Ai2's Climate Emulators:
What do we do when they don't learn the right thing?

Chris Bretherton, Senior Director of Climate Modeling Emeritus
representing the Ai2 Climate Modeling team who did all the work



Chris Bretherton Spencer Clark Brian Henn Anna Kwa Jeremy McGibbon Andre Perkins Oli Watt-Meyer Elynn Wu James Duncan Troy Arcomano

Ai2

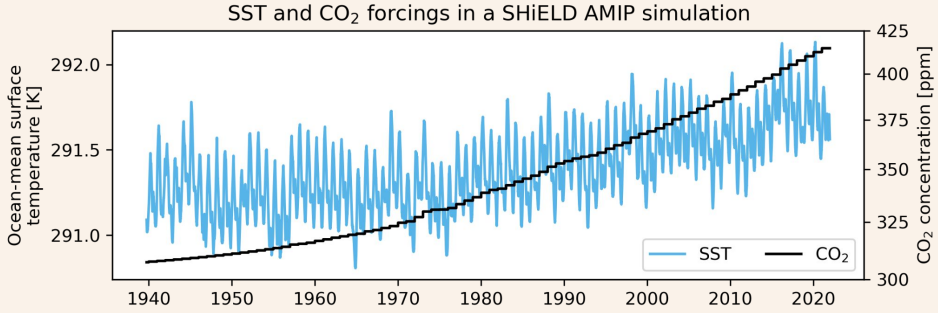
External partners

M²LInES GFDL LLNL NVIDIA

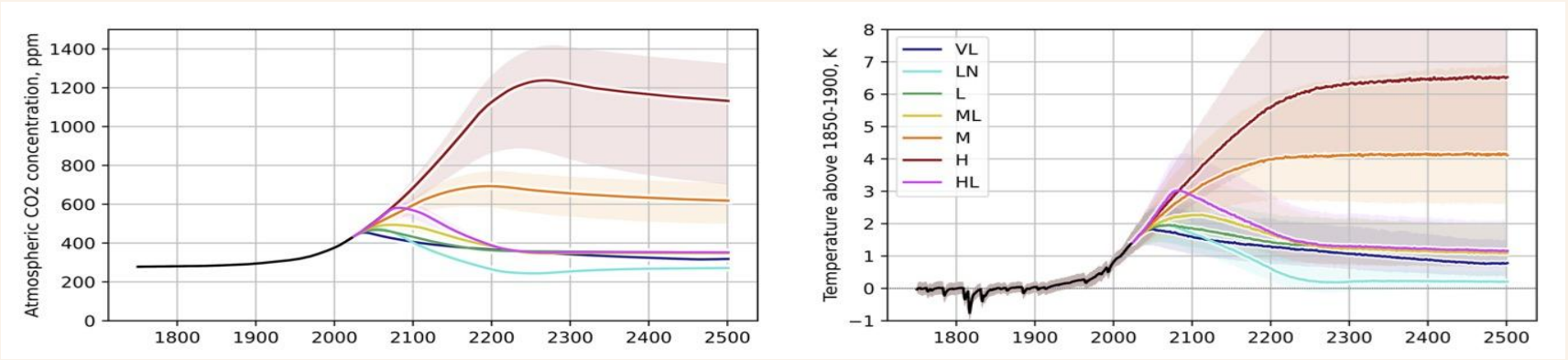
CO_2 effects on climate change: a key challenge for ACE

Focus of this talk:

Disentangling atmospheric response due to CO_2 (rapid adjustment) vs. SST in ACE2, a seamless ML weather-climate emulator of the global atmosphere, when both have similar historical trends



In contrast, many CMIP7 scenarios have periods with decreasing CO_2 but increasing temperature



ScenarioMIP-CMIP7; *van Vooren et al. (2026)*

Ai2

Disentangling AMIP climate trends due to CO_2 vs. SST

Three key CFMIP-inspired tests of physical correctness of climate changes due to CO_2 vs. SST

- Surface warming, stratospheric cooling, and ΔTOA radiation from ΔCO_2 with fixed SST
- Surface warming pattern with a uniform dSST = +4 K
- Abrupt 4xCO_2 with a coupled ocean

