Fine-tuning Earth System Emulators for Accurate Extreme Precipitation Projections

Bouabid et al.

September 8, 2026

Abstract

Earth System Model (ESM) simulations underpin climate adaptation and mitigation choices, but available projections are sparse due to their computational cost and display known biases, particularly for precipitation extremes. ESM emulators offer inexpensive surrogates that can rapidly explore alternative futures, but they inherit the distributional biases of their parent ESM. We fine-tune a generative ESM emulator on reanalysis data to address both limitations simultaneously. To validate this approach, we first fine-tune on a single historical member of a high-resolution ESM ensemble and evaluate against held-out future simulations. We find that fine-tuning corrects precipitation tail biases, and that this correction persists into out-of-sample future warming scenarios. It also recovers the forced distributional response and variability for temperature, relative humidity, and wind speed. This opens a new pathway to fast observation-constrained projections of future physical risk from impact-relevant climate variables.

1 Introduction

Planning adaptation and mitigation measures depends on projections of future climate simulated by Earth System Models (ESMs) (1, 2). ESMs have advanced our understanding of the climate system and of climate change associated with increasing greenhouse gas concentrations (3). However, ESM simulations also display systematic disagreement with trends and variability of certain variables from historical observations (4). This disagreement is especially pronounced in the tails of global precipitation, with large discrepancies and biases in ESMs, suggesting they may not reflect what future extreme precipitation events will look like (5, 6, 7, 8). The computational cost of ESM simulations compounds this problem, as it is prohibitive to simulate the large ensembles needed to resolve rare events. Climate risk is the convolution of hazard probabilities and damages cost, which is dominated by the severity of losses occasioned by extreme events. Mitigating biases in the tails is thus of paramount importance for risk assessment.

Simulations that accurately reproduce the historical climate can be expected more likely to offer reliable future projections. Under this assumption, a long line of work has used historical observations to post-process ESM precipitation projections, either by re-evaluating their individual importance based on agreement with historical observations (e.g. 9, 10, 11, 12, 13, 14, 15), or through bias correction, which realigns grid-cell distributions of simulated precipitation with observed distributions (e.g. 16, 17, 18, 19, 20, 21, 22, 23, 24, 25). These approaches, however, are restricted to correct events already simulated by ESMs, and do not provide a simple way to generate large ensembles of additional future climate realizations needed to study extremes. Climate emulation with machine learning (ML) offers a route for generating such ensembles: once trained, ML emulators are computationally cheap because they do not explicitly resolve climate dynamics.

Autoregressive emulators for example are trained directly on historical observations or reanalysis data to predict the evolution of the atmospheric state in six or twelve hours increments (e.g. 26, 27, 28, 29) and can reproduce observed precipitation statistics. However, their application to study future precipitation under warming has been limited due to lack of extrapolation guarantees in warmer climates (30, 31). A second class of ML emulators are instead trained on ESM simulations (e.g. 32, 33, 34), and can project precipitation beyond the historical warming range, but also inherits the emulated ESM's biases, including in precipitation tails. As such, each class captures only part of what future precipitation projections require: fidelity to observed variability for emulators trained on historical records, and extrapolation beyond the historical climate for emulators trained on ESM simulations. Combining observations with ESM simulations may offer a means to bridge this gap.

An immediate combination strategy is to pre-train an ML emulator on ESM simulations, and finalize the training with observational data. This strategy, known as *fine-tuning* (35), rests on the expectation that patterns acquired from prior training on a comprehensive and imperfect dataset can improve performance when complemented with training on limited, but more accurate, target data. Recent work has shown that ML models pre-trained on Coupled Model Intercomparison Project (CMIP) simulations and fine-tuned on observational or reanalysis products have skill in temperature forecasts at multi-seasonal and decadal timescales (36, 37). A conceptual analogy can be drawn with bias correction, where the pre-trained ESM emulator acts as a background product and refinement with observations acts as a corrective step that preserves learned relationships from ESM simulations, while improving agreement with the historical climate.

