments across alternative climate-policy regimes with multi-gas dynamics and high-fidelity climate response.

KEYWORDS

Surrogate modeling, Climate simulation, Multi-agent reinforcement learning, Climate policy analysis

1 INTRODUCTION

Climate modeling provides the scientific backbone for climate policy exploration, but there exists an inherent trade-off between model fidelity and computational tractability. State-of-the-art Earth System Models (ESMs) resolve intricate physical processes across the atmosphere, ocean, cryosphere, and biosphere, yielding detailed projections of climate variables under specified emission scenarios [4]. However, these models are slow to run - a single ESM simulation can require days to weeks of wall-clock time on high-performance computing systems - severely limiting the number of scenarios or policy strategies one can feasibly evaluate. This computational barrier motivates the use of simpler models for many applications [1].

To enable broader and faster exploration of scenarios, the climate science community relies on reduced-complexity Simple Climate

Models (SCMs) that emulate the climate’s response at far lower computational cost. SCMs are compact models - often energy-balance models with simplified ocean and carbon-cycle components - calibrated to reproduce the behavior of more complex ESMs [20]. For example, MAGICC [25] and CICERO-SCM [12] both use energy-balance formulations coupled to upwelling-diffusion ocean models, providing tractable representations of global climate behavior. Other SCMs such as FaIR take an even more simplified approach, using impulse-response functions to approximate the carbon cycle and temperature response [18, 29]. These approaches sacrifice some process-level detail in exchange for very high computational efficiency. Because SCMs run orders of magnitude faster than ESMs, they have been widely adopted for applications requiring large ensembles or many iterative evaluations, such as probabilistic climate projections or integrated assessment models (IAMs) [1].

IAMs couple the economy, energy-land systems, and society with a climate module to assess mitigation and impact pathways. By linking socio-economic drivers to greenhouse gas emissions and their consequences for the climate system, IAMs translate climate outcomes into economic metrics. Pioneering IAMs like DICE and its regional variant RICE demonstrated this paradigm by combining a highly simplified climate module with a neoclassical economic optimization approach [30–35]. Building on this foundation, more detailed IAM frameworks such as REMIND-MAGPIE, MESSAGE-GLOBIOM, WITCH, and IMAGE employed optimization-based or game-theoretic formulations that incorporate additional realism (e.g. technological detail, land-use, energy systems) [8, 10, 22, 43]. These latter models often make use of socio-economic scenarios (SSPs) to explore uncertainty [37]. In all cases, IAMs rely on fast climate simulators, such as MAGICC, for their climate component, because the computational cost of ESMs precludes their use in large scenario ensembles or within iterative optimization loops. One important limitation of the traditional IAM paradigm is that it typically models the world as a handful of aggregate regions that are internally homogeneous and that optimize toward an equilibrium outcome under strong foresight assumptions. This aggregation and reliance on optimal-control or game-theoretic formulations can mask heterogeneity in preferences, adaptive behavior, and the path-dependent dynamics of collective actions. These limitations have prompted interest in more flexible, simulation-based approaches [21, 40].

Reinforcement learning (RL), and in particular multi-agent RL (MARL), has been proposed as a promising alternative framework for studying climate-economy interactions [26, 38]. In an MARL formulation, multiple agents (e.g. countries or regions) make simultaneous decisions and learn strategies through repeated interaction in a simulated environment, rather than assuming an equilibrium or globally optimized trajectory. This approach can accommodate heterogeneous agents, non-linear dynamics, and bound rational decision-making. An early attempt to couple MARL with an IAM is the RICE-N model introduced by Zhang et al. [45], which extended the RICE integrated assessment model by replacing its decision making with learning agents. However, RICE-N used a highly simplified climate module and basic mechanisms for cooperation, limiting its realism and the policies that could emerge. Subsequent work has begun to enrich this line of research: for example, Rudd-Jones et al. [39] explore more sophisticated cooperation and coalition formation mechanisms in a learning-based climate game, and Heitzig et al. [14] examines the impact of commitment strategies on climate mitigation outcomes. The recent JUSTICE framework by Biswas et al. [3] goes further by integrating the FaIR climate model into a multi-objective MARL setting, improving the fidelity of climate dynamics within the learning environment. Despite this progress, even JUSTICE only allowed agents to control CO₂, retained a low-dimensional action space, and involved only a few agents, limiting policy exploration and undercutting the benefit of a more detailed climate response. In reality, greenhouse gases and aerosols span a spectrum of radiative effects that impact different time horizons. Long-lived gases such as CO₂ and N₂O persist for centuries and set the baseline for long-term warming, methane remains in the atmosphere for about a decade and has a particularly strong near-term warming effect, and very short-lived species such as ozone and aerosol precursors decay within months and often cool the climate. The actions a policymaker can take to reduce emissions (mitigation levers) act on these species heterogeneously where for example decarbonizing the energy sector cuts CO₂^{FF} but also reduces SO₂ emissions, which can lead to short-term warming. Land-use measures such as halting deforestation affect both CO₂^{AFOLU} and CH₄ without the same short-term climate penalty. To explore such interactions, MARL climate games need a climate engine that responds to multi-gas emission patterns rather than just aggregate CO₂. Overall, state-of-the-art MARL climate studies have relied on oversimplified climate dynamics and severely restricted the action space, mainly due to computational constraints.

Scalability is a particularly critical issue as MARL experiments often require on the order of 10⁷ environment interactions for agents to learn effective policies [36]. Even a fast SCM that takes only a few tenths of a second per call can become a bottleneck when invoked millions of times. This explains why, to date, MARL studies have been unable to incorporate more complex or multi-gas climate models - doing so would entail intractable computational running time. A natural next step is to find a way to embedding higher-fidelity climate dynamics into MARL frameworks without incurring a prohibitive computational cost.

One promising approach is to use surrogate modeling and machine learning emulators. Recent work has shown that machine learning surrogates can accelerate the most computationally intensive components of ESMs by orders of magnitude [6], and that deep

neural networks can learn to mimic short-term climate predictions with high speed and reasonable accuracy [44]. These successes suggest that simplified climate models too might be further accelerated by surrogate approaches. If an accurate and faster surrogate could emulate an SCM, it would enable the integration of more detailed climate responses in contexts like MARL or large uncertainty ensembles that require millions of model evaluations.

In this paper, we propose to extend the realism of MARL climate policy games by embedding more complex, multi-gas climate dynamics into the environment while keeping optimization tractable. This is achieved by integrating surrogate models into the environment loop. First, we design a framework that proposes how to integrate surrogate models as a replacement for the climate module of a climate-economy MARL game. Secondly, we introduce a recurrent neural network surrogate of CICERO-SCM, detail the design of our MARL experiment, and outline a method for evaluating policy consistency when running the simulator is intractable. Lastly, we report the results showing how the MARL training time can be reduced by orders of magnitude while maintaining policy fidelity.

2 CONCEPTUAL FRAMEWORK AND CLIMATE ENGINE

A realistic MARL climate environment must expose agents to policy levers that mirror key mitigation and adaptation priorities in climate policy. According with the IPCC Sixth Assessment Report (AR6) on mitigation and on impacts & adaptation [13, 17, 24], the primary interventions include: (i) decarbonizing the energy sector (e.g. coal phase-out, renewable expansion), (ii) targeting methane abatement (e.g. waste management, leak mitigation, livestock strategies), (iii) improving agricultural and land-use practices (e.g. reduced deforestation, fertilizer efficiency, sustainable intensification), and (iv) investing in adaptation or preventive measures that reduce realized damages.

These different policies affect multiple greenhouse gases that have a different effect on the change of climate. Ideally, the emission changes should propagate into a high-fidelity climate model providing a realistic estimate of the temperature increase given the policies. The modeled temperature increase should then propagate into region specific damage functions.

However, when developing MARL climate environments, there is a trade-off between environment complexity and tractability. As mentioned in Section 1, simplifications make experiments tractable but also constrain the policies and dynamics that can be explored.