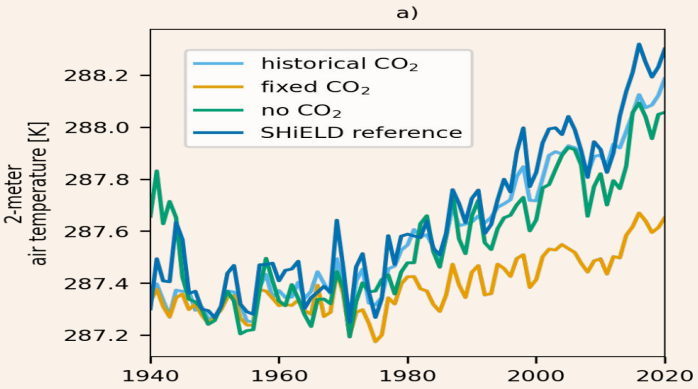
These tests help assess ACE2's suitability as the atmospheric component of a coupled earth system emulator for anthropogenically forced climate change

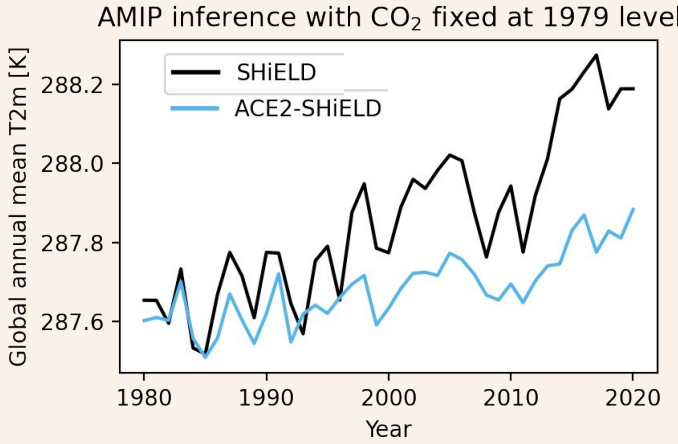
Does ACE2 trained on SHiELD pass these tests? If so, what training data is required?

Could it do so with historical reanalysis and accurately understood physics alone?

Surface temperature response to CO_2 with specified SST

- We trained ACE2 to make 6 hr forecasts of GFDL's SHiELD AGCM, here considered the 'truth'
- Training dataset: 'historical-like' 1940-2020 AMIP simulation including varying CO_2
- Train ACE with and without including CO_2 as an input feature
- Inference testing vs. other SHiELD realizations on varying and constant CO_2 cases
- By using SHiELD rather than ERA5 as the reference, we have 'truth' for these sensitivity tests



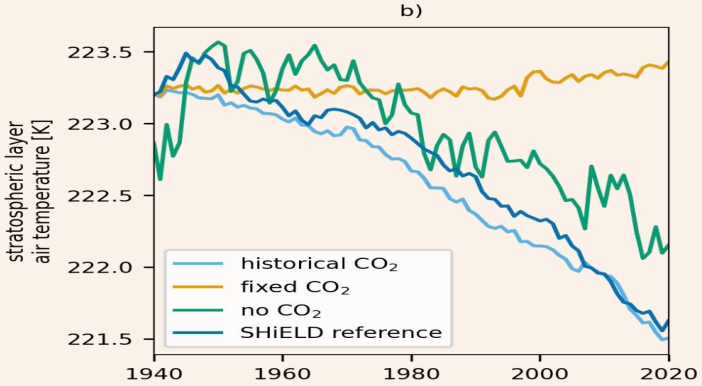


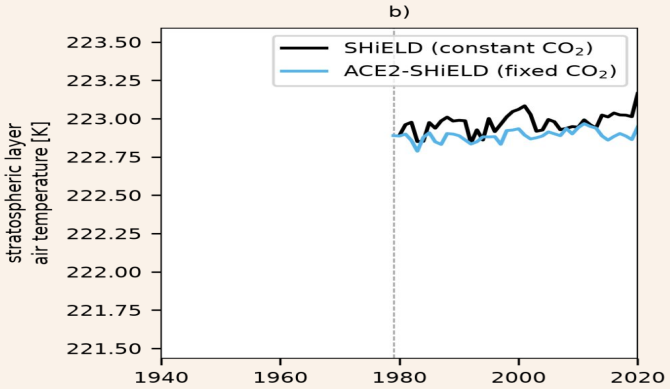
ACE2 gets SHiELD's T2m trend right for training case, even w/o CO_2 input, but overestimates its sensitivity to fixing CO_2

Watt-Meyer et al. (2025)

Ai2

Stratospheric temperature response to CO_2 is better





- ACE2 gets SHiELD's stratospheric temperature trend right for training case, even (qualitatively) without giving CO_2 as an input .
  In the latter case, has ACE2 learned CO_2 as function of SST?