The difficulty in applying fine-tuning to ESM emulators lies in evaluating whether the correction learned over historical climates accurately carries over to warmer climates, where unprecedented extreme events have no observational analogues. For annual temperature, Immorlano et al. (37) address this validation problem by fine-tuning an ML model on historical temperature fields from a held-out CMIP model and evaluating it against that model's future projections. For extreme daily precipitation, however, discrepancies in CMIP models undermines their credibility as proxies to evaluate extrapolation under warming. More broadly, empirical evidence on the out-of-distribution skill of ML methods is mixed: focusing on tropical cyclones, Sun et al. (38) find that a ML emulator cannot generate storms whose severity exceeds that observed in the training data, consistent with the expectation that ML models are unreliable outside their training distribution. At the same time, ML models have been shown to transfer information about extreme events across regions (38, 39) and to leverage short-term variability to inform longer-term trends (40). Such empirical evidences indicate that fine-tuning with historical data might still improve the representation of distributional features of climate beyond the range of observed variability. Yet, this is hardly sufficient to build confidence that the learned relationships are consistent with physical principles governing the climate system. Settling this question requires a controlled experimental setting in which a credible groundtruth in a future climate is available.

Here, we consider a ESM emulator pre-trained on historical and future simulations from a CMIP6 model. We fine-tune it on a single historical member from a high-resolution CESM-HR ensemble. We fine-tune it on a single historical member from a high-resolution CESM-HR ensemble. We choose CESM-HR because it captures the spatial distribution and intensity of daily extreme precipitation more realistically than standard ESMs, in line with present climate observations (41), but also simulates them in future climates (42). We ask whether the corrections learned through historical fine-tuning carry over accurately to future CESM-HR simulations under 8.5 Wm^{-2} forcing, which lie well outside historical levels. Our results show that fine-tuning corrects the distributional biases from the pre-training CMIP6 model. Precipitation tail biases in particular are well corrected over the historical period, and this corrections extrapolate accurately to out-of-distribution future warming, even for unprecedented events.

2 Results

2.1 ESM tail precipitation are biased and inherited by emulators

We pre-train the emulator on daily fields from a coarse-resolution ESM (MPI-ESM1-2-LR (43), 2°) using a score-based generative ML emulation approach (34). The emulator generates joint daily fields of four surface variables: temperature, precipitation, relative humidity, and wind speed. Fine-tuning on the high-resolution ESM (CESM-HR (42), 0.25°) is restricted to 1950–2015, with 2021–2025 held out for hyperparameter calibration (see Methods). To verify whether CESM-HR provides a credible basis for model evaluation, we compare its global land precipitation distribution to reanalysis and observational products regridded to a common resolution. Products show broad mutual agreement on the magnitude of extreme precipitation, while the coarse-resolution ESM displays a pronounced dry bias in the upper tail; inherited by the pre-trained emulator. A higher resolution model such as CESM-HR does not have this bias (Fig. 1b). To evaluate whether fine-tuning on historical climate helps mitigate this bias even under future warming, we fine-tune the emulator on a single CESM-HR historical ensemble member and evaluate the emulator against the CESM-HR 10-member ensemble under the RCP8.5 forcing scenario for the 2080-2100 period. During this period, CESM-HR exceeds its historical GMST levels by more than 3°C, and annual maximum precipitation (Rx1day) exceed historical levels by 50%, placing the evaluation well outside the conditions seen during fine-tuning on the historical period (Fig. 1c).