We introduce a modular framework in which complex but realistic climate components of the environment can be exchanged with fast surrogate emulators. We conceptually divide the environment into three modules:

- (1) **Emissions Module:** a mapping from the agents' chosen actions / policies to emissions of various gases
- (2) **Climate Module:** a climate dynamics function f that takes the current emissions as input and produces the climate's response (e.g. temperature change)
- (3) **Impact Module:** a translation of climate outcomes and chosen actions into economic costs and climate damages

As outlined in Figure 1, one can seamlessly replace a high-fidelity SCM, f_{SCM} , with a learned surrogate model, f_{θ} , without needing

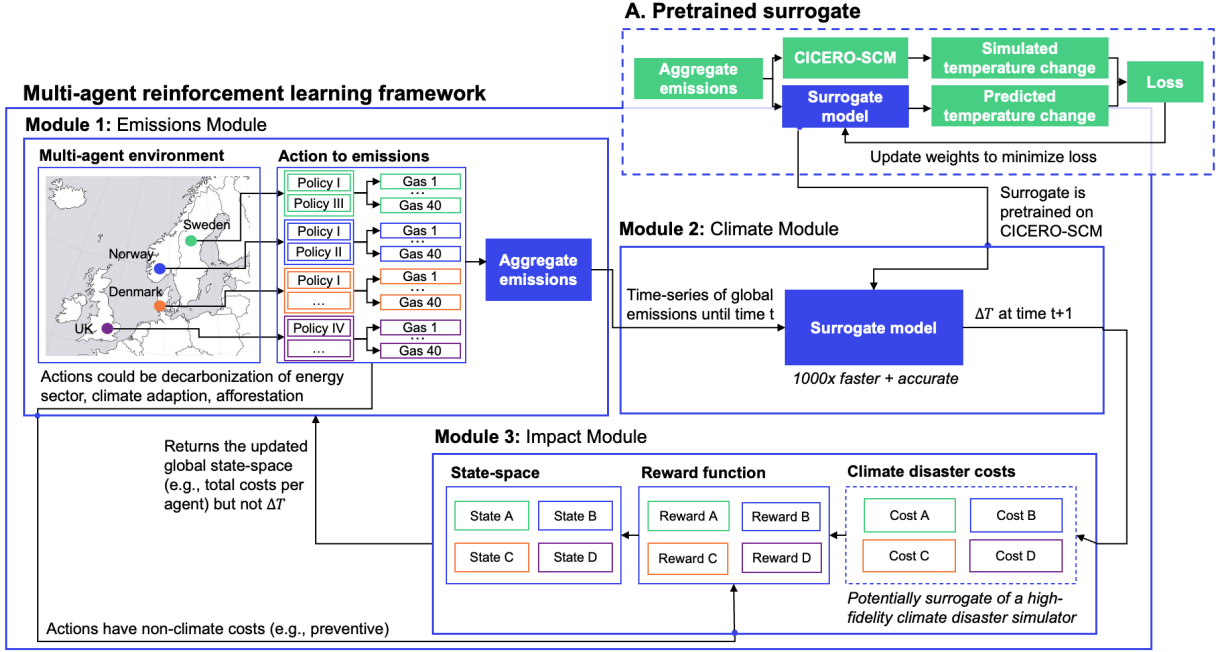


Figure 1: Proposed framework for integrating climate surrogates into MARL environments. In Module 1, agents choose policies that result in emissions, which in Module 2 are translated into temperature change by a pretrained surrogate and in Module 3 converted into costs.

to alter any other parts of the environment. The surrogate thus serves as a replacement that emulates the behavior of the original SCM. This preserves the increased scientific realism from the SCM, using multi-gas pathways with a high-fidelity climate response, while enabling the use of high-speed, hardware-optimized surrogate models within the RL training loop.

The modular structure also provides flexibility to increase fidelity in other parts of the MARL environment without requiring changes to the decision-making loop or learning algorithm. For instance, the impact module can range from a simple damage function that converts global temperature increase into economic damage, to local sea-level rise, agricultural yield changes, or other more realistic region-specific damage functions.

In the remainder of this paper, we instantiate this framework using the CICERO-SCM as our high-fidelity climate engine and a recurrent neural network (RNN) as the surrogate emulator. We substitute it into a multi-agent climate-economic experiment to demonstrate the improvements.

3 METHODOLOGY

In this section, the methodological steps adopted to design a surrogate for CICERO-SCM and embed it into the MARL framework are described.

3.1 Climate dynamics engine: CICERO-SCM

We use the reduced-complexity global climate model CICERO-SCM (v1.1.1) recently implemented in Python [28] as our climate dynamics simulator which maps multi-gas emission trajectories to global

mean surface air temperature. Let \mathcal{G} be the set of gases ($|\mathcal{G}| = 40$) used in CICERO-SCM and let us define the global emissions vector as:

$$E(t) = (E_g(t))_{g \in \mathcal{G}} \in \mathbb{R}^{|\mathcal{G}|} \quad (1)$$

$$E_{1:t} = (E(\tau))_{\tau=1}^t \in \mathbb{R}^{t \times |\mathcal{G}|} \quad (2)$$

where the notation $(E_g(t))_{g \in \mathcal{G}}$ defines a vector of dimension $|\mathcal{G}|$ containing all the elements in \mathcal{G} . CICERO-SCM evolves annually as a dynamical system that updates its internal state based on the emission history. The model can be written as a recursive mapping:

$$\Delta T(t) = f_{\text{SCM}}(E_{1:t}) \quad (3)$$

where $\Delta T(t)$ is the simulated global mean surface air temperature change in year t , and $E_{1:t}$ represents the full emissions history up to that year. Internally, the model couples (i) a semi-empirical carbon-cycle module converting CO_2 emissions to atmospheric concentrations, (ii) exponential-decay schemes for other long-lived gases such as CH_4 and N_2O , and (iii) an upwelling-diffusion energy-balance model linking total radiative forcing to transient temperature response. Radiative efficiencies follow [9], and parameters controlling ocean heat uptake and radiative forcing are calibrated to Earth System Models (ESMs) and observations. CICERO-SCM thus provides a computationally tractable yet physically consistent mapping. However, each call still takes ≈ 0.4 s, making complex MARL games with millions of steps impractical.

3.2 Surrogate model of CICERO-SCM

Generation of emission trajectories for model training. To capture the range of emission pathways that could arise in our MARL setup (Section 3.3), we generate an ensemble of trajectories by perturbing the year-over-year growth rates of the SSP2-4.5 baseline scenario [11] from 2015-2075. We smooth the year-over-year growth to reduce short-term volatility and better match the smoother trajectories typical of learned policies.

We first compute the baseline year-over-year growth factor for each gas:

$$\delta_g^{\text{base}}(t) = \frac{E_g^{\text{base}}(t)}{E_g^{\text{base}}(t-1)} \quad (4)$$

where $E_g^{\text{base}}(t)$ is the SSP2-4.5 baseline emissions for gas g .