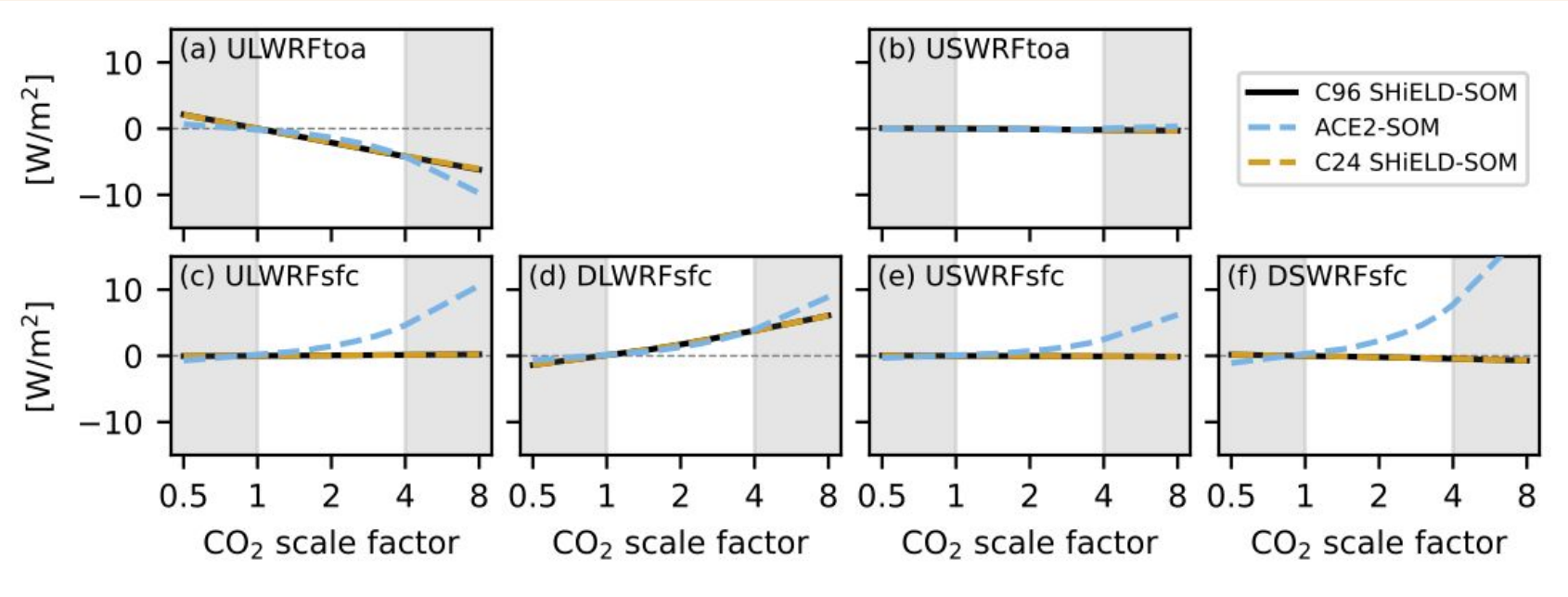
- ACE2 with CO_2 as an input also correctly predicts no stratospheric temperature trend for fixed CO_2

Ai2

Watt-Meyer et al. (2025)

Radiative sensitivity to CO_2 (when a feature) with fixed SST

'Truth' comes from SHiELD radiation multi-calls (doesn't include rapid adjustment)