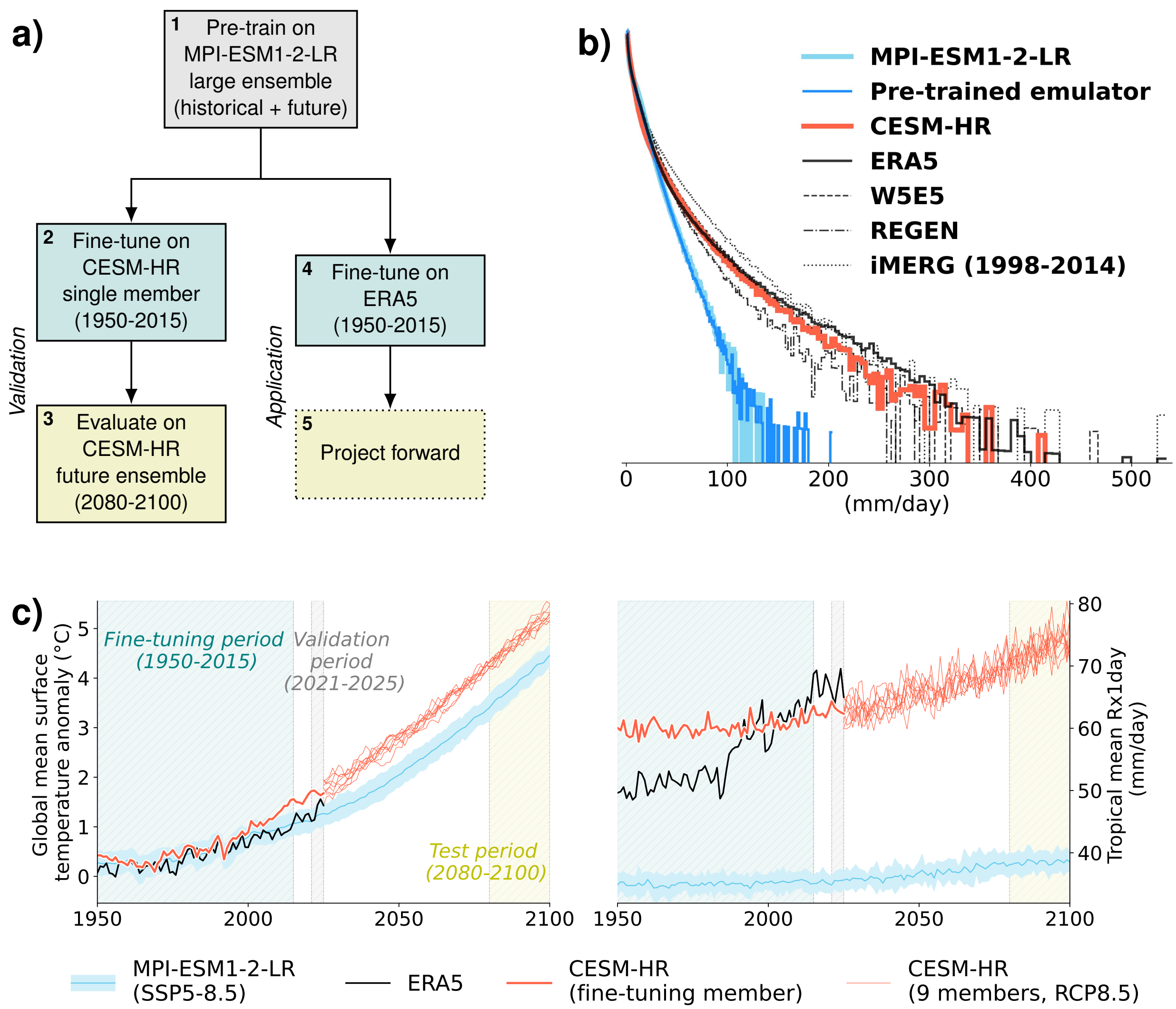

Figure 1: Experimental setup and precipitation tail biases in ESM outputs. (a) Schematic of the experimental design: a score-based generative emulator is pre-trained on coarse resolution (MPI-ESM1-2-LR) historical and future simulations, then fine-tuned on single high-resolution (CESM-HR) historical ensemble member for evaluation, or on reanalysis data (ERA5) for projections. (b) Distribution of daily precipitation over land grid cells during 1979–2014 on a common 2° grid for the coarse-resolution model and its pre-trained emulator, the high-resolution model, reanalysis, bias-corrected reanalysis (W5E5) and two observational products (iMERG, REGEN); high-resolution simulations, observations, and reanalysis products cluster together whereas the coarse-resolution model and its emulator underestimate precipitation tail frequencies. (c) GMST and tropical mean Rx1day trajectories for the coarse-resolution model, reanalysis, high-resolution training member and full high-resolution future ensemble; shaded regions indicate the fine-tuning period (1950–2015), validation period (2021-2025) and out-of-distribution evaluation period (2080–2100).

2.2 Fine-tuning corrects precipitation tails under warming

After fine-tuning, the emulator shows alignment with the CESM-HR wet-day precipitation distribution over land in 2080-2100, including in the upper tail (Fig. 2a). Over the same period, MPI-ESM1-2-LR and the pre-trained emulator are significantly drier. CESM-HR simulates a departure from historical precipitation levels under warming (see SI-X), which in particular is visible for the historical CESM-HR member used for fine-tuning (dashed histogram in Fig. 2a). The fine-tuned emulator reproduces this departure, providing a correction that extends beyond the range of historical precipitation. It suggests that a distributional correction is learned from CESM-HR, while preserving a forced response learned from pre-training on MPI-ESM1-2-LR; under warming, the emulator does not reproduce the historical CESM-HR member used for fine-tuning, but projects a shift in the precipitation distribution. This