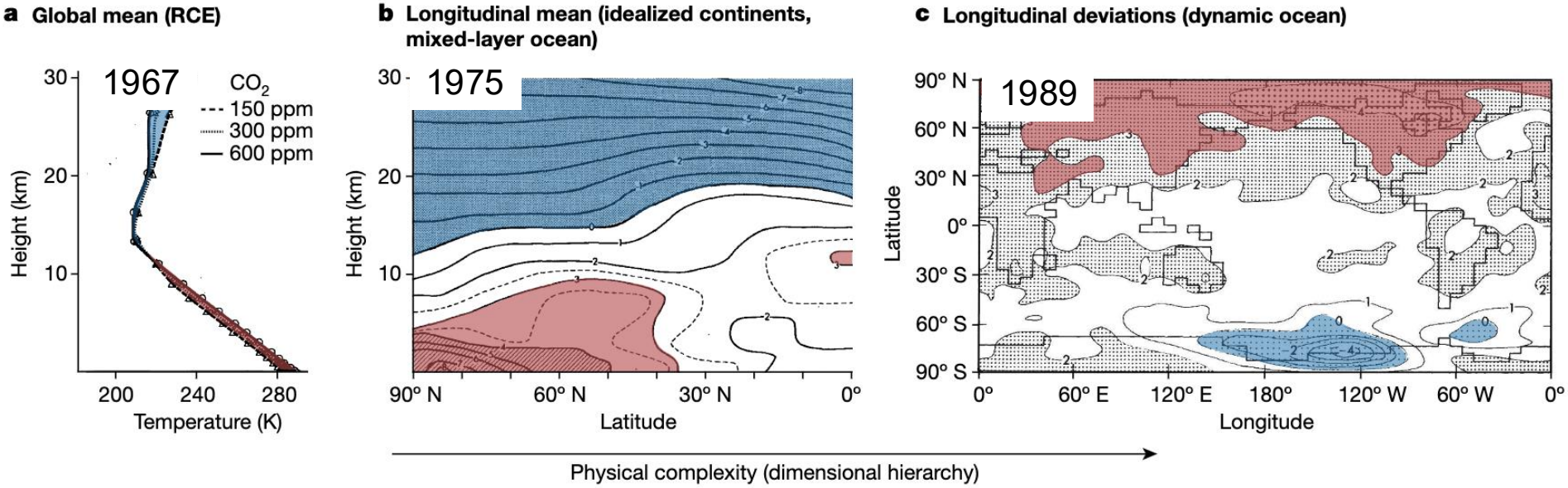
Right answer for the right reason



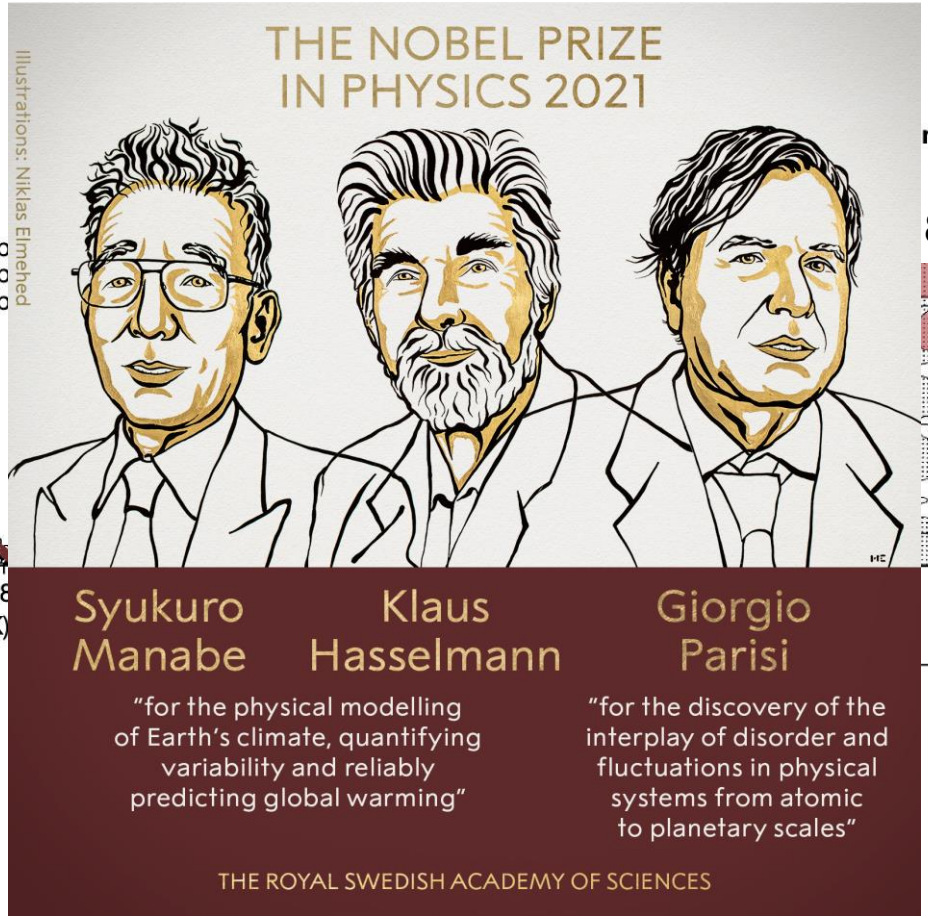
Tiffany A. Shaw
The University of Chicago

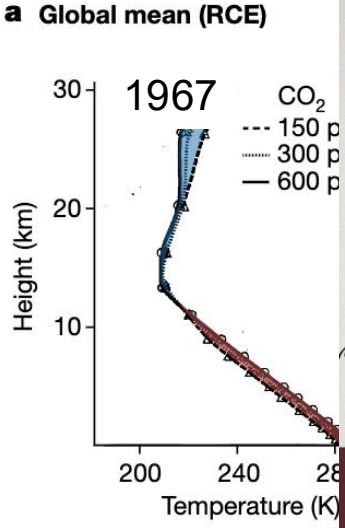
Manabe got the right answer for the right reason

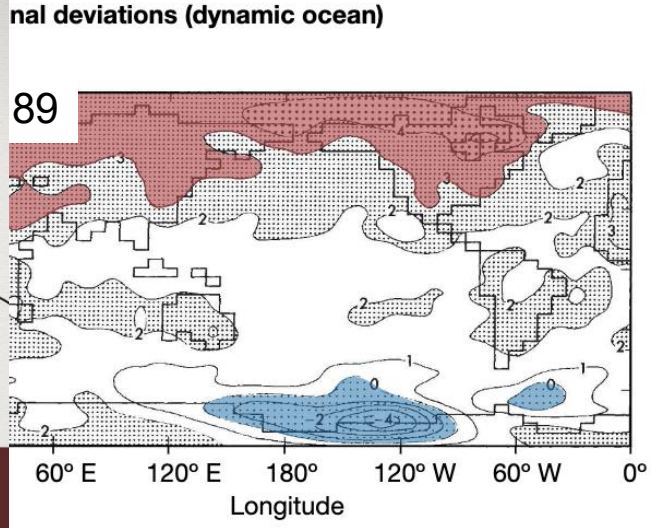


Shaw & Stevens (2025, Nature)

Manabe got the right answer for the right reason





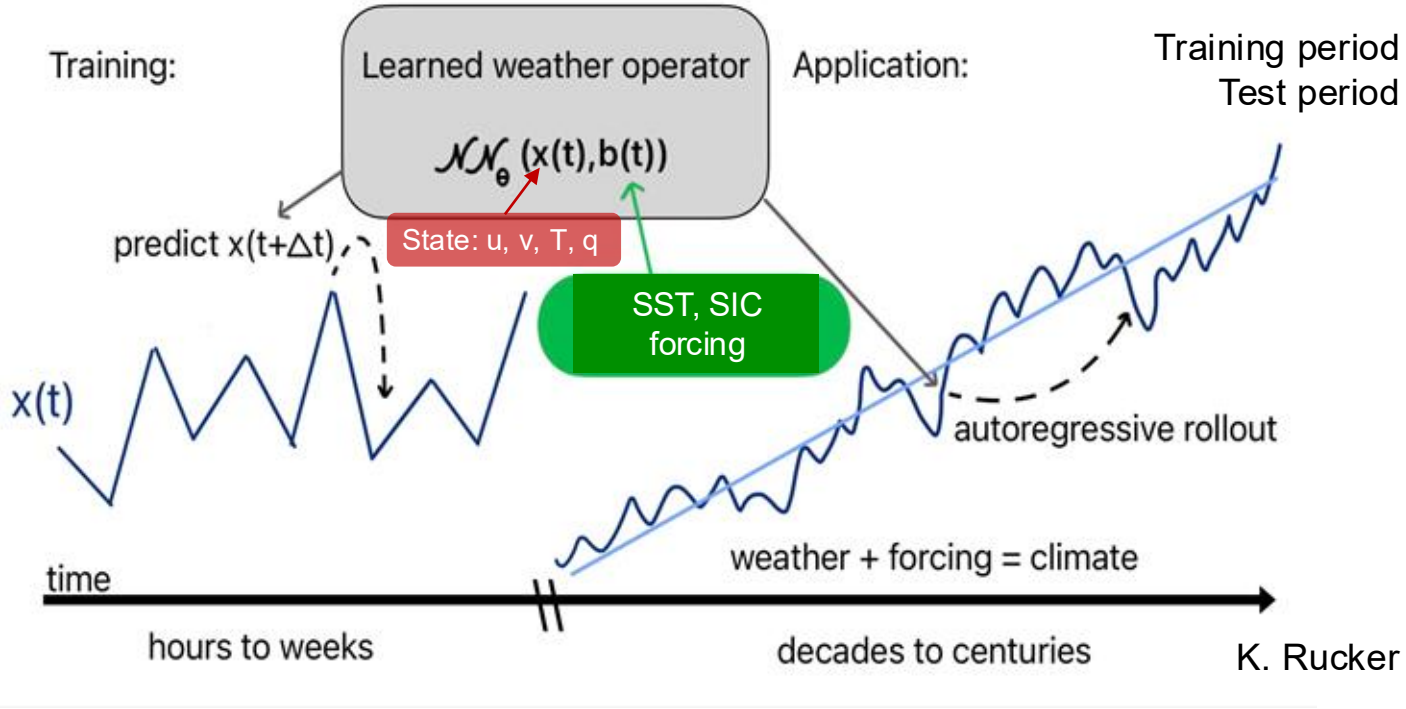


Shaw & Stevens (2025, Nature)