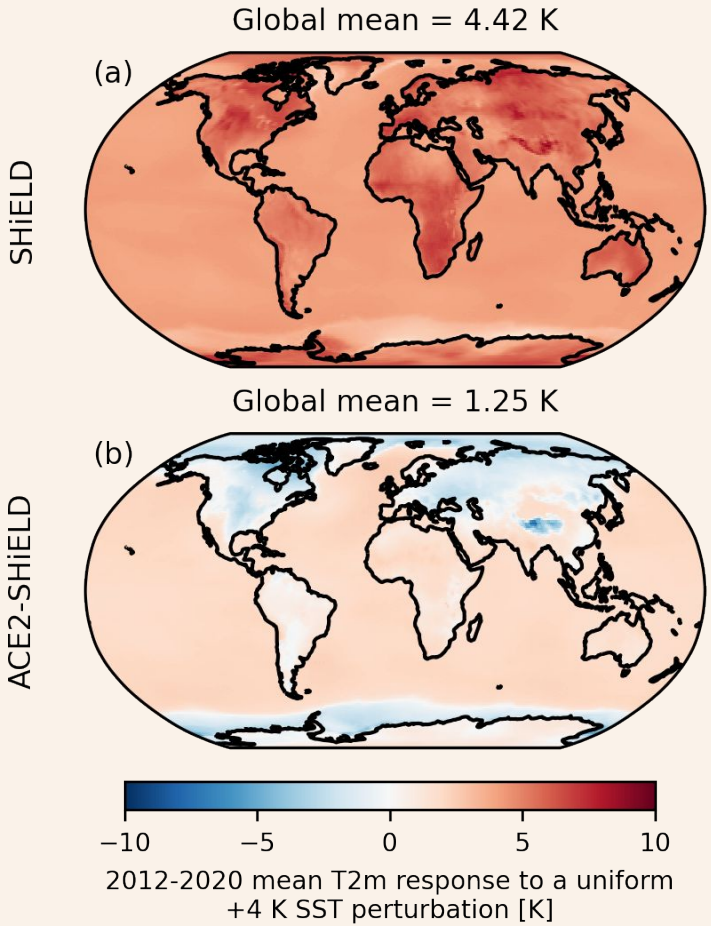
ACE2 mis-estimates shortwave and longwave flux sensitivities to CO_2 at TOA and surface

Ai2

Clark et al. (2025)

ACE2 +4K dSST response

- ACE2-SHiELD trained on 1940-2020 AMIP reference simulation generalizes poorly in this out-of-sample test.
- So far, we can reduce but not eliminate this bias with small architectural tweaks
- With historical training alone, ACE2 again incorrectly separates climate responses to CO_2 and SST.

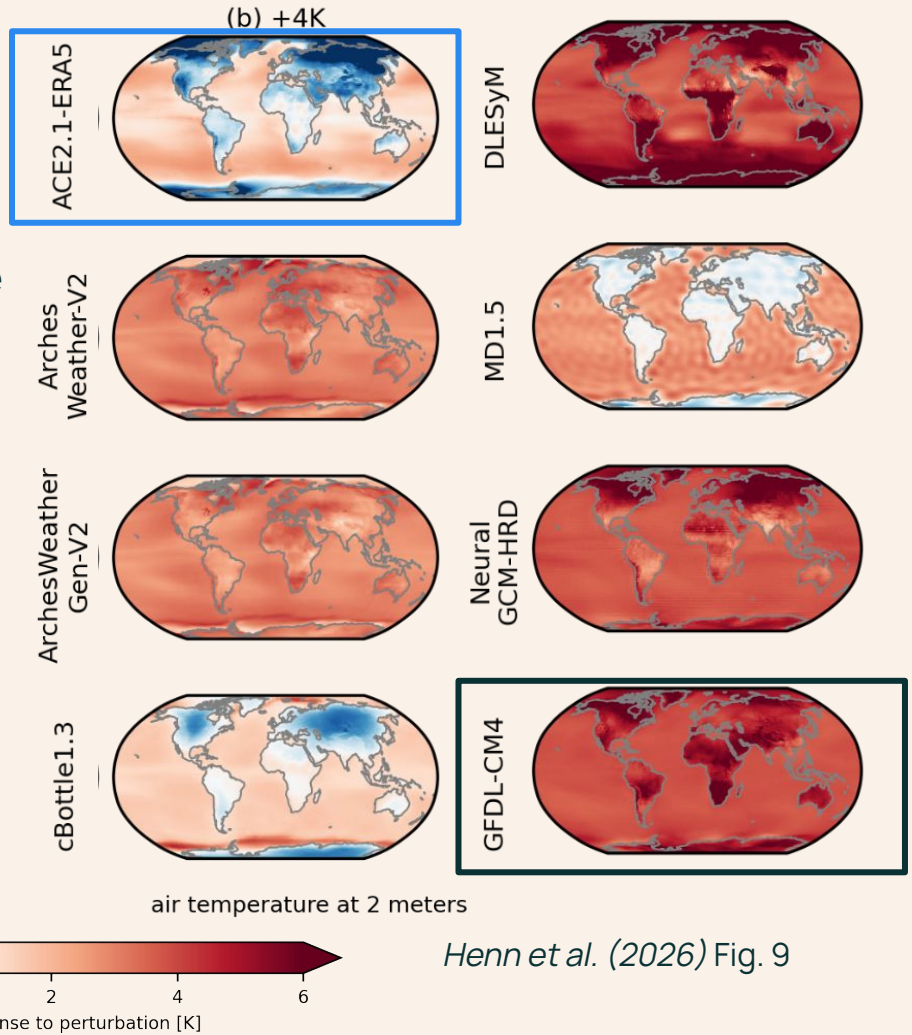


Clark et al. (2026) Fig. 6

Ai2

ACE2 +4K dSST response: AIMIP intercomparison

- AIMIP: ACE2 trained on **1979-2014 ERA5 without CO_2 as an input** generalizes worse in this out-of-sample test (with similar biases to ACE2-SHiELD) vs. some other AI models with different architectures.
- The CM4 GCM comparison has no CO_2 change, but arguably should include SST-correlated CO_2 increase