out-of-sample behavior is sensitive to fine-tuning hyperparameters (see Methods). The emulator also produces extreme events that exceed the simulated precipitation range, at both pre-training and fine-tuning stages. Similar behavior has been reported with generative models, with samples generated outside the support of the training distribution (**?**). We therefore expect an overestimation of record-breaking floods frequency for the emulator fine-tuned on ERA5.

The spatial organization of extremes precipitation for the fine-tuned emulator also shows alignment with CESM-HR (Fig. 2b). Relative to the MPI-ESM1-2-LR pre-training data, the emulator displays stronger extremes in the tropics, a sharper change in extremes intensity across the Himalayas, and drier regions in the South Atlantic. Emulator samples and power spectra analysis (see SI-X) tell the same story: fine-tuning modifies the emulator precipitation spatial structure to align it with its fine-tuning target. Figure 2b also shows statistical artifacts in the emulated Rx1day, with a grid cell noise that is absent from the CESM-HR Rx1day; the pre-trained emulator also displays this, suggesting an artifact from the generative model (see SI-X). The emulated precipitation extremes at at an individual grid cell may therefore be subject to a random under- or overestimation bias. Regional aggregation however helps mitigate this bias. When the emulated precipitation is pooled over progressively larger regions in Nigeria, the emulator distribution aligns closely with CESM-HR (Fig. 2c). Therefore, a regional assessment of extreme precipitation can be envisaged with the fine-tuned emulator.

2.3 The correction extends to near-surface variables

The emulator is pre-trained to sample precipitation jointly with near-surface temperature, relative humidity, and wind speed. Fine-tuning is also performed for all four variables jointly. Figure 3 shows it stably modifies the 2080-2100 global land distributions of these near-surface variables (red), shifting them away from the distribution of the MPI-ESM1-2-LR pre-training simulations (blue), so they align better with CESM-HR (black).