For each scenario $s \in [1, \dots, S]$ where $S = 20,000$, we generate gas-specific multiplicative changes $\zeta_g^s(t)$ by drawing numbers from a uniform distribution within bounds (ℓ_g, u_g) :

$$\zeta_g^s(t) \sim \mathcal{U}(\ell_g, u_g) \quad (5)$$

To reduce year-to-year short-term volatility, we apply exponential smoothing in log space, which corresponds to a geometric exponential moving average (EMA) on the growth factors:

$$\zeta_g^s(t) = (\zeta_g^s(t-1))^\alpha (\zeta_g^s(t))^{1-\alpha}, \quad \zeta_g^s(2015) = 1 \quad (6)$$

where $\alpha = 0.8$. We then perturb the baseline growth as:

$$\delta_g^s(t) = \delta_g^{\text{base}}(t) \zeta_g^s(t) \quad (7)$$

which defines emissions recursively:

$$E_g^s(t) = E_g^{\text{base}}(t) \quad t \leq 2015, \quad (8)$$

$$E_g^s(t) = E_g^s(t-1) \delta_g^s(t) \quad t \geq 2016. \quad (9)$$

For five gases to be controlled in the MARL experiment (Section 3.3), we define a subset $C \subseteq \mathcal{G}$ of gases:

$$C = \{\text{CO}_2^{\text{FF}}, \text{CO}_2^{\text{FOLU}}, \text{CH}_4, \text{N}_2\text{O}, \text{SO}_2\} \quad (10)$$

and define bounds dependent on whether the gas is in the subset:

$$(\ell_g, u_g) = \begin{cases} (0.925, 1.075) & g \in C \\ (1, 1) & g \in \mathcal{G} \setminus C \end{cases} \quad (11)$$

which implies that gases $g \in \mathcal{G} \setminus C$ follow the baseline emission growth and gases $g \in C$ follow a perturbed growth with $\pm 7.5\%$ changes in the growth rate per year. For each scenario s and year t , we define the emissions vectors as:

$$E^s(t) = (E_g^s(t))_{g \in \mathcal{G}} \in \mathbb{R}^{|\mathcal{G}|} \quad (12)$$

$$E_C^s(t) = (E_g^s(t))_{g \in C} \in \mathbb{R}^{|C|} \quad (13)$$

where $E^s(t)$ and $E_C^s(t)$ contains the sampled emissions in year t for $t \geq 2015$, while for $t < 2015$ it contains the historical baseline emissions $E^{\text{base}}(t)$.

Simulated temperature responses using CICERO-SCM. The simulator produces a projection of global mean surface air temperature change $\Delta T^s(t)$ over $t \in [1900, 2075]$ where Δ indicates the change compared to pre-industrial temperature at year 1900.

Running the model for all $S = 20,000$ scenarios yields the temperature ensemble in Figure 2. The ensemble shows a narrow spread in near-term warming due to the dominance of the shared historical

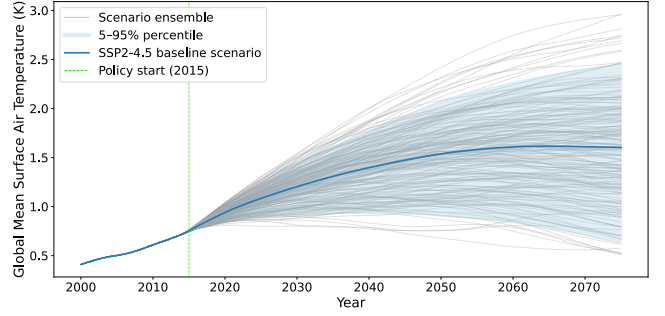


Figure 2: Global mean surface air temperature change for the generated emission trajectories.

emissions, but diverges progressively towards 2075 as post-2015 emission differences accumulate. The SSP2-4.5 baseline lies close to the ensemble median, indicating that the perturbation design generates futures consistent with the reference scenario while still spanning substantial variation in end-of-period warming.

Data processing for surrogate modeling. Using the emission trajectories, $E_C^s(t)$ and the temperature outputs, $\Delta T^s(t)$, we reformat the data into supervised learning samples suitable for training an RNN-based surrogate model:

$$X^s(t) = [E_C^s(t-W), \dots, E_C^s(t)] \quad (14)$$

$$y^s(t) = \Delta T^s(t) \quad (15)$$

where W is the window length (in years) and X^s has shape $(W+1) \times |C|$. We choose $W = 65$ to ensure that the input contains sufficient historical context to capture slow climate system responses and long-lived greenhouse gas effects, while remaining computationally efficient. Temperature is not used as an input, and hence the model is not autoregressive in temperature.

We construct one training sample per scenario and per target year $t \in \{2015, \dots, 2075\}$, with the input window spanning $[t-W, \dots, t]$ (may start before 2015). Targets with $t < 2015$ are excluded because pre-2015 emissions are identical across scenarios and provide no policy-relevant variation. This yields 61 data points per scenario (2015–2075) via rolling windows and across the 20,000 generated scenarios, this results in 1.22 million samples for the surrogate model.

The dataset is split by scenario into training (70%), validation (15%), and testing (15%), ensuring no temporal leakage between splits. Splitting by scenario, rather than by time, prevents the model from implicitly learning from past or future years of the same emission pathway.

Architecture of RNN-based surrogate model. We develop a surrogate with an RNN-based architecture where the input is the emissions window $X(t) \in \mathbb{R}^{(W+1) \times |C|}$ and the task is to predict the temperature change $\Delta T(t)$. The architecture comprises three modules: (i) an RNN encoder, (ii) a skip connection, and (iii) a prediction head as shown in Figure 3.

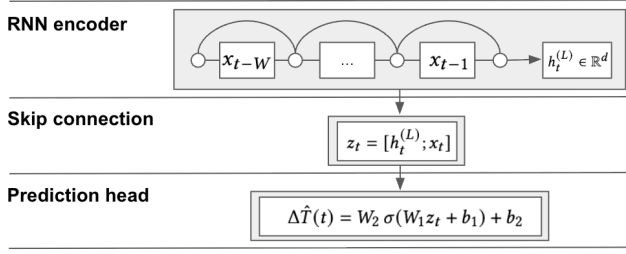


Figure 3: Architecture of the RNN-based surrogate.

For the RNN encoder, let $X_{\text{hist}}(t) = [x_{t-W}, \dots, x_{t-1}]$ with $x_\tau \in \mathbb{R}^{|C|}$. Stacked recurrent layers map $X_{\text{hist}}(t)$ to a hidden representation $h_t^{(L)}$ summarizing the historical dynamics:

$$h_t^{(L)} = \text{RNN}_\theta([x_{t-W}, \dots, x_{t-1}]), \quad h_t^{(L)} \in \mathbb{R}^d, \quad (16)$$

where $\text{RNN}_\theta(\cdot)$ denotes the stacked recurrent encoder parameterized by θ . In our experiments, we tested Long Short-Term Memory (LSTM) [16], a Gated Recurrent Unit (GRU) [5] and a Temporal Convolutional Network (TCN) [23] as the recurrent encoder. The TCN replaces the recurrence with convolutions but preserves the same input-output structure. Importantly, the RNN encoder operates only on the historical window $x_{t-W:t-1}$ as the current-year emissions are stored separately and concatenated via a skip connection:

$$z_t = [h_t^{(L)}; x_t] \in \mathbb{R}^{d+|C|} \quad (17)$$

so that short-horizon signals in x_t are preserved alongside the long-horizon summary $h_t^{(L)}$ before the prediction head.

The prediction head maps z_t to the surrogate output via a two-layer MLP:

$$\Delta \hat{T}(t) = W_2 \sigma(W_1 z_t + b_1) + b_2 \quad (18)$$

where σ is a GELU or SiLU nonlinearity [7, 15].

3.3 MARL Climate Mitigation Experiment

We consider a finite-horizon Markov game for climate mitigation, played annually from 2016 to 2050 ($H=35$) among N agents (countries). Each year, agents select policy actions that influence greenhouse gas emissions and adaptation levels. These choices determine the multi-gas emission pathways passed to the climate module (f_{SCM} or f_{NET}), which produces next-year temperatures and affects the cost of damages. The episode length and action space are chosen to remain within the surrogate’s training distribution.