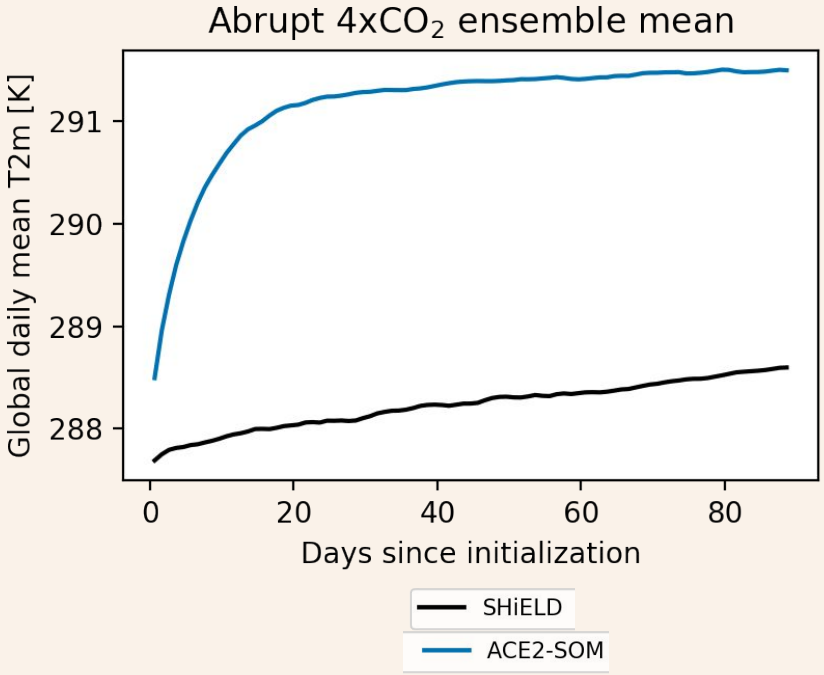


Henn et al. (2026) Fig. 9

Ai2

Abrupt 4xCO_2 response of ACE2 coupled to slab ocean model

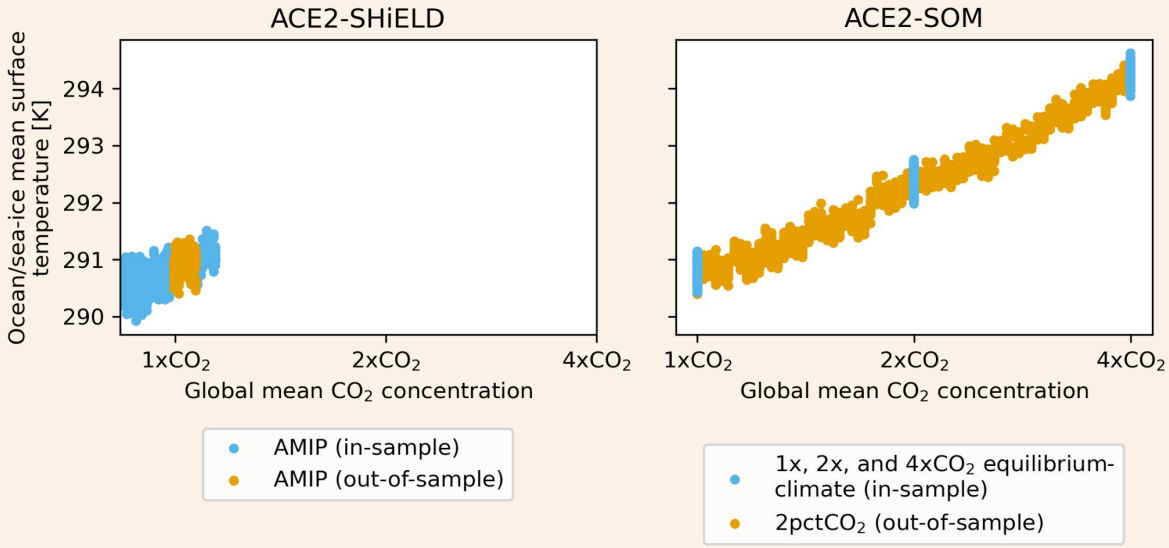
- ACE2-SOM is trained on SHiELD-SOM run in equilibrium (1,2,4)xCO_2 climates
- An 'out-of-sample' test due to the initial combination of 4xCO_2 and unperturbed SST
- ACE atmospheric temperature adjusts far too rapidly, although ultimately the climate reaches the correct equilibrium warming.