For temperature, CESM-HR simulates in 2080-2100 a clear increase in land temperatures relative to its historical levels (compare Fig. 3a,d dashed and black). While fine-tuned with historical simulations, the emulator reproduces a similar increase, drawing from the forced response learned by pre-training. Similarly for relative humidity, the emulator reproduces the shift of probability mass toward lower relative humidities on land that CESM-HR simulates with warming (Fig. 3b compare dashed, black, red). However, the emulated future land relative humidity distribution conserves the shape learned from CESM-HR historical simulations, and does not project the apparition of a second mode in CESM-HR future simulations. The fine-tuning also corrects the upper tail of relative humidity distribution (Fig. 3e), although it exceeds values consistent with observations, a known feature of ESMs (44). For wind speed on land, CESM-HR simulates little change in the dis-

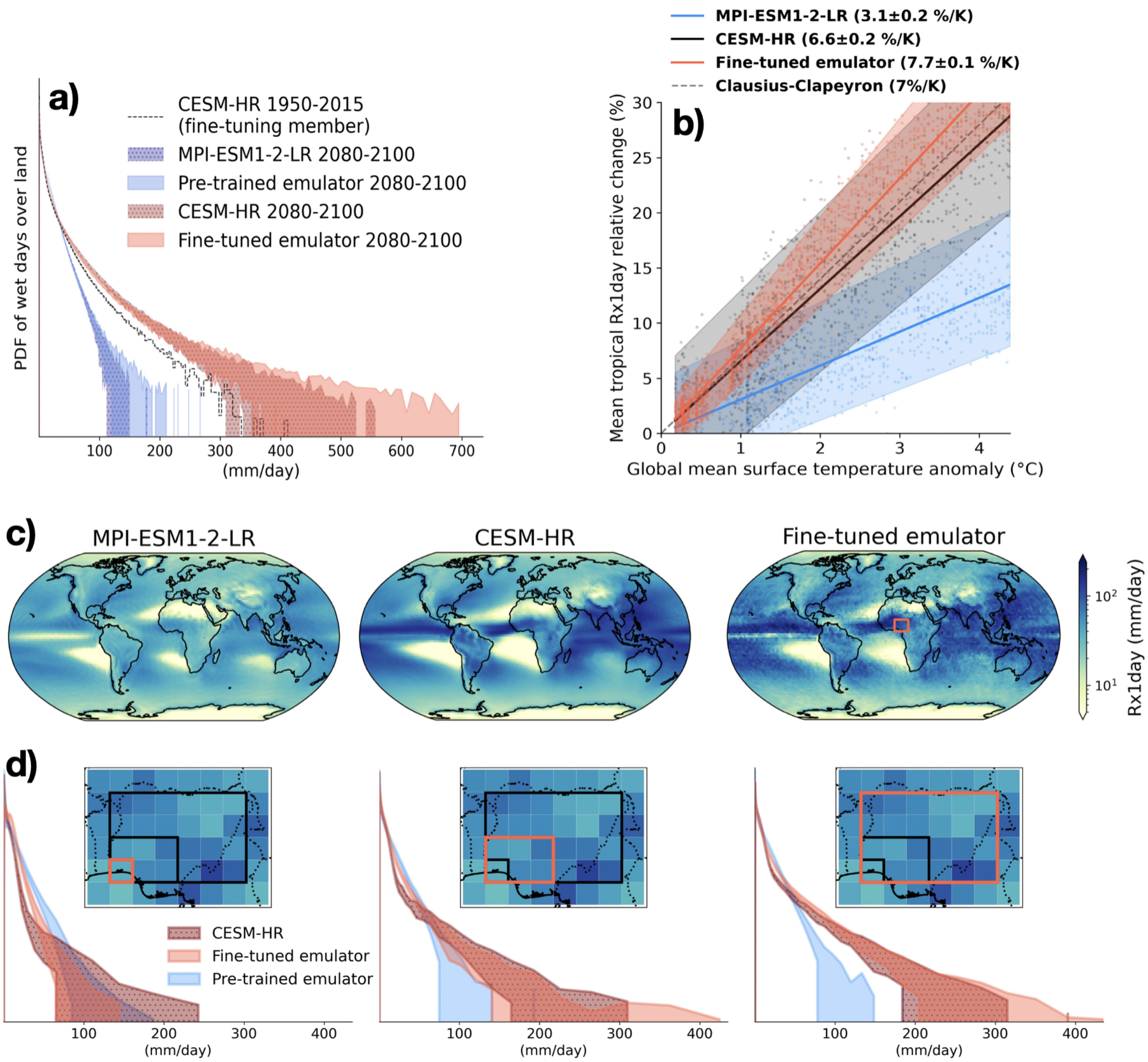

Figure 2: Fine-tuning corrects precipitation tail biases out of sample under future warming. (a) Distribution of wet days over land during 2080-2100 for the MPI-ESM1-2-LR, CESM-HR, pre-trained emulator and fine-tuned emulator, alongside the distribution of the CESM-HR historical member used for fine-tuning; the fine-tuned emulator aligns with CESM-HR 2080-2100 despite being fine-tuned only on the historical period; shaded areas represent 95% confidence range. (b) Maximum annual precipitation by grid-cell averaged over 2080–2100 for MPI-ESM1-2-LR, CESM-HR and fine-tuned emulator; the fine-tuned emulator shows agreement with CESM-HR in the