Actions. In each year t , all agents act simultaneously. Agent i selects a discrete action vector:

$$a_{i,t} = (e_{i,t}, m_{i,t}, l_{i,t}, p_{i,t}) \quad (19)$$

where $e_{i,t} \in \{0, 0.5, 1.0\}$ is the effort level for energy decarbonization, $m_{i,t} \in \{0, 0.5, 1.0\}$ for methane abatement, $l_{i,t} \in \{0, 0.5, 1.0\}$ for agricultural and land-use measures, and $p_{i,t} \in \{0, 0.03, 0.08\}$ for preventive investment (representing climate adaptation measures). The three effort levers for mitigation $(e_{i,t}, m_{i,t}, l_{i,t})$ map to growth deviations in the controllable gases C through a fixed policy matrix $M \in \mathbb{R}^{3 \times |C|}$. Each row of M specifies how one policy lever affects

the growth of all gases in C . Given an agent’s effort vector for mitigation $k_{i,t} = (e_{i,t}, m_{i,t}, l_{i,t})$, the induced growth deviations are:

$$\tilde{\delta}_{i,g}(t) = \begin{cases} [k_{i,t} M]_g & g \in C \\ 0 & g \in \mathcal{G} \setminus C \end{cases} \quad (20)$$

and the effective growth factors are then given by:

$$\delta_{i,g}^{\text{eff}}(t) = \delta_g^{\text{base}}(t) (1 + \tilde{\delta}_{i,g}(t)) \quad (21)$$

where $\delta_g^{\text{base}}(t)$ is from Equation (4) and $\delta_{i,g}^{\text{eff}}(t)$ is defined for all $g \in \mathcal{G}$ and for all agents $i \in [1, \dots, N]$. The adaptation action $p_{i,t}$ accumulates into a prevention stock P_i that multiplicatively attenuates climate damages in the reward function.

Climate engine. Let $E^{\text{base}}(t) \in \mathbb{R}^{|\mathcal{G}|}$ be baseline global emissions and $S_i \in \mathbb{R}^{|\mathcal{G}|}$ per-gas shares for each agent with $\sum_{i=1}^N S_i = \mathbf{1}_{|\mathcal{G}|}$. We start per-agent realized emissions at the last historical year and then compound using the effective growth $\delta_i^{\text{eff}}(t) \in \mathbb{R}^{|\mathcal{G}|}$ defined above:

$$\bar{E}_i(2015) = S_i \odot E^{\text{base}}(2015) \quad (22)$$

$$\bar{E}_i(t) = \bar{E}_i(t-1) \odot \delta_i^{\text{eff}}(t), \quad (t = 2016, \dots, 2050) \quad (23)$$

$$\Delta T(t) = f\left(\sum_{i=1}^N \bar{E}_i(t)\right) \quad (24)$$

where \odot denotes the elementwise product. We use f either as CICERO-SCM (f_{SCM}) or a learned RNN-based surrogate (f_{NET}). The climate simulator (f_{SCM}) maps the $|\mathcal{G}|$ -dimensional global emissions $\bar{E}(t)$ to temperature change $\Delta T(t)$. The climate surrogate (f_{NET}) maps the $|C|$ -dimensional global emissions vector $\bar{E}^C(t)$ to temperature change $\Delta T(t)$. A filtering of gases is made inside the step-function depending on the climate engine. Both climate engines are initialized with historical emissions, and in each step the current emissions $\bar{E}(t)$ are added to their internal state.

Observation. All agents receive the same centralized vector:

$$O(t) = [\Delta T(t-1), \tau(t), \bar{E}_i^C(t-1), D_i^C(t-1), P_i(t-1)] \quad (25)$$

where $\tau(t) \in [0, 1]$ is a normalized year index, $\bar{E}_i^C(t-1)$ are last-year realized emissions for the controllable gases (C) per agent, $D_i^C(t-1)$ are cumulative deviations from baseline emissions for the controllable gases (C) per agent, and $P_i(t-1) \in [0, P_{\text{max}}]$ are prevention stocks per agent. The two summary quantities are defined as:

$$D_i^C(t) = \sum_{u=2016}^t \left(\bar{E}_i^C(u) - S_i^C \odot E^{\text{base},C}(u) \right) \quad (26)$$

$$P_i(t) = \min\{P_{\text{max}}, P_i(t-1) \phi + p_{i,t}\} \quad (27)$$

where P_{max} is the maximum effect prevention can have on the climate cost and $\phi \in [0, 1]$ is a decay rate on the preventive investments. All components in $O(t)$ with i subscript are vectors and flattened, so every agent receives all information about other agents’ previous actions.

Reward. At each year t , the reward for agent i is the negative of three types of costs: $C_i^c(t)$ climate disaster costs, $C_i^k(t)$ policy costs, and $C_i^p(t)$ prevention costs:

$$r_i(t) = -\eta \left(C_i^c(t) + C_i^k(t) + C_i^p(t) \right) \quad (28)$$

$$C_i^c(t) = c_i^c \psi (\Delta T(t))^4 (1 - P_i(t)) \quad (29)$$

$$C_i^k(t) = (c_i^k)^\top (k_{i,t} \odot k_{i,t}) \quad (30)$$

$$C_i^p(t) = c_i^p \cdot p_{i,t} \quad (31)$$

where $c_i^c, c_i^k = (c_i^e, c_i^m, c_i^l)$, and c_i^p are the agent-specific climate cost, policy costs, and prevention cost respectively. $P_i(t) \in [0, P_{\max}]$ is the prevention stock, $\psi = 0.003$ is the base climate damage calibration parameter, and $\eta = 10^{-1}$ is the coefficient for reward normalization.

At the terminal step we add a look-ahead cost for near-term climate damage. From the terminal state, $t = 2050$, we roll the climate model forward for $U = 15$ years with baseline emission growth $\delta_g^{\text{base}}(t)$ used to calculate $\bar{E}(t)$ using equation (23, 24) and decaying prevention $P_i(t+u) = \min\{P_{\max}, P_i(t) \phi^u\}$ for $[u, \dots, U]$, we define the terminal penalty:

$$r_i^{\text{term}}(t) = \sum_{u=1}^U C_i^c(t+u) \quad (32)$$

and adjust the terminal reward as

$$r_i(t) \leftarrow r_i(t) - r_i^{\text{term}}(t) \eta. \quad (33)$$

This terminal look-ahead serves as a pragmatic reward-shaping term to reflect imminent damages in the terminal step.

Optimization. We train independent recurrent policies using Proximal Policy Optimization (PPO) [41]. Each agent i has parameters θ_i and seeks to maximize its expected discounted return:

$$\theta_i^* = \arg \max_{\theta_i} J_i(\theta_i) \quad (34)$$

$$J_i(\theta_i) = \mathbb{E}_{\tau \sim \pi_{\theta_i}} \left[\sum_{t=0}^{H-1} \gamma^t r_i(t) \right], \quad \gamma = 0.999 \quad (35)$$

where $r_i(t)$ is the per-step reward defined in equations (28–31). Policies are implemented as LSTM actor-critics trained over complete $H=35$ -year episodes, with the hidden state carried forward through time.

Policy scenarios. We consider two climate-policy games that differ primarily in how rapidly policies can be learned from the available reward signals.

(i) *Tractable scenario.* This scenario uses four homogeneous agents ($N = 4$) with identical damage and mitigation cost parameters. The only effective lever is energy decarbonization and the other actions are either prohibitively expensive or have negligible impact. These design choices yield strong gradient signals and hence fast convergence. The purpose of this setting is to empirically investigate if the surrogate and simulator learn identical policies. Full numerical parameters are listed in Appendix D.

(ii) *Intractable scenario.* The other scenario includes more agents $N = 10$ with heterogeneous damage sensitivities, emission shares, and mitigation costs. Several mitigation levers produce similar climate outcomes, making it difficult for the agents to discern which

actions are most effective and hence having weaker gradient signals. As a result, discovering high-quality policies requires far more environment interactions. Training this scenario to convergence with CICERO-SCM would be prohibitively slow (millions of simulator calls), so we instead train with the surrogate and evaluate policy consistency using the proposed replay-based method described in Section 3.4. Full numerical parameters are listed in Appendix D.

3.4 Evaluation criteria

We evaluate surrogates by one-step inference speed, predictive accuracy on the test data, MARL training acceleration, and policy consistency (whether policies learned with f_{NET} match those from f_{SCM}).

Accuracy. We measure predictive accuracy of the surrogate f_{NET} relative to CICERO-SCM f_{SCM} using the root-mean-square error (RMSE) and coefficient of determination R^2 between predicted and true temperature increments $\Delta \hat{T}(t)$ and $\Delta T(t)$ over a held-out test set.