Clark et al. (2025), Fig. 10

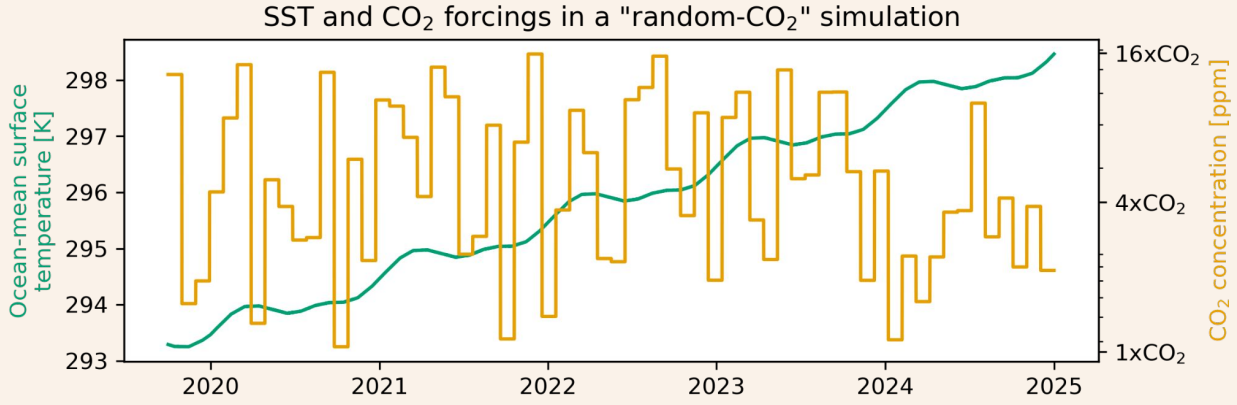
AGCM emulation solution: Add custom training datasets

- Training data for ACE2-SHiELD & ACE2-SOM lack samples with mismatched SST and CO_2.
- This is OK for out-of-sample scenarios along the same line as the training data in the SST-CO_2 phase space.
- But climate change can deviate from this line



Random-CO_2 simulations efficiently span SST-CO_2 space

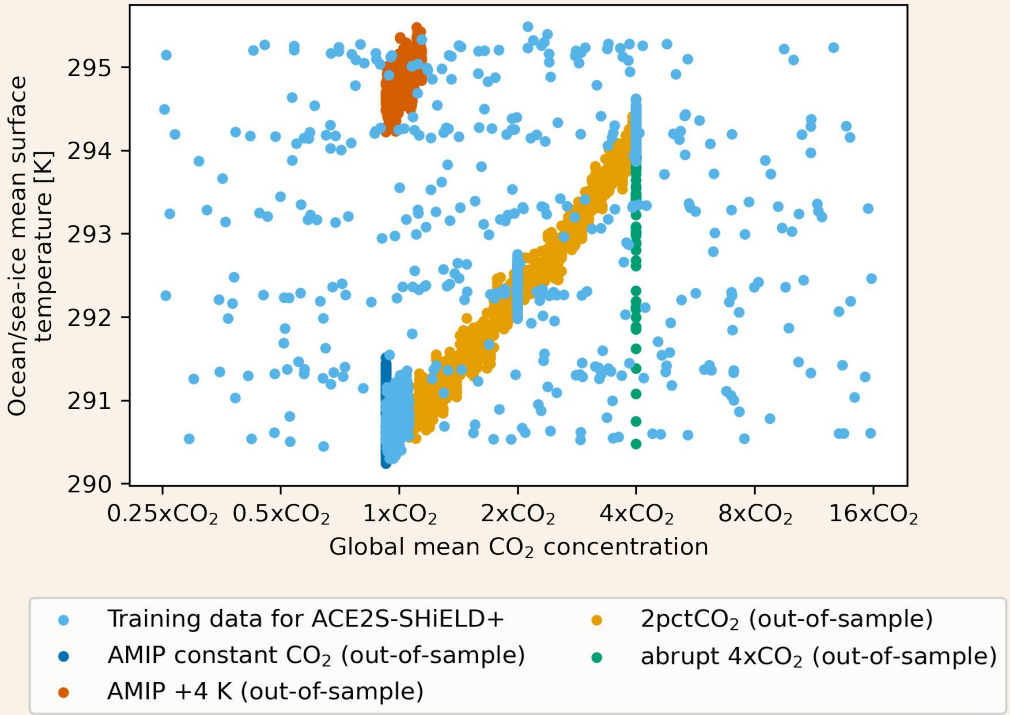
Random-CO_2 SHiELD simulations are run with annually repeating climatological SST with a uniform 1 K/year trend superposed and CO_2 that randomly changes every 30 days.