spatial pattern of extreme precipitation but is noisy at grid-cell level. (c) Distribution of wet days over an increasing large region in Nigeria; the fine-tuned emulator produces distributions that align better with CESM-HR when pooled over larger regions.

tribution with warming, and fine-tuning the emulator produces a distribution that aligns well with CESM-HR (Fig. 3b,f), only with a small positive bias. Note that biases may reflect our fine-tuning hyperparameter selection strategy, which favors alignment in precipitation tails (see Methods), and can be modified to prioritize other variables.

The emulator preserves correlations across variables and thus can be used to project how extreme precipitation covary with other variables after fine-tuning. This enables sampling from future compound events, such as hot-wet or extreme precipitation-wind events (45), with distributions calibrated against observational data. In addition, fine-tuning also corrects the spatial characteristics of the generated near-surface variables (see comparative samples in SI-X). Figure 3g-i show that for spatial features larger than synoptic scales (e.g. heatdome, moonsoon) the power spectra of MPI-ESM1-2-LR and CESM-HR are already comparable and well reproduced by the emulator. However, because CESM-HR runs at horizontal resolution ten times finer than MPI-ESM1-2-LR, it has greater energy at scales close to the numerical meshgrid. After fine-tuning, the emulator aligns with the CESM-HR power spectra for these fine scales. Therefore, the emulator produces spatially coherent future fields, with spatial characteristics close to the fine-tuning target.