Inference per-step time. For both f_{NET} and f_{SCM} we construct a class that is initialized with historical emissions. The class includes a step-method that takes an emissions vector $E(t)$ as argument, appends it to the history, and executes the climate engine to produce $\Delta T(t)$. We count the time it takes to append $E(t)$, normalize it (for f_{NET}), and run a forward pass generating $\Delta T(t)$. We evaluate CICERO-SCM on CPU and the RNN-based surrogate on CPU and GPU. We compute the one-step prediction inference time for 100,000 inference steps and report the mean inference time.

MARL per-step time. We compare per environment step time under two otherwise identical environments (scenario (i)) that differ only in the climate backend $f \in \{f_{\text{SCM}}, f_{\text{NET}}\}$. CICERO-SCM is executed on CPU whereas the RNN-based surrogate is executed on GPU.

Policy consistency. An ideal surrogate should induce policies indistinguishable from those obtained with the original simulator. Let Π denote the policy class and $J_f(\pi) = \mathbb{E}[\sum_{t=0}^H \gamma^t r_t \mid f, \pi]$ denote the expected discounted return under climate engine f and policy $\pi \in \Pi$. We define policy consistency as:

$$\text{sign}[\Delta J_{f_{\text{NET}}}(\pi_1, \pi_2)] \approx \text{sign}[\Delta J_{f_{\text{SCM}}}(\pi_1, \pi_2)], \quad \forall \pi_1, \pi_2 \in \Pi \quad (36)$$

$$\nabla_{\theta} J_{f_{\text{NET}}}(\pi_{\theta}) \approx \nabla_{\theta} J_{f_{\text{SCM}}}(\pi_{\theta}), \quad \forall \theta \in \mathcal{N}(\theta_{\text{SCM}}^*) \quad (37)$$

where $\Delta J_f(\pi_1, \pi_2) = J_f(\pi_1) - J_f(\pi_2)$ is the reward difference between two policies, $\theta_{\text{SCM}}^* = \arg \max_{\theta} J_{f_{\text{SCM}}}(\pi_{\theta})$ are the optimal policy parameters under f_{SCM} , and $\mathcal{N}(\theta_{\text{SCM}}^*)$ is a neighborhood around the optimum. This definition is conceptually related to model-based RL analyses of return and gradient alignment between learned dynamics and true environments [19]. The first condition requires that both environments induce approximately the same preference ordering over candidate policies, while the second ensures that local ascent directions around the optimum align, leading to convergence toward the same equilibrium under gradient- or exploration-based updates. This parallels the theoretical findings of Shen et al. [42], who show that when a learned dynamics model's predictions match the real environment along a policy's visitation distribution, the resulting return discrepancies, and hence policy

rankings and gradients, are negligible. Although Shen et al. [42] study a model-based RL setting, their analysis of return discrepancies provides conceptual support for our consistency criteria. Nonetheless, to evaluate these criteria, it requires access to policies trained directly with f_{SCM} , which can be intractable to get.

We propose a tractable empirical alternative. Let’s first define the feasible set of emissions trajectories attainable in our MARL game \mathcal{K} as:

$$\mathcal{K} := \{ \bar{E}(\cdot) \mid \bar{E}(t), \tilde{\delta}_{\mathcal{G} \setminus C}(t) = 0, \tilde{\delta}_C(t) \in \mathcal{D}(M, \mathcal{L}) \} \quad (38)$$

where the link between $\tilde{\delta}$ and \bar{E} is described in equations (21-24) and \mathcal{D} is the set of controllable growth deviations induced by the lever levels $k_{i,t} \in \mathcal{L}$ and the policy matrix M (Section 3.3).

Let $\mathcal{S} \subseteq \mathcal{K}$ be defined as the emissions trajectories stored during training using f_{NET} more formally defined as:

$$\mathcal{S} = \{ \bar{E}^{(k)}(\cdot) \}_{k=1}^K \quad (39)$$

$$\bar{E}^{(k)}(\cdot) = (\bar{E}^{(k)}(1), \dots, \bar{E}^{(k)}(H+U)) \quad (40)$$

$$\bar{E}^{(k)}(t) \in \mathbb{R}^{|\mathcal{G}|} \quad (41)$$

where K denotes the total number of training episodes, $H = 35$ is the episode length, and $U = 15$ is the rollout length. Let then $\tilde{\mathcal{S}} \subseteq \mathcal{K}$ denote the set of emission trajectories that would be stored during training using f_{SCM} if that would have been tractable. Trivially, $\mathcal{S} = \tilde{\mathcal{S}}$ if $f_{\text{NET}} = f_{\text{SCM}}$ (all else equal).

More generally, if the surrogate uniformly approximates the simulator on \mathcal{K} :

$$\sup_{\bar{E}(\cdot) \in \mathcal{K}} \|f_{\text{NET}}(\bar{E}(\cdot)) - f_{\text{SCM}}(\bar{E}(\cdot))\|_{\infty} \leq \varepsilon, \quad (42)$$

then for sufficiently small ε one expects the sets of policy-induced trajectories to satisfy:

$$\|\mathcal{S} - \tilde{\mathcal{S}}\| \rightarrow 0 \text{ as } \varepsilon \rightarrow 0. \quad (43)$$

While we do not prove this formally, the intuition is that a surrogate that accurately approximates the simulator will have policy-induced emission trajectories \mathcal{S} that are arbitrarily close to those of the true simulator $\tilde{\mathcal{S}}$. Moreover, because MARL exploration introduces stochasticity in the agents’ policy updates, the sampled trajectories may already overlap substantially with $\tilde{\mathcal{S}}$ even for moderate approximation errors ($\varepsilon > 0$).

Therefore, we propose to randomly sample N trajectories from \mathcal{S} (uniform without replacement) and replay them through f_{SCM} to obtain $\Delta T^{\text{SCM}}(\cdot)$. While $\mathcal{S} \neq \tilde{\mathcal{S}}$ in general, a sufficiently accurate surrogate implies substantial overlap between \mathcal{S} and $\tilde{\mathcal{S}}$, so samples from \mathcal{S} provide a reasonable proxy for the policy-induced emission trajectory distribution under the true simulator $\tilde{\mathcal{S}}$. This intuition mirrors the transition-occupancy-matching approach of Ma et al. [27], which learns a dynamics model by matching the distribution of transitions experienced by the current policy in the real environment and in the model. By focusing on policies along the convergence trajectory rather than random states, we evaluate the surrogate in the regions of the state-action space most relevant for converging to the same policies.

Consequently, computing RMSE between $\Delta T^{\text{NET}}(\cdot)$ and $\Delta T^{\text{SCM}}(\cdot)$ for N sampled trajectories in \mathcal{S} provides an empirical evaluation of the surrogate’s accuracy on policy-induced emission paths. To

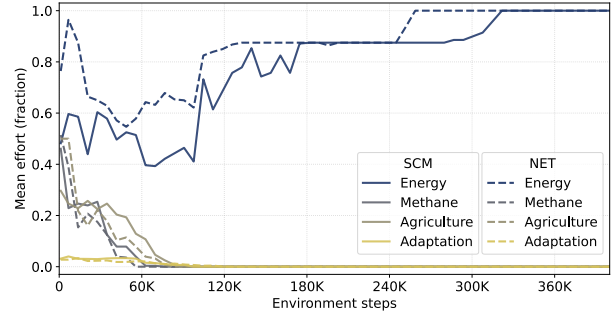


Figure 4: Comparison of learned policies in tractable scenario (i) between CICERO-SCM and the GRU-based surrogate.

test preservation of preference ordering (eq. 36), we also compute Kendall’s τ rank-consistency between discounted temperature-based returns defined as:

$$r_k^f = - \sum_t \gamma^t \Delta T_k^f(t), \quad f \in \{f_{\text{SCM}}, f_{\text{NET}}\} \quad (44)$$

where $k = [1, \dots, N]$. Evaluating only $N \ll |\mathcal{S}|$ trajectories keeps the evaluation tractable when training π^{SCM} to convergence is intractable.