Emerging trend signals are a test for AI models

Signal	Physical Mechanism	Failure mode
Tropical upper-tropospheric warming		
Southwest US drying		
Poleward Jet shift		
……		

Apply test to AI atmosphere emulators - AIMIP

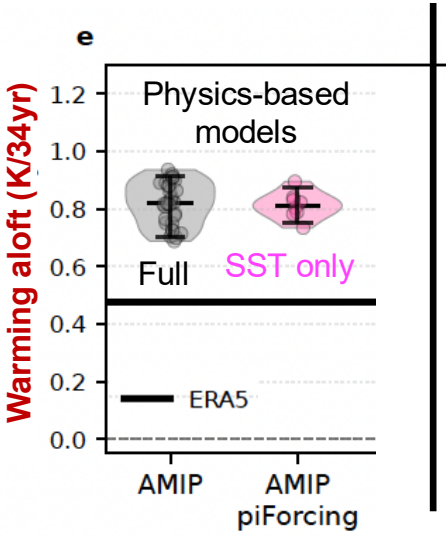


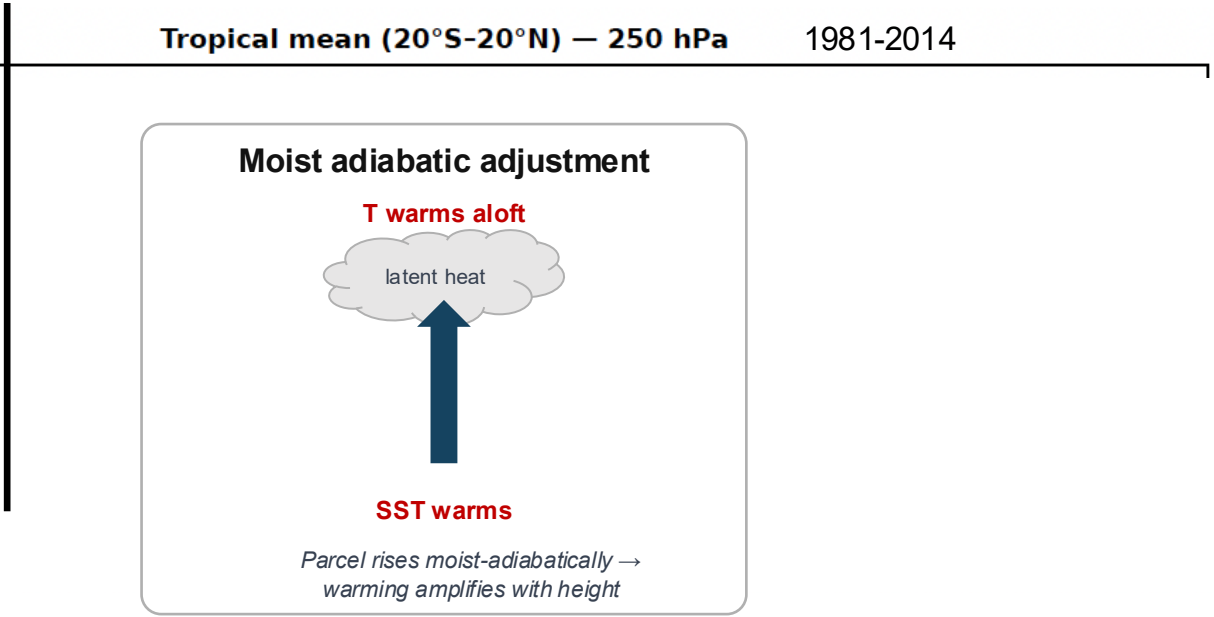
Training period (1979-2014)
Test period (2015-2024)

7 models to sample epistemic (structural) uncertainty
5 member ensembles to sample aleatoric (initial condition) uncertainty

Henn et al. (2026, arXiv)

Physics-based models show SST forcing drives upper tropospheric warming



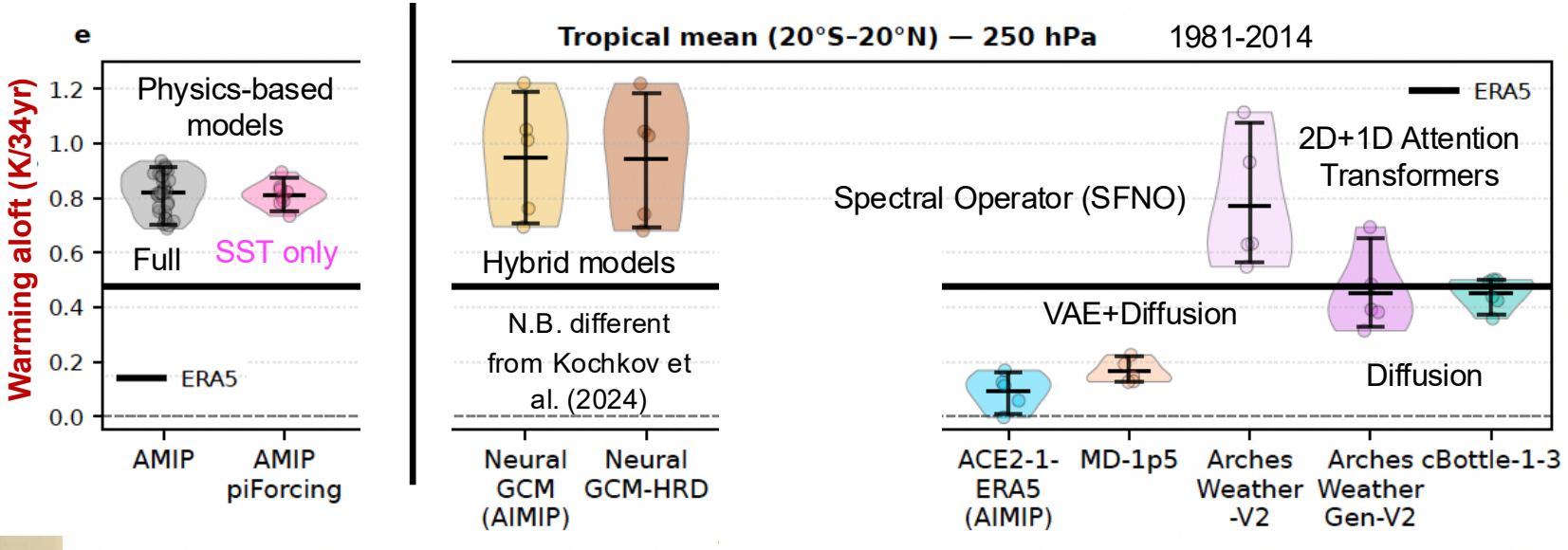




24 models sample epistemic (structural) uncertainty
103 member ensembles sample aleatoric (initial condition) uncertainty

Baxter et al. (2026, in prep)

Large architecture-dependent spread reveals inconsistent physical learning of SST forcing

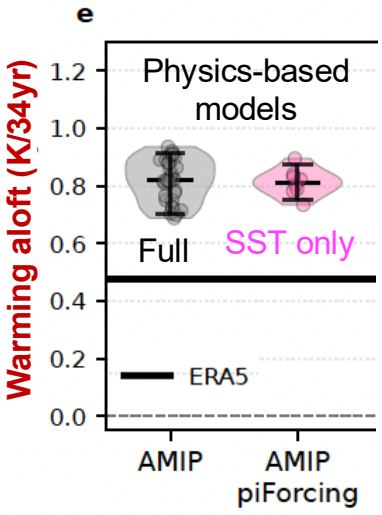


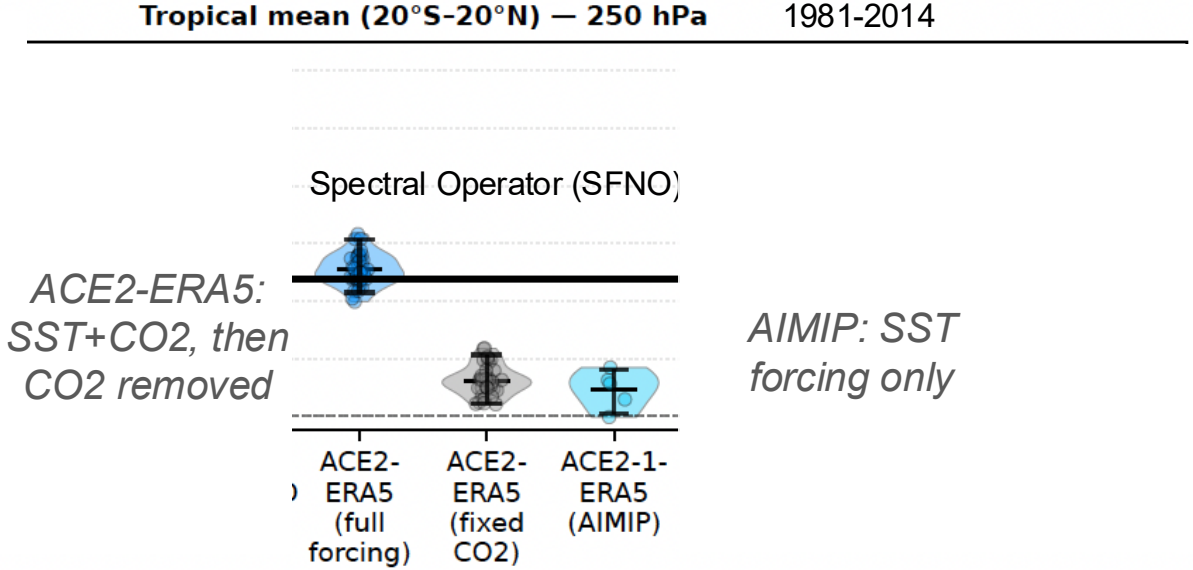


Training on ERA5 does not guarantee trend skill

Baxter et al. (2026, in prep)