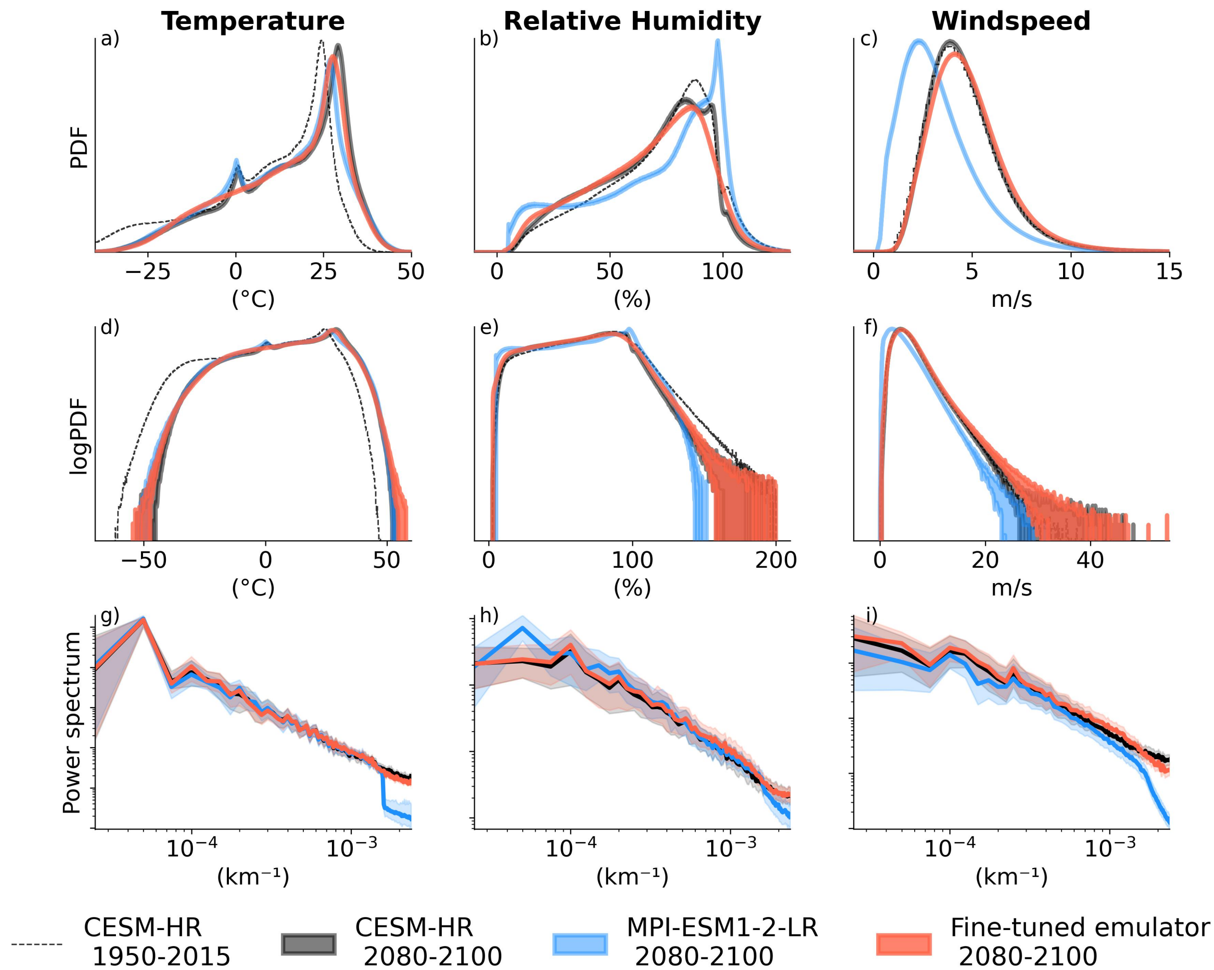

Figure 3: Fine-tuning corrects distributional biases and spatial characteristics of near-surface climate variables. (a–c) Distribution of daily temperature, relative humidity, and wind speed over 2080–2100 for MPI-ESM1-2-LR, CESM-HR, the fine-tuned emulator, alongside the distribution of the CESM-HR historical member used for fine-tuning; shaded areas represent residuals 95% confidence range. (d–f) Same as (a–c) with logarithmic y-axis to emphasize tail behavior; (g–i) Global zonally averaged spherical power spectra computed using days from the 2080-2100 period for each variable; full lines indicate the mean, shaded areas span 95% range across different days.