4 RESULTS

4.1 Surrogate model performance

The left side of Table 1 summarizes the surrogate accuracy and computational efficiency on held-out data. All three surrogates achieve nearly perfect accuracy with the GRU variant being the most accurate, with an RMSE of 3.7×10^{-4} K and the TCN variant being the least accurate with RMSE of 6.8×10^{-4} K. Given that the training trajectories span 0–2.5K (Figure 2), these errors are very small. The LSTM and GRU variants have the fastest one-step inference on GPU with ~ 0.0004 seconds translating into ~ 200 - $1100\times$ faster than CICERO-SCM. This three-order-of-magnitude gain makes it feasible to embed climate dynamics in MARL experiments that would otherwise be intractable. The TCN is not as fast as it required 5 layers to achieve accuracy in the order of 10^{-4} whereas LSTM and GRU only required 1 layer.

The one-step inference speed-ups translate into $> 100\times$ faster per-environment call time during MARL training for the LSTM and GRU variant. The speed-up in MARL training is not one-to-one with the speed-up in one-step inference as other components of the MARL loop (policy networks, synchronization, communication overhead) begin to dominate once the surrogate is fast enough. In addition, we are running 32 environments per runner which further hides step-function latency. This makes the comparison practical but conservative, since the apparent speed-up is smaller than it would be without this parallelism.

4.2 Policy consistency

High test-set accuracy does not necessarily mean that the surrogate will induce the same policies as CICERO-SCM. Agents trained with different climate engines may visit different state trajectories, and small prediction errors can compound over long horizons.

Table 1: Surrogate performance on held-out test data and policy-induced trajectories and acceleration of inference speed and MARL environment step in scenario (i). Speed-up is measured relative to CICERO-SCM.

Climate engine	Test data performance & inference speed			Policy-induced performance & MARL speed			
	Test data (RMSE, R^2)	Mean inference [s] (CPU/GPU)	Speed-up (CPU/GPU)	Scenario (i) (RMSE, rank- τ)	Scenario (ii) (RMSE, rank- τ)	Mean env-step [s]	Speed-up
CICERO-SCM	–	464.4×10^{-3} / –	–	–	–	217.7×10^{-3}	–
LSTM	$(4.7 \times 10^{-4}, 0.99)$	1.1×10^{-3} / 0.4×10^{-3}	442 \times / 1161 \times	$(5.9 \times 10^{-4}, 0.996)$	$(3.2 \times 10^{-4}, 0.990)$	1.6×10^{-3}	137 \times
GRU	$(3.7 \times 10^{-4}, 0.99)$	2.3×10^{-3} / 0.4×10^{-3}	202 \times / 1161 \times	$(3.9 \times 10^{-4}, 0.996)$	$(2.0 \times 10^{-4}, 0.997)$	1.6×10^{-3}	137 \times
TCN	$(6.8 \times 10^{-4}, 0.99)$	3.3×10^{-3} / 1.3×10^{-3}	140 \times / 357 \times	$(21.1 \times 10^{-4}, 0.994)$	$(10.3 \times 10^{-4}, 0.982)$	4.5×10^{-3}	49 \times

To assess whether policies learned with f_{NET} remain consistent with those that would emerge under f_{SCM} , we replay $N = 1000$ emission trajectories sampled from the policy-induced distribution visited during training. For each trajectory, we compare the resulting temperature paths from the surrogate and simulator using the RMSE and the Kendall’s τ rank-consistency between discounted temperature-based returns. These metrics quantify pointwise accuracy along realistic, policy-relevant trajectories and whether the surrogate preserves the preference ordering over policies implied by the simulator.

(i) *Tractable scenario.* In the simple homogeneous tractable scenario, both the surrogate model and CICERO-SCM can be trained to convergence. Table 1 shows that the LSTM and GRU surrogates maintain low RMSE on the replayed trajectories, and Kendall’s τ confirms that returns remain correctly ordered. These results are empirically supporting that the policy consistency criteria formulated in equations (36-37) are satisfied. Figure 4 confirms convergence to the same optimal actions under both climate engines, indicating that when learning signals are strong, the surrogate reproduces the simulator’s optimal behavior.

(ii) *Intractable scenario.* The heterogeneous setting ($N = 10$ agents) requires many more environment interactions before policies stabilize and is therefore intractable to train with the simulator. We train the surrogate for $> 1\text{M}$ environment steps until all agents reach optimal reward. As reported in Table 1, RMSE on replayed policy-induced trajectories is even lower than in the tractable case, with similarly high Kendall’s τ rank-consistency. A plausible explanation is that with more agents and heterogeneous preferences, each country controls a smaller share of global emissions and optimal policies become less extreme, pulling global emissions closer to the SSP2-4.5 baseline and toward the center of the surrogate’s training distribution. We recognize that the proposed replay-based evaluation does not provide a formal guarantee of policy consistency, but it indicates small errors near the relevant parts of the trajectory space and preserves the ordering of returns across policies. Consequently, local gradients around the optimum are likely aligned, implying that training with the surrogate would converge to the same optimal policy as training with the simulator.

5 LIMITATIONS AND FUTURE RESEARCH

The results presented in this paper demonstrate the benefits of using surrogate climate models within climate-economic MARL

settings, however, there are several limitations and directions for future research.

Refinement of MARL experiment. Future work should refine mitigation levers and agent heterogeneity based on latest climate science. The MARL setup should include cooperation mechanisms and heterogeneous damage functions, preferably via surrogates of local high fidelity simulators. This would enable large scale comparative studies in the climate community and help analyze emerging behaviors under different policy designs.

Uncertainty-aware surrogates. We did not perturb the structural (calibration) parameters of CICERO-SCM’s differential-equation core when training the RNN surrogates. Conditioning the surrogate on these parameters would propagate structural uncertainty through the MARL loop and enable distributional or risk-sensitive objectives [2], which are especially relevant under tipping-point and tail-risk scenarios.

Policy consistency. While we propose an empirical method to test policy consistency, a formal proof of equations (42)–(43) would further substantiate the claim. Our conclusions do not depend on such a proof, and we leave it for future work.

6 CONCLUSION

We introduced a framework for integrating high-fidelity climate dynamics into scalable multi-agent reinforcement learning by replacing the climate module with a learned surrogate. We developed an RNN-based emulator of the CICERO-SCM climate model trained on 20,000 multi-gas emission trajectories. The surrogate achieves global-mean temperature RMSE of $< 0.0004\text{K}$ and approximately 1000 \times faster one-step inference, translating into end-to-end MARL training speed-ups $> 100\times$ relative to CICERO-SCM.

We show that, by using the surrogate within a MARL framework, we converge to the same set of policies in a computationally tractable experiment. When complexity precludes direct validation of policy consistency, we propose a methodology that replays policy-induced emission trajectories through the simulator, providing a tractable validation path when simulator-based convergence is infeasible.

Together, these results demonstrate that high-fidelity, multi-gas climate response models can be faithfully approximated and deployed as components of reinforcement learning environments - removing a major computational barrier to scalable research on cooperative climate-policy design and uncertainty propagation.

ACKNOWLEDGMENTS

The work presented in this article is supported by Novo Nordisk Foundation grant NNF23OC0085356.