ACE2 matches the warming trend under full forcing but not once CO2 is removed

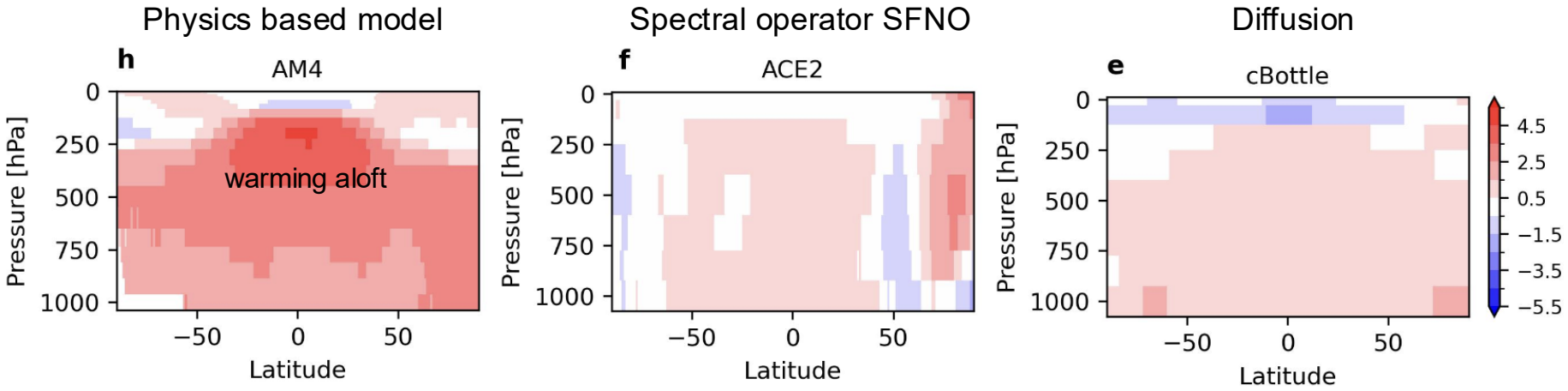






Baxter et al. (2026, in prep)
Rucker et al. (2026, GRL)

No upper-tropospheric warming under counterfactual +2K SST indicates malattribution



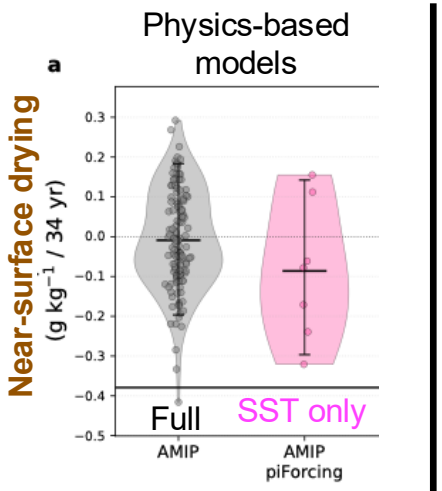
*Similar issue for land temperature trends & + 2K

Zhang & Merlis (2026, JAMES)

Emerging trend signals are a test for AI models

Signal	Physical Mechanism	Failure mode
Tropical upper-tropospheric warming	Moist adiabatic adjustment	**Architecture dependent, Right trend, wrong reason; collapses under +2K counterfactual**
Southwest US drying		
Poleward jet shift		
……		

Physics-based models don’t capture SW US drying driven by land-atmosphere coupling



US Southwest 700 hPa Specific Humidity Trends (1981-2014)

Land–atmosphere coupling

free troposphere (dry)

LCL rises

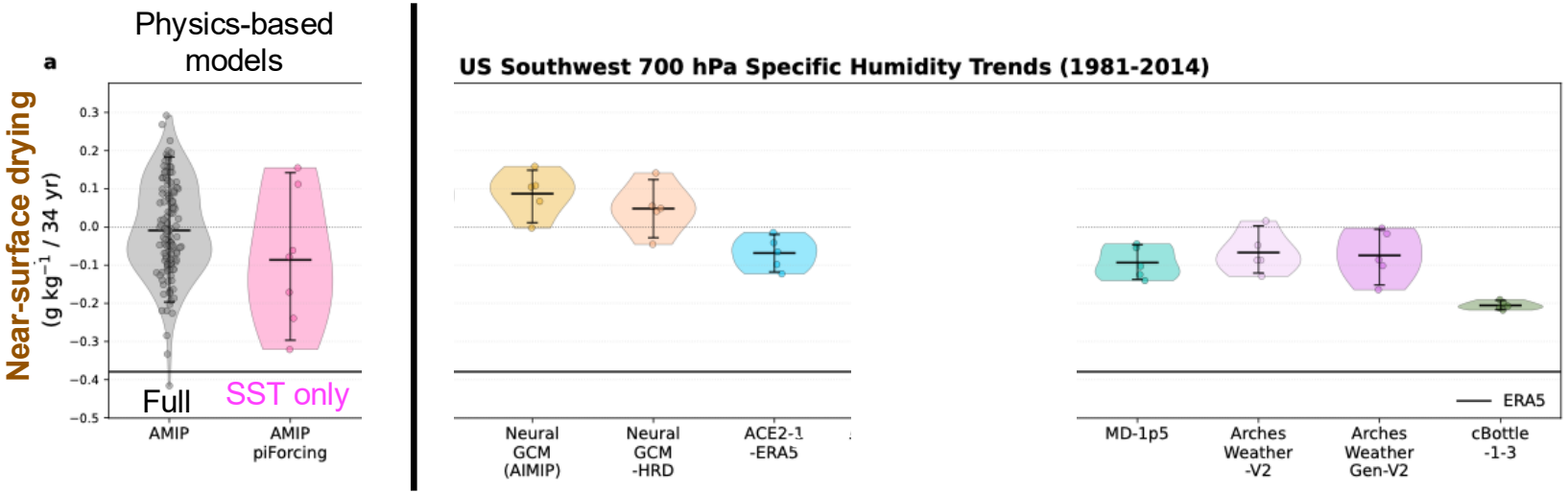
drier soil / surface dries

Entrainment aloft (blue) and soil drying (brown) both raise the LCL — same signature, either way.



Baxter et al. (2026, in prep)

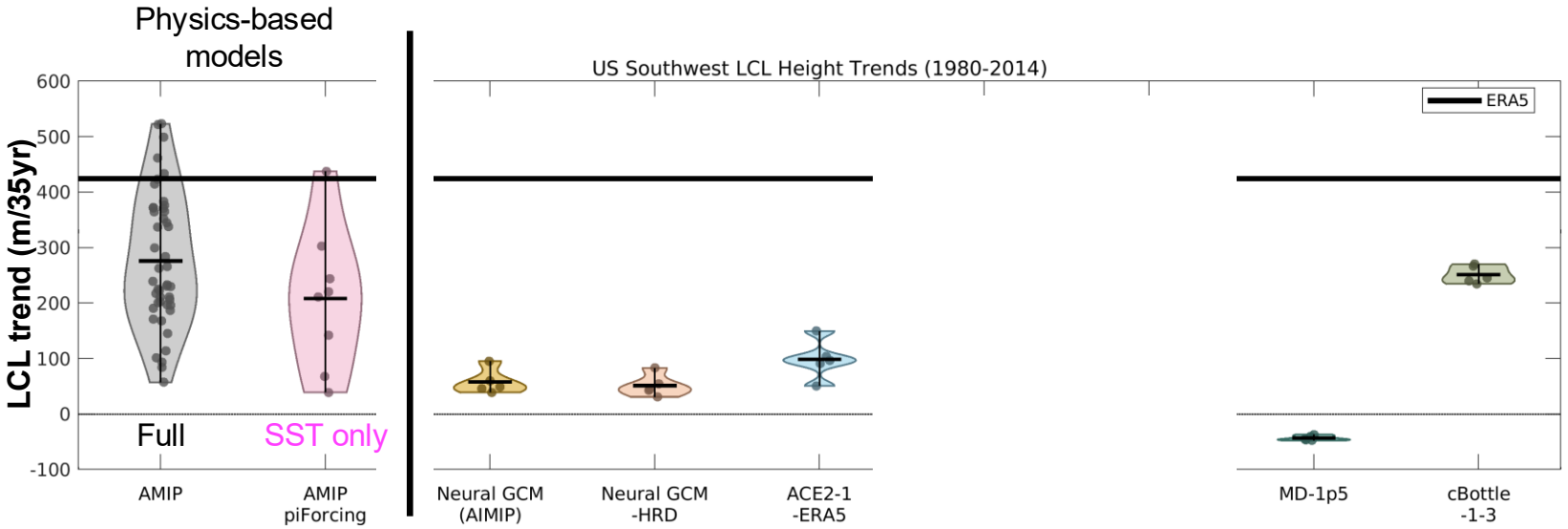
No architecture captures SW US drying from SST forcing





Baxter et al. (2026, in prep)

No architecture learns LCL trend from SST forcing



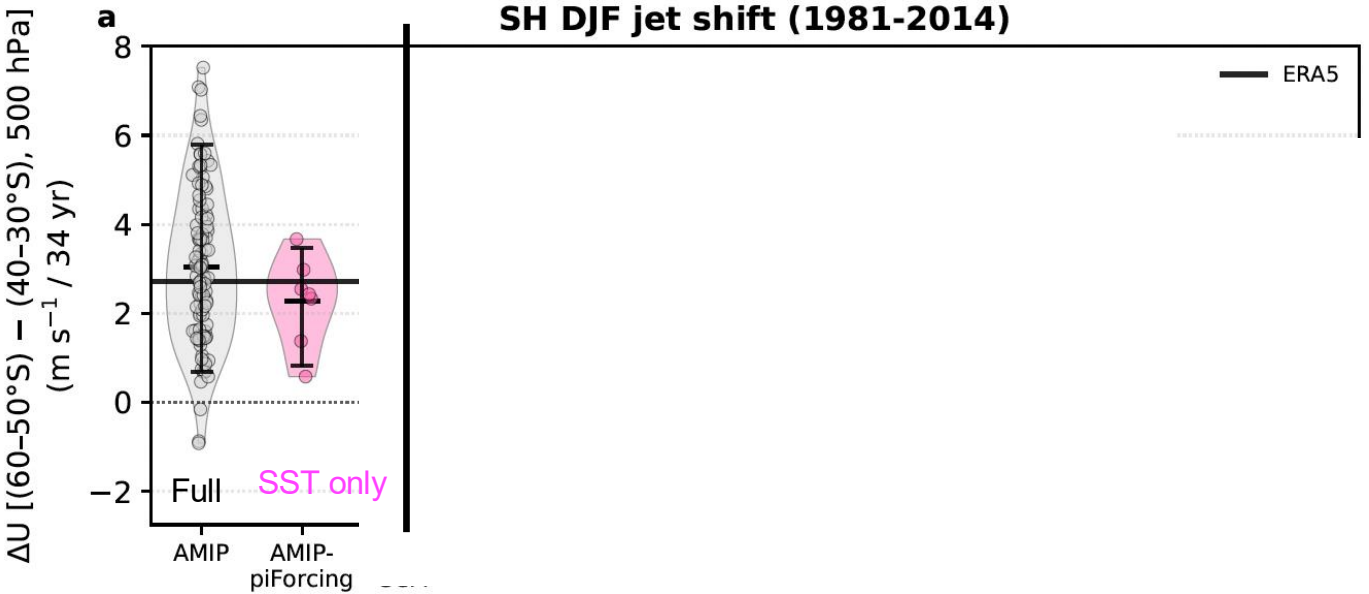


Baxter et al. (2026, in prep)