Simplified Fig. 1d of *Clark et al. (2026), arXiv, in review at JGR-ML*

We train a new ACE2S-SHiELD+ model on an equal mixture of AMIP, equilibrium-climate, and random-CO_2 data

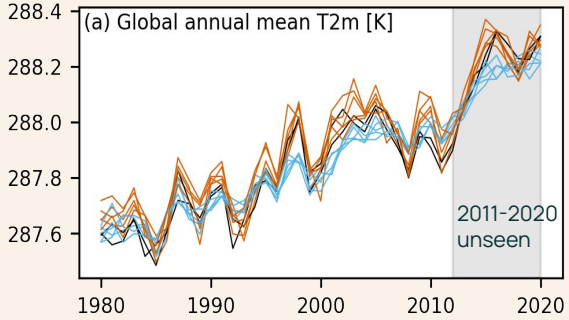
New training dataset spans SST/CO_2 combinations of interest



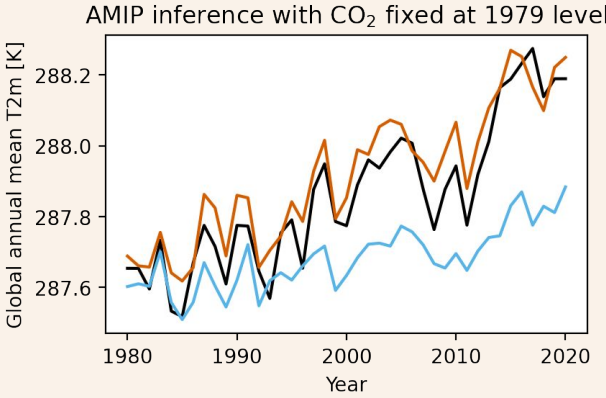
Clark et al. (2026), Fig. 2b

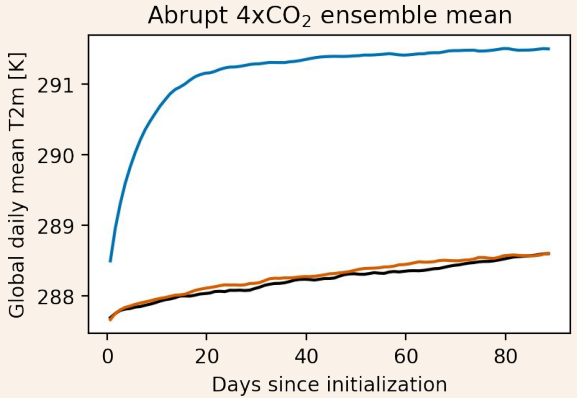
Ai2

Random-CO_2 multiclimate training disentangles effects of SST and CO_2 so ACE2S-SHiELD+ generalizes to unseen scenarios

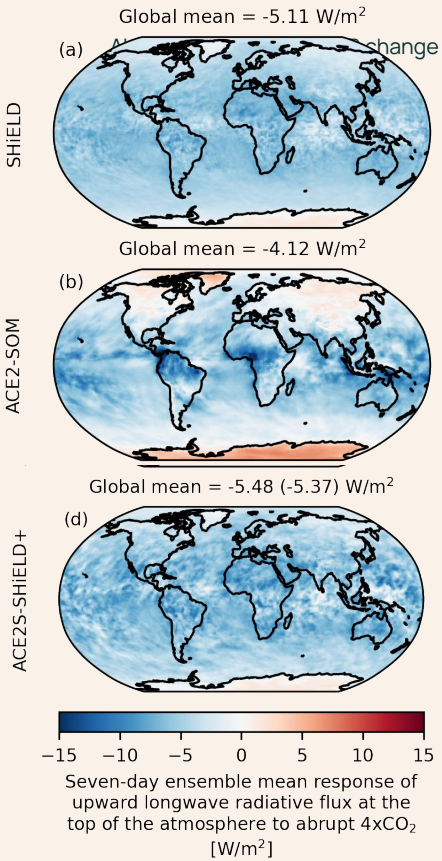


Clark et al. (2026),
Figs. 3,5,6,S10









Abrupt 4xCO_2

'Truth'