REFERENCES

- [1] Romero-Prieto Alejandro, Mathison Camilla, and Smith Chris. 2025. Review of climate simulation by simple climate models. *EGUsphere* 2025 (2025), 1–80. doi:10.5194/egusphere-2025-2691
- [2] Marc G. Bellemare, Will Dabney, and Rémi Munos. 2017. A Distributional Perspective on Reinforcement Learning. arXiv:1707.06887 [cs.LG] <https://arxiv.org/abs/1707.06887>
- [3] Palok Biswas, Zuzanna Osika, Isidoro Tamassia, Adit Whorra, Jazmin Zatarain-Salazar, Jan Kwakkel, Frans A. Oliehoek, and Pradeep K. Murukannaiah. 2025. Exploring Equity of Climate Policies using Multi-Agent Multi-Objective Reinforcement Learning. arXiv:2505.01115 [cs.LG] <https://arxiv.org/abs/2505.01115>
- [4] Gordon B. Bonan and Scott C. Doney. 2018. Climate, ecosystems, and planetary futures: The challenge to predict life in Earth system models. *Science* 359, 6375 (2018), eaam8328. doi:10.1126/science.aam8328
- [5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv:1412.3555 [cs.NE] <https://arxiv.org/abs/1412.3555>
- [6] Lu Dan and Ricciuto Daniel. 2019. Efficient surrogate modeling methods for large-scale Earth system models based on machine-learning techniques. *Geoscientific Model Development* 12, 5 (2019), 1791–1807. doi:10.5194/gmd-12-1791-2019
- [7] Stefan Elfving, Eiji Uchibe, and Kenji Doya. 2017. Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning. arXiv:1702.03118 [cs.LG] <https://arxiv.org/abs/1702.03118>
- [8] Johannes Emmerling, Lara A. Reis, Michela Bevione, Loïc Berger, Valentina Bosetti, Samuel Carrara, Giacomo Marangoni, Fabio Sferra, Massimo Tavoni, Jan Witajewski-Baltviks, and Petr Havlik. 2016. The WITCH 2016 Model - Documentation and Implementation of the Shared Socioeconomic Pathways. *SSRN Electronic Journal* (01 2016). doi:10.2139/ssrn.2800970
- [9] M. Etminan, G. Myhre, E. J. Highwood, and K. P. Shine. 2016. Radiative forcing of carbon dioxide, methane, and nitrous oxide: A significant revision of the methane radiative forcing. *Geophysical Research Letters* 43, 24 (2016), 12,614–12,623. doi:10.1002/2016GL071930
- [10] Oliver Fricko, Petr Havlik, Joeri Rogelj, Zbigniew Klimont, Mykola Gusti, Nils Johnson, Peter Kolp, Manfred Strubegger, Hugo Valin, Markus Amann, Tatiana Ermolieva, Nicklas Forsell, Mario Herrero, Chris Heyes, Georg Kindermann, Volker Krey, David L. McCollum, Michael Obersteiner, Shonali Pachauri, Shilpa Rao, Erwin Schmid, Wolfgang Schoepp, and Keywan Riahi. 2017. The marker quantification of the Shared Socioeconomic Pathway 2: A middle-of-the-road scenario for the 21st century. *Global Environmental Change* 42 (2017), 251–267. doi:10.1016/j.gloenvcha.2016.06.004
- [11] Oliver Fricko, Petr Havlik, Joeri Rogelj, Zbigniew Klimont, Mykola Gusti, Nils Johnson, Peter Kolp, Manfred Strubegger, Hugo Valin, Markus Amann, Tatiana Ermolieva, Nicklas Forsell, Mario Herrero, Chris Heyes, Georg Kindermann, Volker Krey, David L. McCollum, Michael Obersteiner, Shonali Pachauri, Shilpa Rao, Erwin Schmid, Wolfgang Schoepp, and Keywan Riahi. 2017. The marker quantification of the Shared Socioeconomic Pathway 2: A middle-of-the-road scenario for the 21st century. *Global Environmental Change* 42 (2017), 251–267. doi:10.1016/j.gloenvcha.2016.06.004
- [12] Jan S. Fuglestedt and Terje K. Berntsen. 1999. *A Simple Model for Scenario Studies of Changes in Global Climate: Version 1.0*. CICERO Working Paper 1999-02. Center for International Climate and Environmental Research (CICERO). <https://pub.cicero.oslo.no/cicero-xmlui/handle/11250/192444>
- [13] Nabuurs G.-J., Mrabet R., Abu Hatab A., Bustamante M., Clark H., Havlik P., House J., Mbow C., Ninan K. N., Popp A., Roe S., Sohngen B., and S. Towprayoon. 2022. Agriculture, Forestry and Other Land Uses (AFOLU). In *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, P. R. Shukla, Jim Skea, R. Slade, A. Al Khouradajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, and J. Malley (Eds.). Cambridge University Press, Cambridge, UK and New York, NY, USA. doi:10.1017/9781009157926.009
- [14] Jobst Heitzig, Jörg Oechsler, Christoph Pröschel, Niranjana Ragavan, and Richie YatLong Lo. 2023. Improving International Climate Policy via Mutually Conditional Binding Commitments. arXiv:2307.14266 [cs.CY] <https://arxiv.org/abs/2307.14266>
- [15] Dan Hendrycks and Kevin Gimpel. 2023. Gaussian Error Linear Units (GELUs). arXiv:1606.08415 [cs.LG] <https://arxiv.org/abs/1606.08415>
- [16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9 (11 1997), 1735–1780. doi:10.1162/neco.1997.9.8.1735
- [17] Intergovernmental Panel on Climate Change (IPCC). 2022. Summary for Policy-makers. In *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, Portner H.-O., Roberts D. C., Poloczanska E.S., Mintenbeck K., Tignor M., Alegría A., Craig M., Langsdorf S., Loschke S., Moller V., and Okem A. (Eds.). Cambridge University Press, Cambridge, UK and New York, NY, USA, 3–33. doi:10.1017/9781009325844.001
- [18] Smith Christopher J., Forster P. M., Allen M., Leach N., Millar R. J., Passerello G. A., and Regayre L. A. 2018. FAIR v1.3: a simple emissions-based impulse response and carbon cycle model. *Geoscientific Model Development* 11, 6 (2018), 2273–2297. doi:10.5194/gmd-11-2273-2018
- [19] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. 2021. When to Trust Your Model: Model-Based Policy Optimization. arXiv:1906.08253 [cs.LG] <https://arxiv.org/abs/1906.08253>
- [20] Dorheim Kalyn, Link Robert, Hartin Corinne, Kravitz Ben, and Snyder Abigail. 2020. Calibrating Simple Climate Models to Individual Earth System Models: Lessons Learned From Calibrating Hector. *Earth and Space Science* 7, 11 (2020), e2019EA000980. doi:10.1029/2019EA000980 e2019EA000980
- [21] Konstantinos Koasidis, Alexandros Nikas, and Haris Doukas. 2023. Why integrated assessment models alone are insufficient to navigate us through the polycrisis. *One Earth* 6, 3 (2023), 205–209. doi:10.1016/j.oneear.2023.02.009
- [22] Elmar Krieglger, Nico Bauer, Alexander Popp, Florian Humpenöder, Marian Leimbach, Jessica Strefler, Lavinia Baumstark, Benjamin Leon Bodirsky, Jérôme Hilaire, David Klein, Ioanna Mouratiadou, Isabelle Weindl, Christoph Bertram, Jan-Philipp Dietrich, Gunnar Luderer, Michaja Pehl, Robert Pietzcker, Franziska Piontek, Hermann Lotze-Campen, Anne Biewald, Markus Bensch, Anastasis Giannousakis, Ulrich Kreidenweis, Christoph Müller, Susanne Rolinski, Anselm Schultes, Jana Schwanitz, Miodrag Stevanovic, Katherine Calvin, Johannes Emmerling, Shinichiro Fujimori, and Ottmar Edenhofer. 2017. Fossil-fueled development (SSP5): An energy and resource intensive scenario for the 21st century. *Global Environmental Change* 42 (2017), 297–315. doi:10.1016/j.gloenvcha.2016.05.015
- [23] Colin Lea, Michael D. Flynn, Rene Vidal, Austin Reiter, and Gregory D. Hager. 2016. Temporal Convolutional Networks for Action Segmentation and Detection. arXiv:1611.05267 [cs.CV] <https://arxiv.org/abs/1611.05267>
- [24] Clarke Leon, Y.-M. Wei, Ana De La Vega Navarro, Amit Garg, Andrea N. Hammann, Smail Khennas, Azevedo Ines M. L., Loschel Andreas, Singh A. K., Steg Linda, Strbac Goran, and Wada Keiichi. 2022. Energy Systems. In *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, P.R. Shukla, Jim Skea, R. Slade, A. Al Khouradajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, and J. Malley (Eds.). Cambridge University Press, Cambridge, UK and New York, NY, USA. doi:10.1017/9781009157926.008
- [25] Meinshausen M., Raper S. C. B., and Wigley T. M. L. 2011. Emulating coupled atmosphere-ocean and carbon cycle models with a simpler model, MAGICC6 – Part 1: Model description and calibration. *Atmospheric Chemistry and Physics* 11, 4 (2011), 1417–1456. doi:10.5194/acp-11-1417-2011
- [26] Strnad Felix M., Barfuss Wolfram, Donges Jonathan F., and Heitzig Jobst. 2019. Deep reinforcement learning in World-Earth system models to discover sustainable management strategies. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 29, 12 (12 2019), 123122. doi:10.1063/1.5124673
- [27] Yecheung Jason Ma, Kausik Sivakumar, Joesan Yan, Osbert Bastani, and Dinesh Jayaraman. 2023. Learning Policy-Aware Models for Model-Based Reinforcement Learning via Transition Occupancy Matching. In *Proceedings of the 5th Annual Learning for Dynamics and Control Conference (Proceedings of Machine Learning Research, Vol. 211)*, Nikolai Matni, Manfred Morari, and George J. Pappas (Eds.). PMLR, 259–271. <https://proceedings.mlr.press/v211/ma23a.html>
- [28] Sandstad Maria, Aamaas B., Johansen A. N., Lund M. T., Peters G. P., Samset B. H., Sanderson B. M., and Skeie R. B. 2024. CICERO Simple Climate Model (CICERO-SCM v1.1.1) – an improved simple climate model with a parameter calibration tool. *Geoscientific Model Development* 17, 17 (2024), 6589–6625. doi:10.5194/gmd-17-6589-2024
- [29] Zebdeed R. J. Nicholls, M. Meinshausen, J. Lewis, R. Gieseke, D. Dommgenet, K. Dorheim, C.-S. Fan, J. S. Fuglestedt, T. Gasser, U. Golüke, P. Goodwin, C. Hartin, A. P. Hope, E. Krieglger, N. J. Leach, D. Marchegiani, L. A. McBride, Y. Quilcaille, J. Rogelj, R. J. Salawitch, B. H. Samset, M. Sandstad, A. N. Shiklomanov, R. B. Skeie, C. J. Smith, S. Smith, K. Tanaka, J. Tsutsui, and Z. Xie. 2020. Reduced Complexity Model Intercomparison Project Phase 1: introduction and evaluation of global-mean temperature response. *Geoscientific Model Development* 13, 11 (2020), 5175–5190. doi:10.5194/gmd-13-5175-2020
- [30] William Nordhaus. 2014. Estimates of the Social Cost of Carbon: Concepts and Results from the DICE-2013R Model and Alternative Approaches. *Journal of the Association of Environmental and Resource Economists* 1, 1 (None 2014), 000. doi:10.1086/676035
- [31] William Nordhaus. 2018. Projections and Uncertainties about Climate Change in an Era of Minimal Climate Policies. *American Economic Journal: Economic*