Emerging trend signals are a test for AI models

Signal	Physical Mechanism	Failure mode
Tropical upper-tropospheric warming	Moist adiabatic adjustment	**Architecture dependent, Right trend, wrong reason; collapses under +2K counterfactual**
SW US drying	Atmosphere land coupling, rising LCL	**Architecture independent, missing LCL coupling mechanism**
Poleward jet shift		
……		

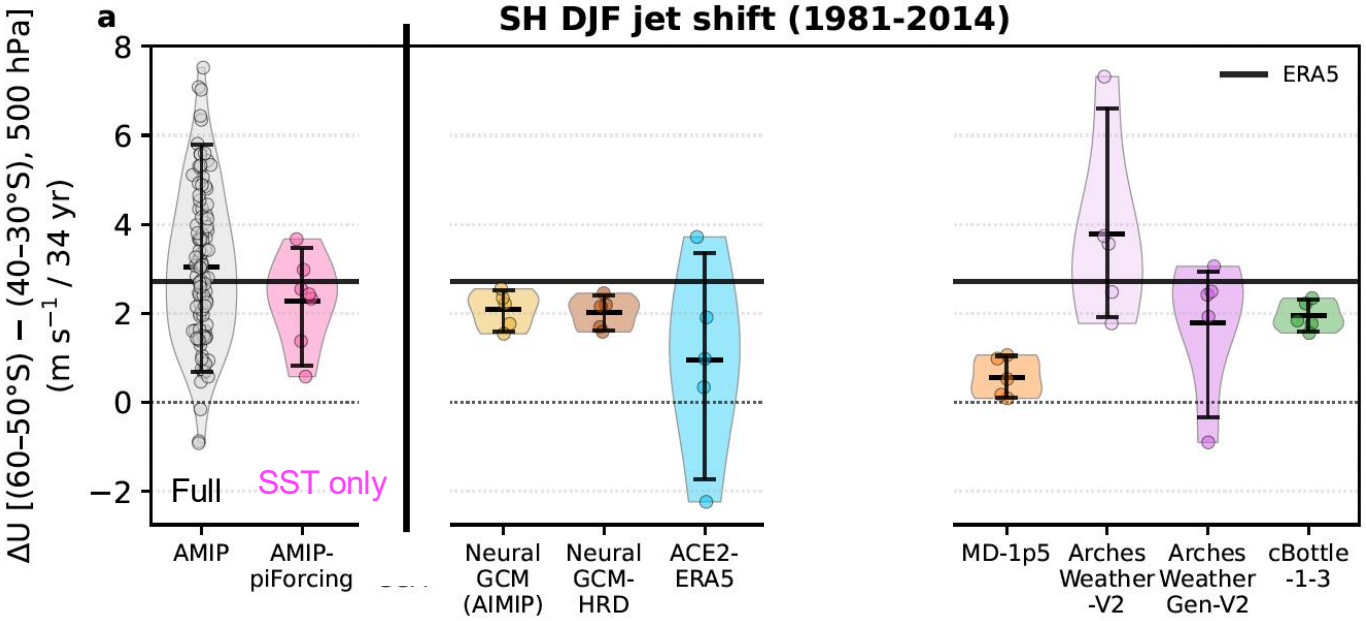
Physics-based models capture poleward shift driven by SST forcing





Baxter et al. (2026, in prep)

Large architecture-dependent spread reveals inconsistent physical learning of the jet shift

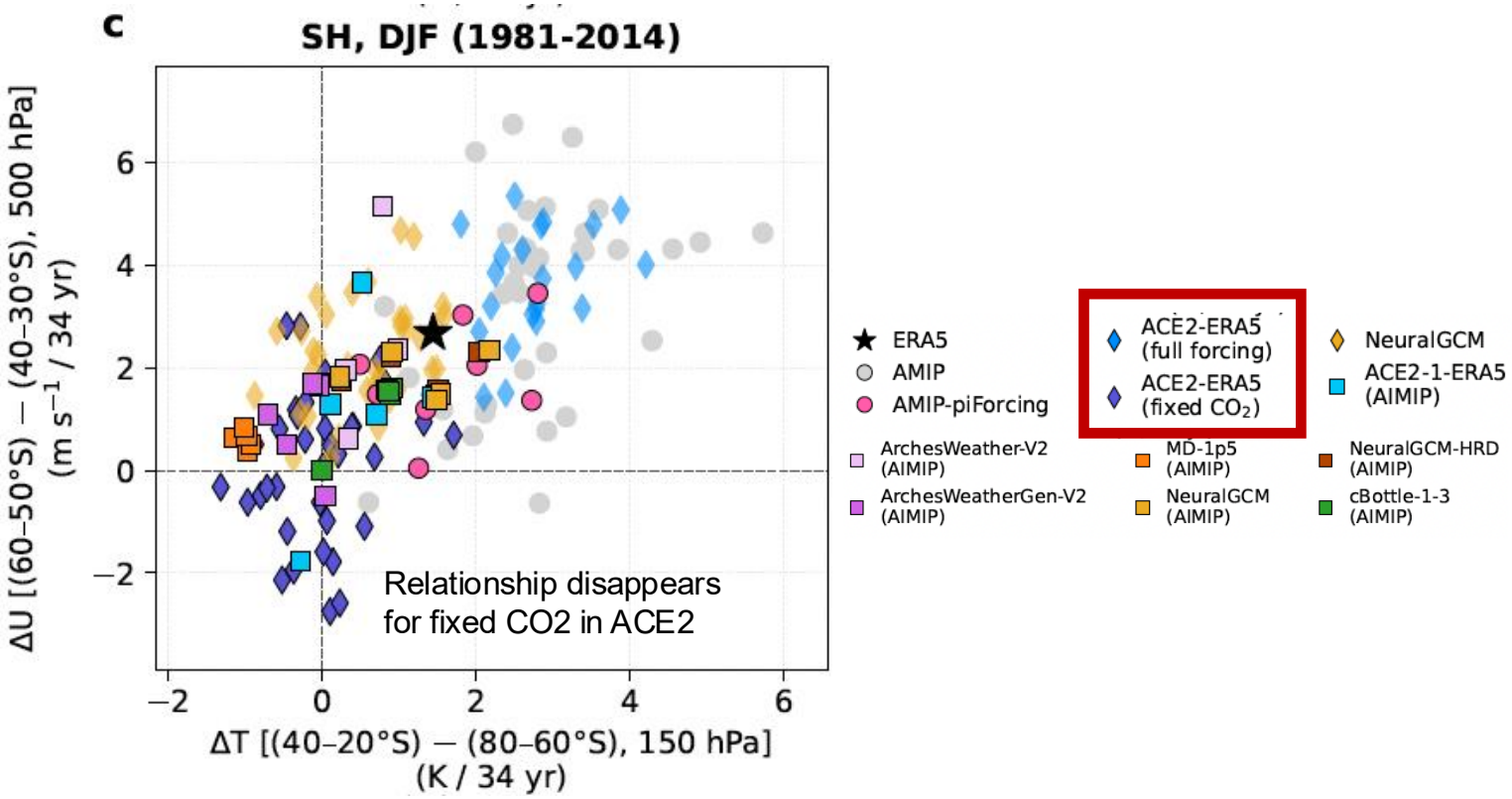




Baxter et al. (2026, in prep)

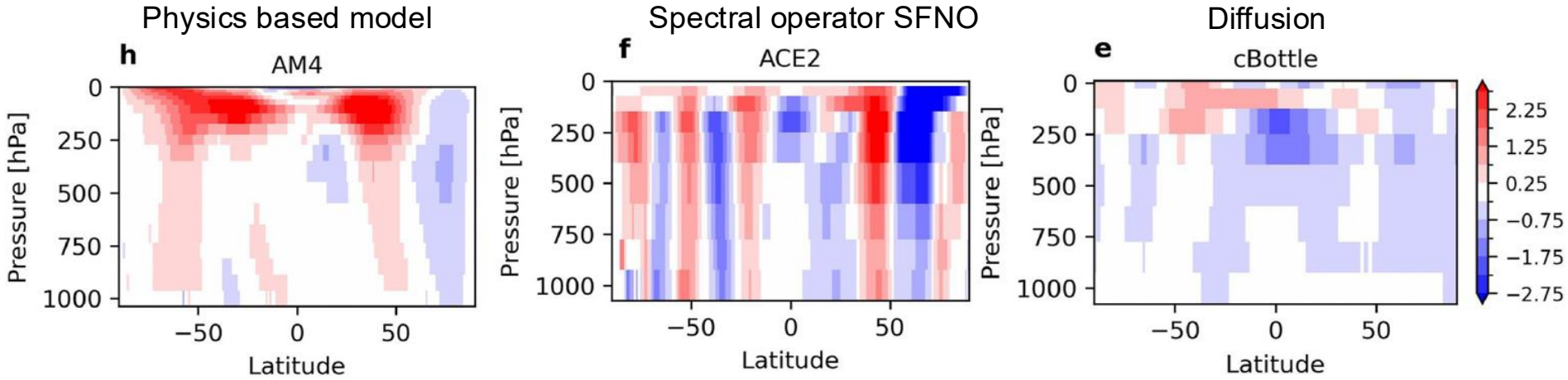
Poleward shift connected to temperature gradient