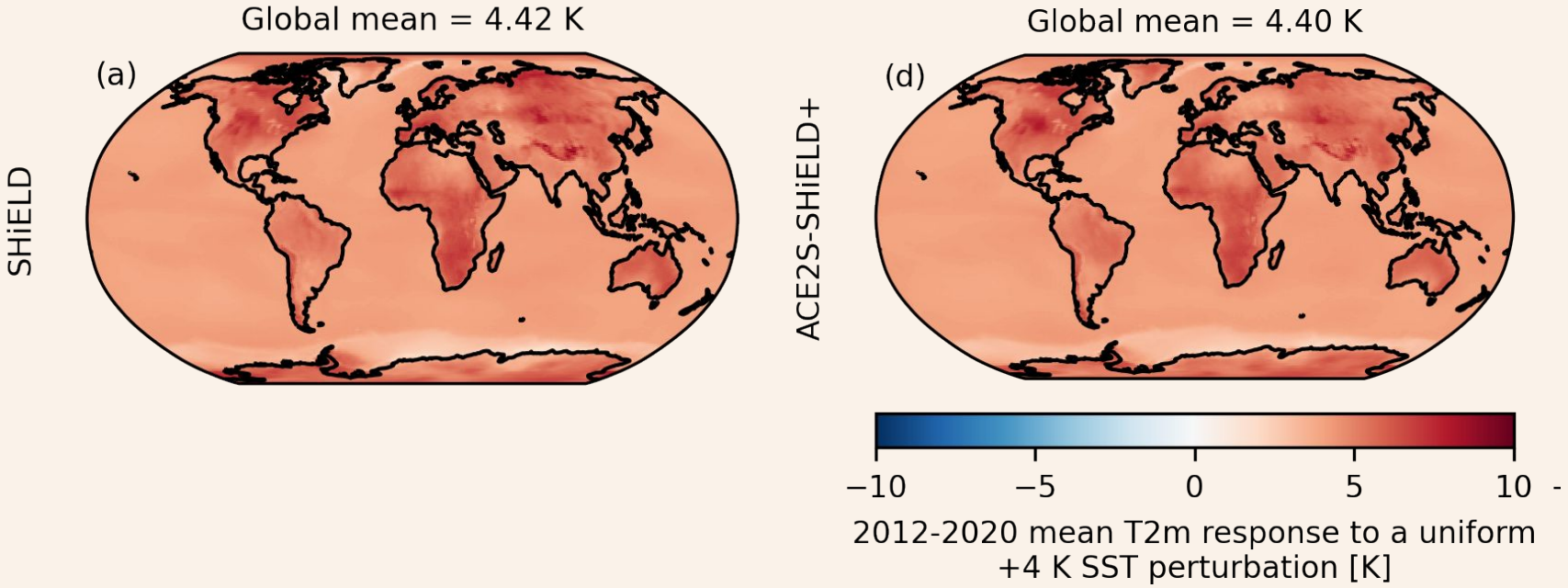
Large OLR biases

Small OLR biases

Ai2

Random-CO_2 multiclimate training disentangles effects of SST and CO_2 so ACE2S-SHiELD+ generalizes to unseen scenarios

dSST = 4 K response is now nearly perfect!



Clark et al. (2026), Fig. 7

Ai2

Conclusions

For our ACE2 weather-climate emulator:

- Extrapolation to unseen climates is unreliable
- Idealized CFMIP tests for physics-based models are physically interpretable by climate scientists and isolate issues with our ML model
- ML solution: Custom training data separating SST and CO_2 sensitivity across timescales
- Unfortunately, this solution won't apply to an ML emulator trained only on historical data
- The grand challenge of an AI weather-climate model trained only on historical data that generalizes reliably to future climate change remains to be solved.
- Ai2 Climate Modeling is hiring 3 people for a 3-year project to extend NeuralGCM into a coupled ocean-atmosphere hybrid model for multidecadal climate forecasting trained on historical data + known physics. A plum for postdocs and new PhDs with strong AI background and physical science training : https://allenai.org/careers

Ai2