- Policy* 10, 3 (August 2018), 333–60. doi:10.1257/pol.20170046
- [32] William Nordhaus and Zili Yang. 1996. A Regional Dynamic General-Equilibrium Model of Alternative Climate-Change Strategies. *American Economic Review* 86 (02 1996), 741–65.
- [33] William D. Nordhaus. 1992. An Optimal Transition Path for Controlling Greenhouse Gases. *Science* 258, 5086 (1992), 1315–1319. doi:10.1126/science.258.5086.1315
- [34] William D. Nordhaus. 1993. Reflections on the Economics of Climate Change. *Journal of Economic Perspectives* 7, 4 (December 1993), 11–25. doi:10.1257/jep.7.4.11
- [35] William D. Nordhaus. 2010. Economic aspects of global warming in a post-Copenhagen environment. *Proceedings of the National Academy of Sciences* 107, 26 (2010), 11721–11726. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1005985107 doi:10.1073/pnas.1005985107
- [36] Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. 2021. Benchmarking Multi-Agent Deep Reinforcement Learning Algorithms in Cooperative Tasks. arXiv:2006.07869 [cs.LG] <https://arxiv.org/abs/2006.07869>
- [37] Keywan Riahi, Detlef P. van Vuuren, Elmar Kriegler, Jae Edmonds, Brian C. O'Neill, Shinichiro Fujimori, Nico Bauer, Katherine Calvin, Rob Dellink, Oliver Fricko, Wolfgang Lutz, Alexander Popp, Jesus Crespo Cuaresma, Samir KC, Marian Leimbach, Leiwen Jiang, Tom Kram, Shilpa Rao, Johannes Emmerling, Kristie Ebi, Tomoko Hasegawa, Petr Havlik, Florian Humpenöder, Lara Aleluia Da Silva, Steve Smith, Elke Stehfest, Valentina Bosetti, Jiyong Eom, David Gernaat, Toshihiko Masui, Joeri Rogelj, Jessica Strefler, Laurent Drouet, Volker Krey, Gunnar Luderer, Mathijs Harmsen, Kiyoshi Takahashi, Lavinia Baumstark, Jonathan C. Doelman, Mikiko Kainuma, Zbigniew Klimont, Giacomo Marangoni, Hermann Lotze-Campen, Michael Obersteiner, Andrzej Tabeau, and Massimo Tavoni. 2017. The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global Environmental Change* 42 (2017), 153–168. doi:10.1016/j.gloenvcha.2016.05.009
- [38] James Rudd-Jones, Mirco Musolesi, and María Pérez-Ortiz. 2025. Multi-Agent Reinforcement Learning Simulation for Environmental Policy Synthesis. arXiv:2504.12777 [cs.MA] <https://arxiv.org/abs/2504.12777>
- [39] James Rudd-Jones, Fiona Thendean, and María Pérez-Ortiz. 2024. Crafting desirable climate trajectories with RL explored socio-environmental simulations. arXiv:2410.07287 [physics.soc-ph] <https://arxiv.org/abs/2410.07287>
- [40] Ivan Savin, Creutzig Felix, Filatova Tatiana, Foramitti Joël, Konc Théo, Niamir Leila, Safarzynska Karolina, and van den Bergh Jeroen. 2023. Agent-based modeling to integrate elements from different disciplines for ambitious climate policy. *WIREs Climate Change* 14, 2 (2023), e811. arXiv:https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcc.811 doi:10.1002/wcc.811
- [41] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347 [cs.LG] <https://arxiv.org/abs/1707.06347>
- [42] Jian Shen, Hang Lai, Minghuan Liu, Han Zhao, Yong Yu, and Weinan Zhang. 2023. Adaptation Augmented Model-based Policy Optimization. *Journal of Machine Learning Research* 24, 218 (2023), 1–35. <http://jmlr.org/papers/v24/22-0606.html>
- [43] Elke Stehfest, Detlef Vuuren, Tom Kram, Alexander Bouwman, Rob Alkemade, Michel Bakkenes, H. Biemans, A. Bouwman, Michel Elzen, Jan Janse, Jelle van Minnen, M. Müller, and Anne Prins. 2014. *Integrated Assessment of Global Environmental Change with IMAGE 3.0. Model description and policy applications*. PBL Publishers.
- [44] Weber Theodore, Corotan A., Hutchinson B., Kravitz B., and Link R. 2020. Technical note: Deep learning for creating surrogate models of precipitation in Earth system models. *Atmospheric Chemistry and Physics* 20, 4 (2020), 2303–2317. doi:10.5194/acp-20-2303-2020
- [45] Tianyu Zhang, Andrew Williams, Soham Phade, Sunil Srinivasa, Yang Zhang, Prateek Gupta, Yoshua Bengio, and Stephan Zheng. 2022. AI for Global Climate Cooperation: Modeling Global Climate Negotiations, Agreements, and Long-Term Cooperation in RICE-N. arXiv:2208.07004 [cs.LG] <https://arxiv.org/abs/2208.07004>