Baxter et al. (2026, in prep)

Jet shift response collapses under +2K SST warming

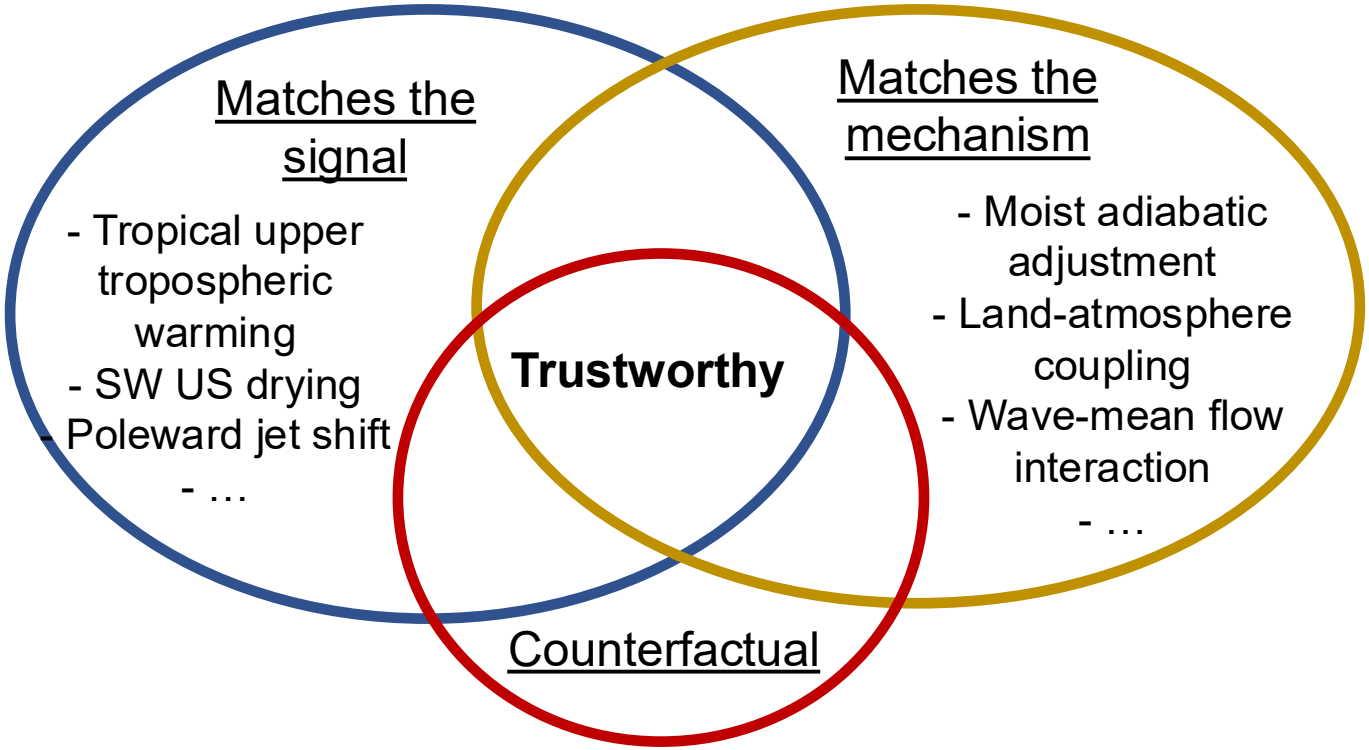


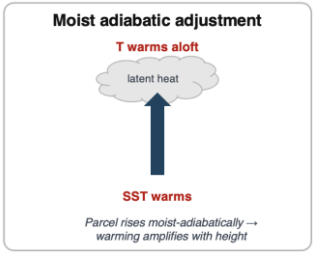
Zhang & Merlis (2026, JAMES)

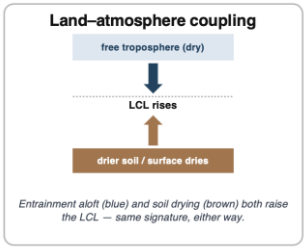
Emerging trend signals are a test for AI models

Signal	Physical Mechanism	Failure mode
Tropical upper-tropospheric warming	Moist adiabatic adjustment	**Architecture dependent, Right trend, wrong reason; collapses under +2K counterfactual**
SW US drying	Atmosphere land coupling, rising LCL	**Architecture independent, missing LCL coupling mechanism**
Poleward jet shift	Temperature gradient, wave-mean flow interaction	**Architecture dependent, Right trend, wrong reason; collapses under +2K counterfactual**
……		

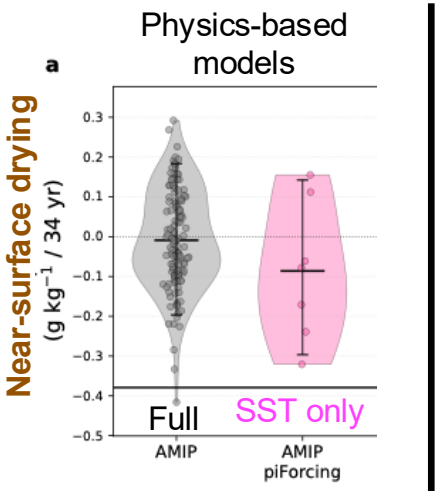
Getting the right answer requires testing the signal, mechanism & verification via counterfactual

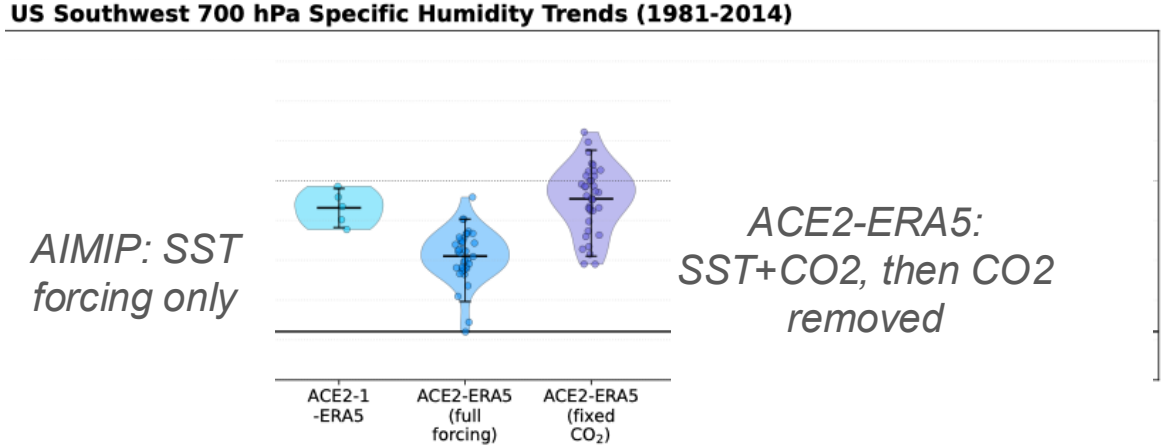






ACE2 gets closer to drying trend under full forcing — but not once CO2 is removed

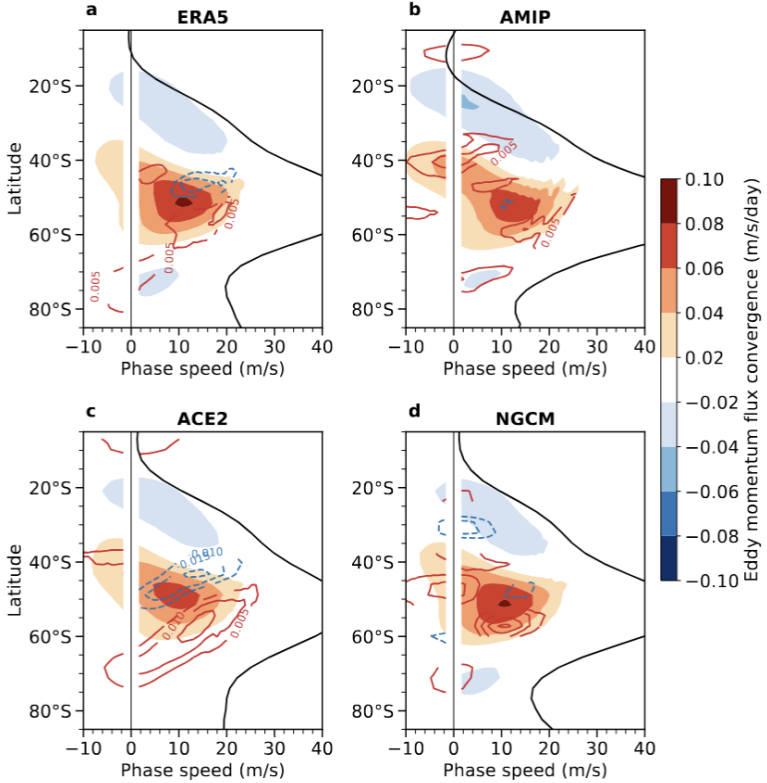




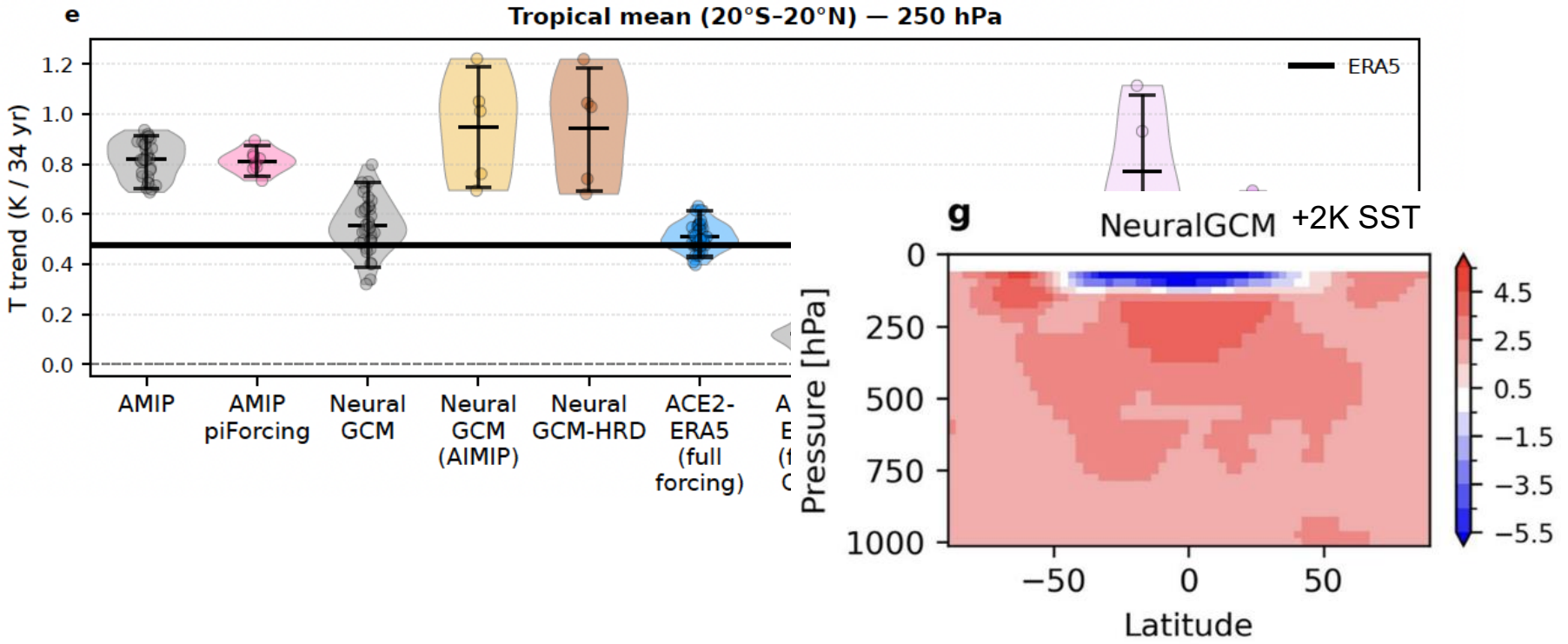


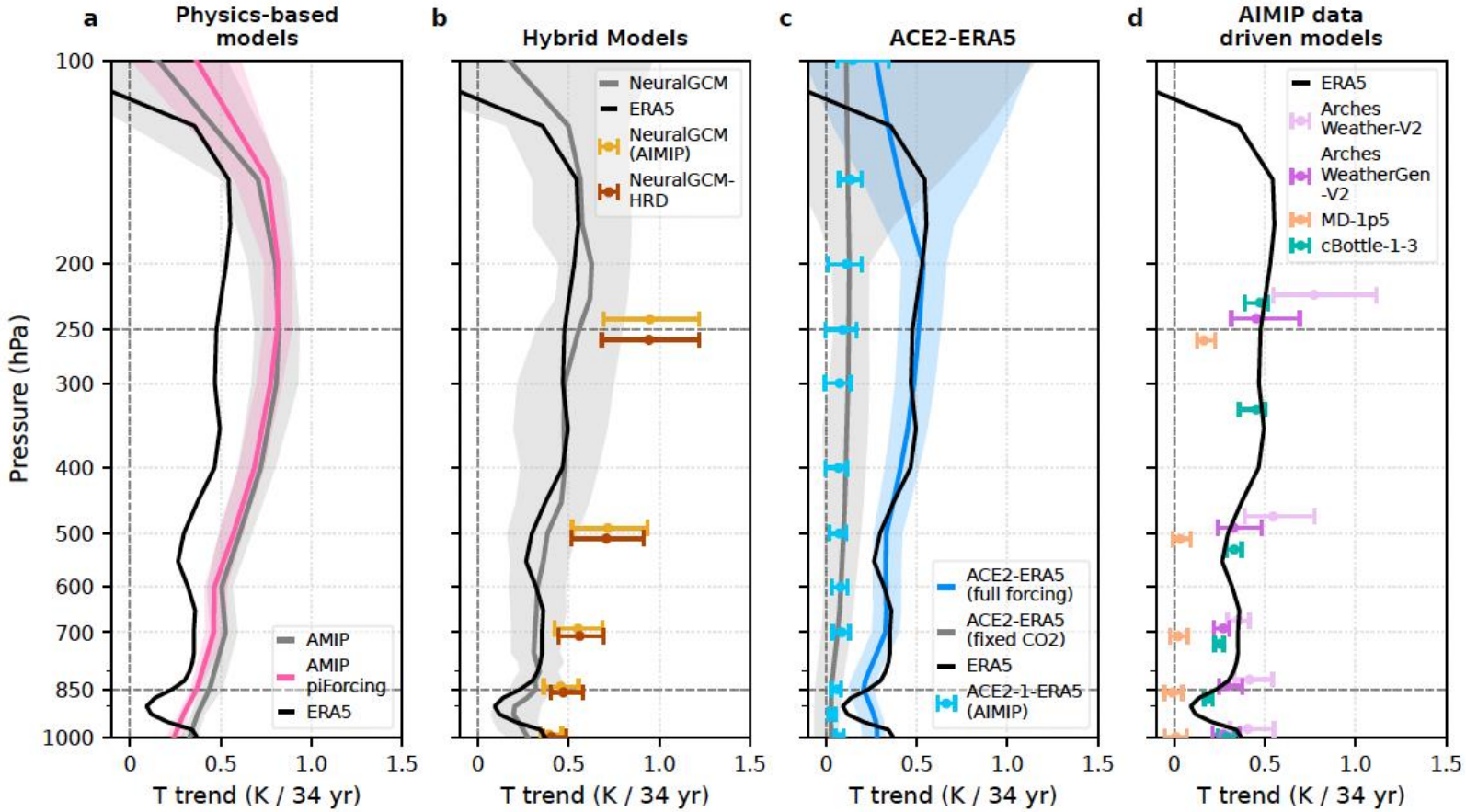
Baxter et al. (2026, in prep)

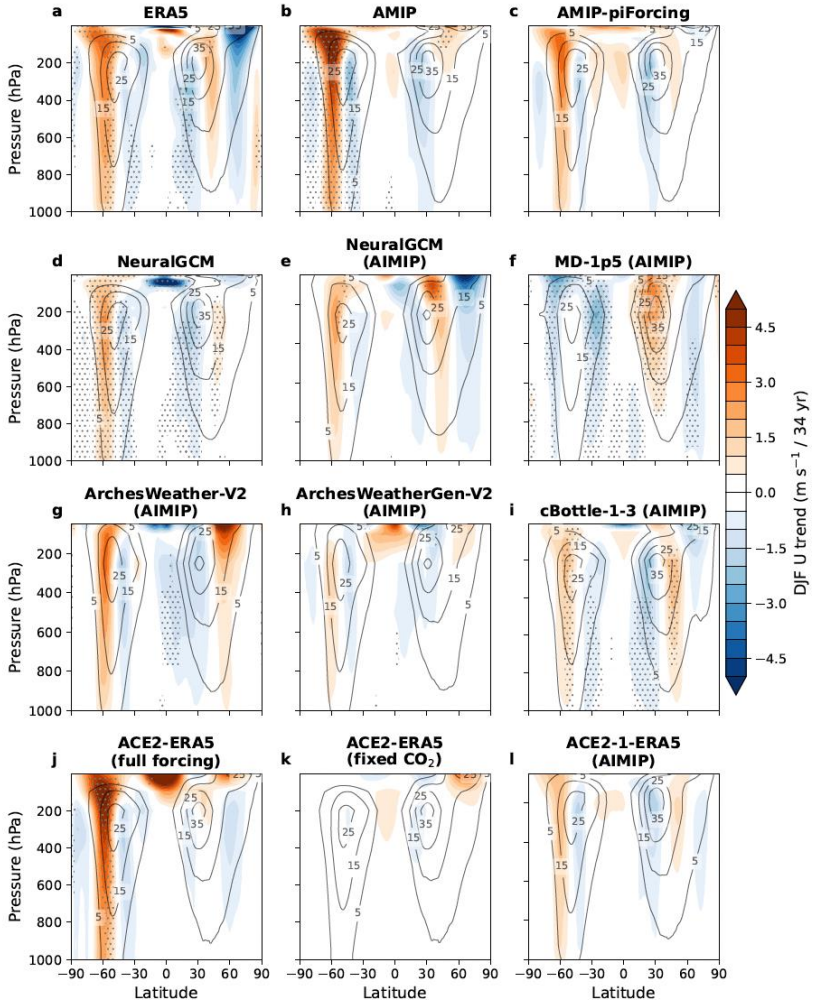
Poleward jet shift consistent with critical line and eddy momentum flux trends



Neural GCM reproducibility issue: Kochkov vs AIMIP

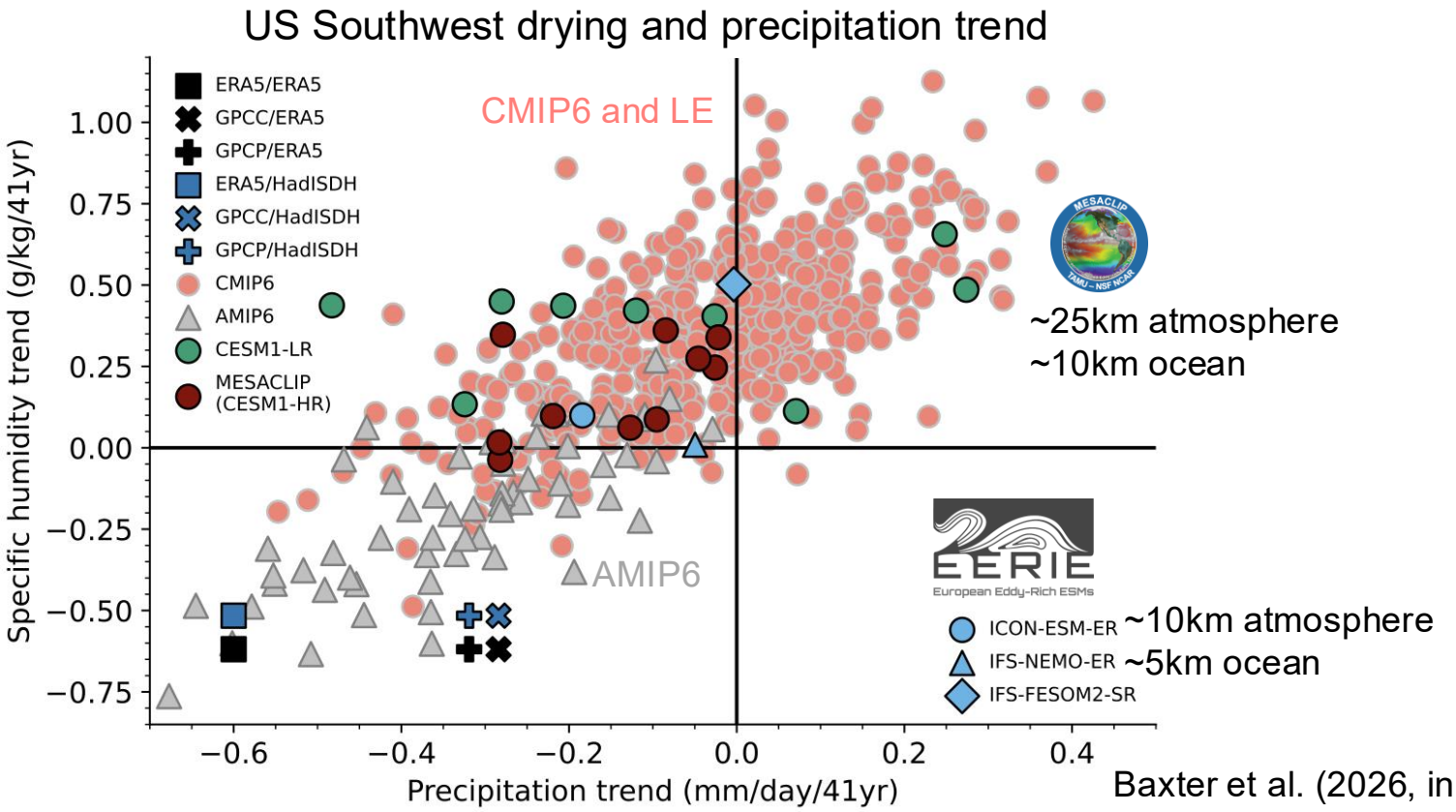



a
Physics-based models
b
Hybrid Models
c
ACE2-ERA5
d
AIMIP data driven models
Pressure (hPa)
100
200
250
300
400
500
600
700
850
1000
0
0.5
1.0
1.5
T trend (K / 34 yr)
AMIP
AMIP piForcing
ERA5
NeuralGCM
ERA5
NeuralGCM (AIMIP)
NeuralGCM-HRD
ACE2-ERA5 (full forcing)
ACE2-ERA5 (fixed CO2)
ERA5
ACE2-1-ERA5 (AIMIP)
ERA5
Arches Weather-V2
Arches WeatherGen -V2
MD-1p5
cBottle-1-3


a ERA5
b AMIP
c AMIP-piForcing
d NeuralGCM
e NeuralGCM (AIMIP)
f MD-1p5 (AIMIP)
g ArchesWeather-V2 (AIMIP)
h ArchesWeatherGen-V2 (AIMIP)
i cBottle-1-3 (AIMIP)
j ACE2-ERA5 (full forcing)
k ACE2-ERA5 (fixed CO_2)
l ACE2-1-ERA5 (AIMIP)
Pressure (hPa)
Latitude
DJF U trend (m s^{-1} / 34 yr)

Km-scale shows no significant improvement in US Southwest drying discrepancy



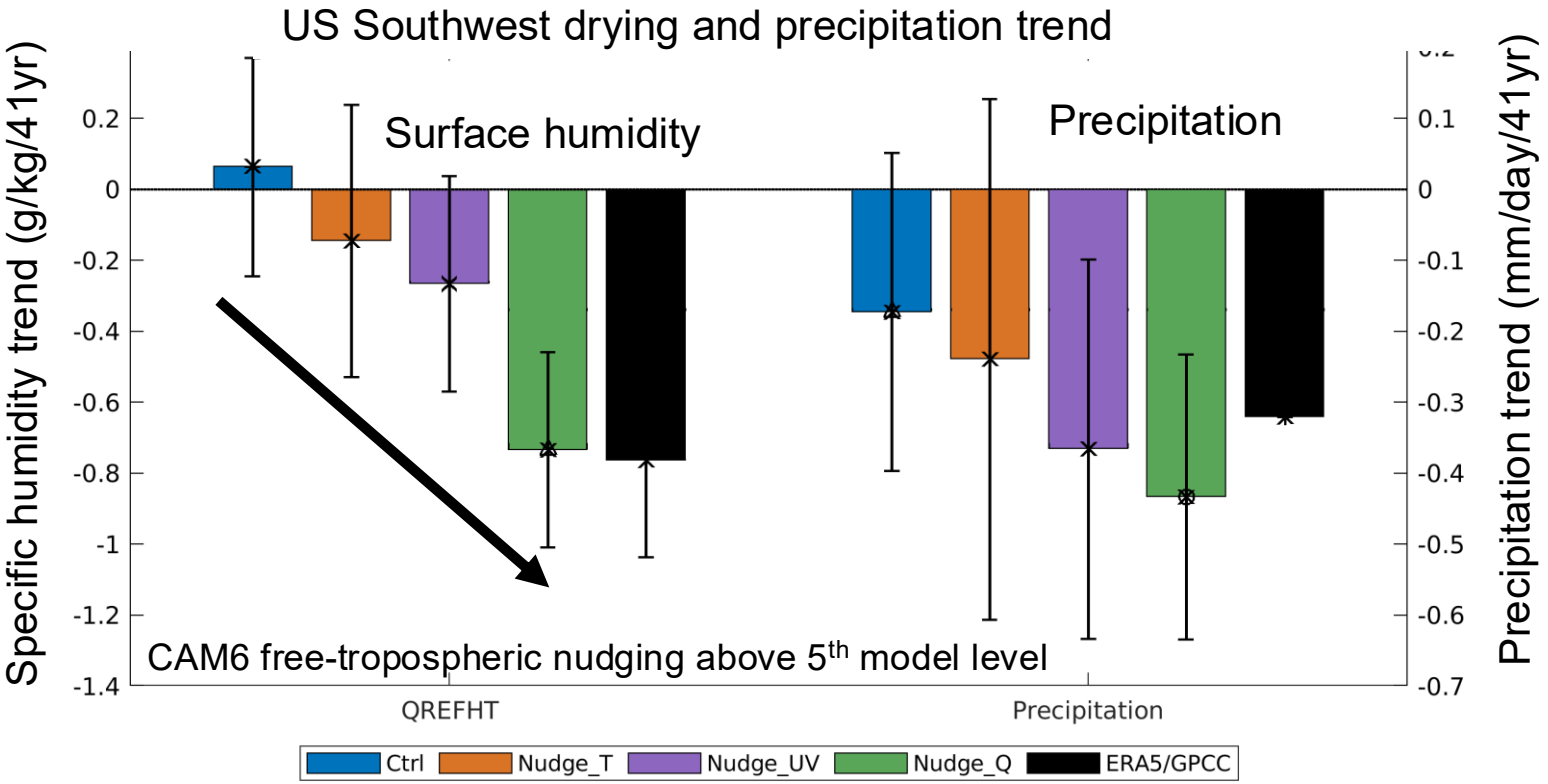


Baxter et al. (2026, in prep)

Nudging wind and humidity above the boundary layer significantly improves drying discrepancy





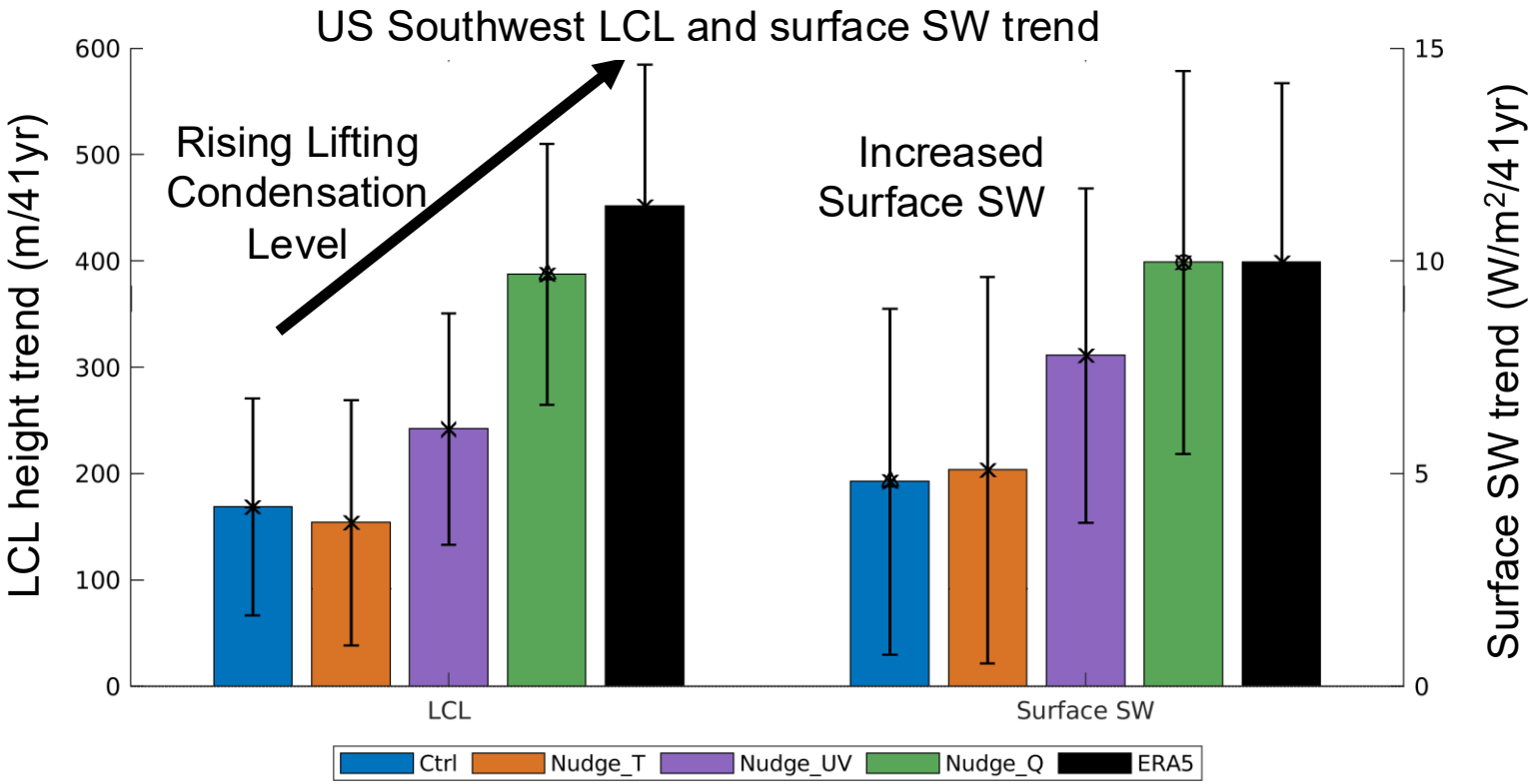


Baxter et al. (2026, in prep)

Correct boundary layer response involves rising LCL and increase surface shortwave

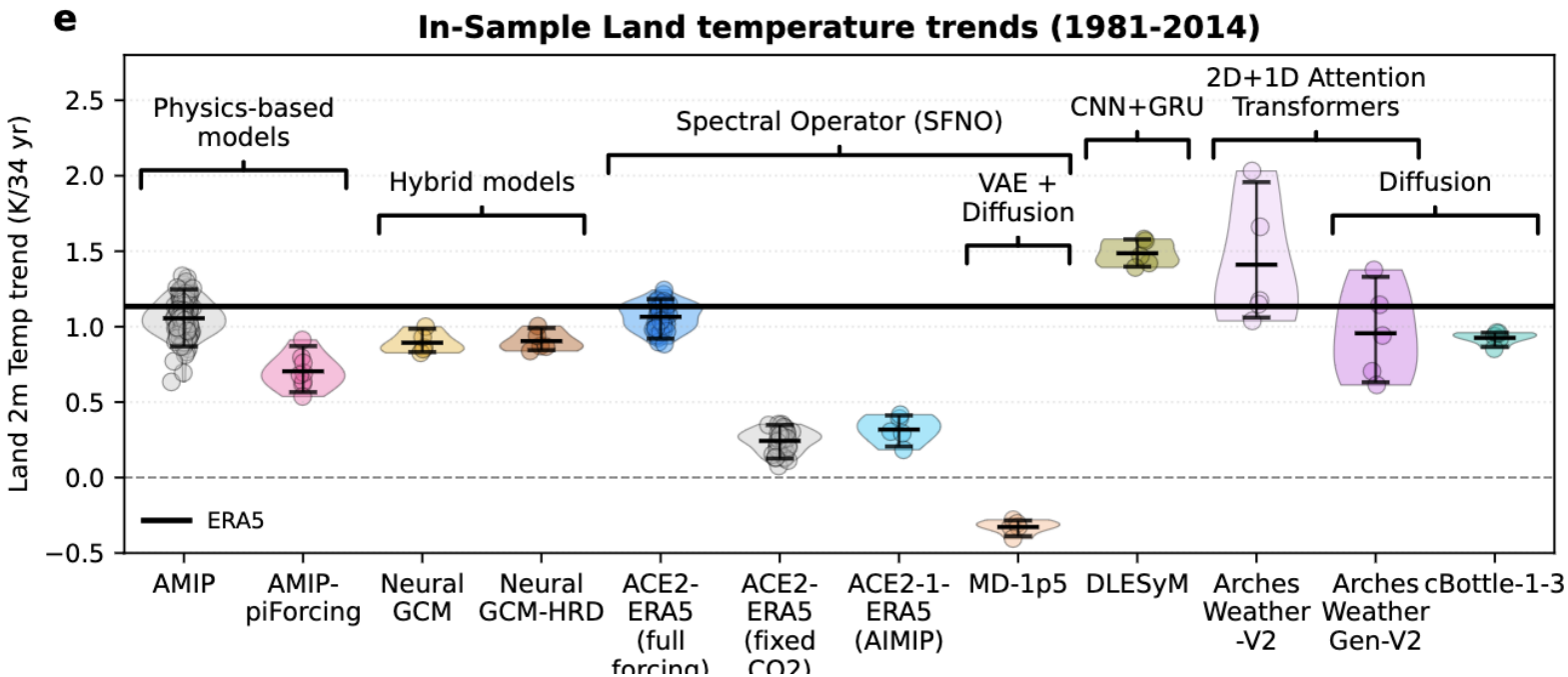






Baxter et al. (2026, in prep)

Large architecture-dependent spread reveals inconsistent physical learning

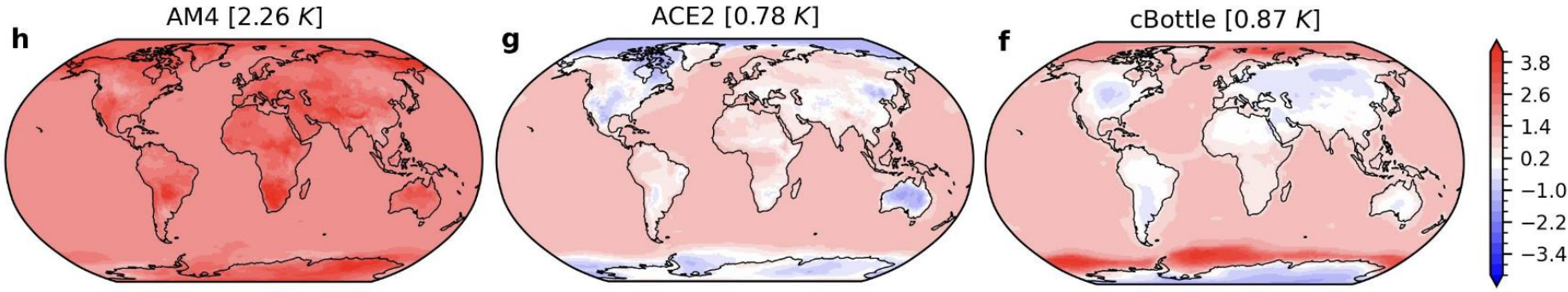


Malattribution leads to getting the wrong answer to counterfactual changes

Physics based model

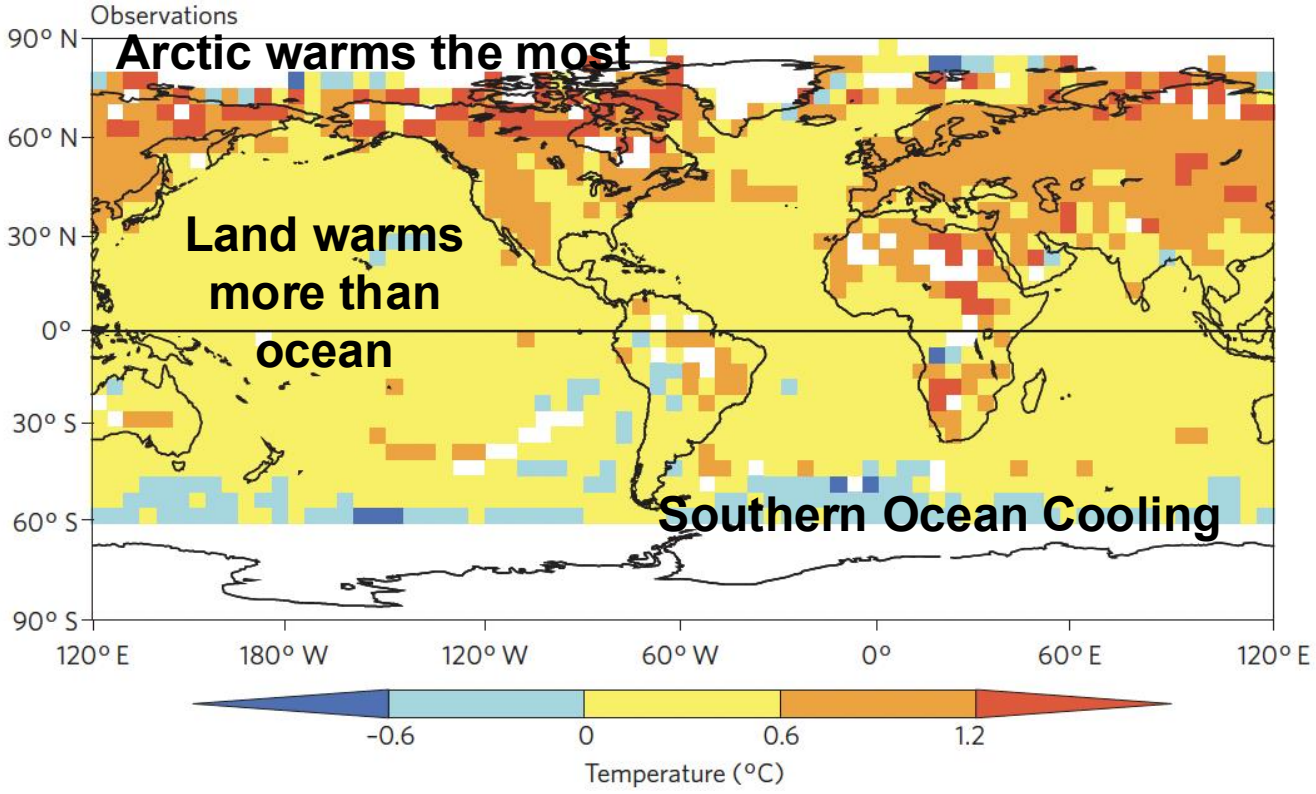
Spectral operator SFNO

Diffusion



AIMIP Henn et al. (2026, arXiv)
Zhang & Merlis (2026, JAMES)

Manabe got the right answer for the right reason



Stouffer & Manabe (2017, Nature Clim. Ch.)