2.4 Pre-training constrains the spatial pattern of change

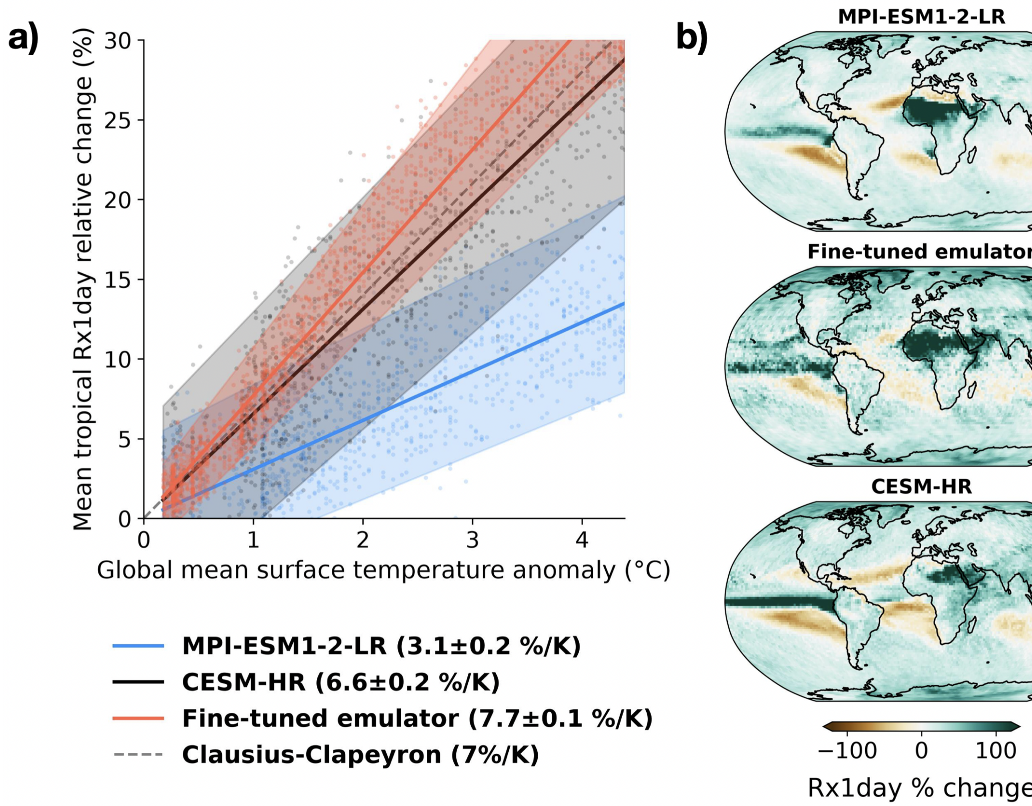

Figure 4: Fine-tuning corrects extreme tropical precipitation trend but conserves spatial pattern of change from MPI-ESM1-2-LR. (a) Relative change in Rx1day averaged over 30°S-30°N function of global mean surface temperature for MPI-ESM1-2-LR, CESM-HR and fine-tuned emulator; the fine-tune emulator gets closer to CESM-HR despite fine-tuning on the historical period only; line indicate least-square regression, shading 95% confidence range for values and dots values. (b) Relative change in Rx1day per grid-cell between 2080-2100 and 1950-2000 periods; the fine-tuned emulator projects a spatial pattern of change closer to the MPI-ESM1-2-LR simulations it was pre-trained on.

Fine-tuning helps correct the scaling of extreme tropical precipitation with additional warming. MPI-ESM1-2-LR shows a weak increase in Rx1day with GMST, of about 3% K^{-1}, where CESM-HR shows a trend closer to the Clausius–Clapeyron rate of $\sim$7% K^{-1} (Fig. 4a). After fine-tuning on one historical CESM-HR member, the emulator shows a Rx1day trend of 7.7% K^{-1} which exceeds CESM-HR simulations but remains close to the CESM-HR ensemble spread, and markedly departs from the MPI-ESM1-2-LR pretraining simulation data. It also shows an underdispersion of the emulated Rx1day at a given warming level relative to CESM-HR, which may be due to fine-tuning on a single member making the emulator underestimate variability in extremes.

Inspecting the spatial pattern of Rx1day change however shows that it remains largely inherited from the MPI-ESM1-2-LR simulations (Fig. 4b). The historical period used for fine-tuning is insufficient to inform the emulator on the spatial pattern of future change in Rx1day with warming. Therefore fine-tuning on a CESM-HR historical member adjusts the magnitude and spatial features of absolute extreme precipitation, the global rate of change with warming of extreme precipitation, but the spatial pattern of change remains inherited from pre-training. The choice of pre-training ESM therefore assumes a spatial pattern for future extreme precipitation change in projections.

2.5 Future extreme precipitation fine-tuned on reanalysis

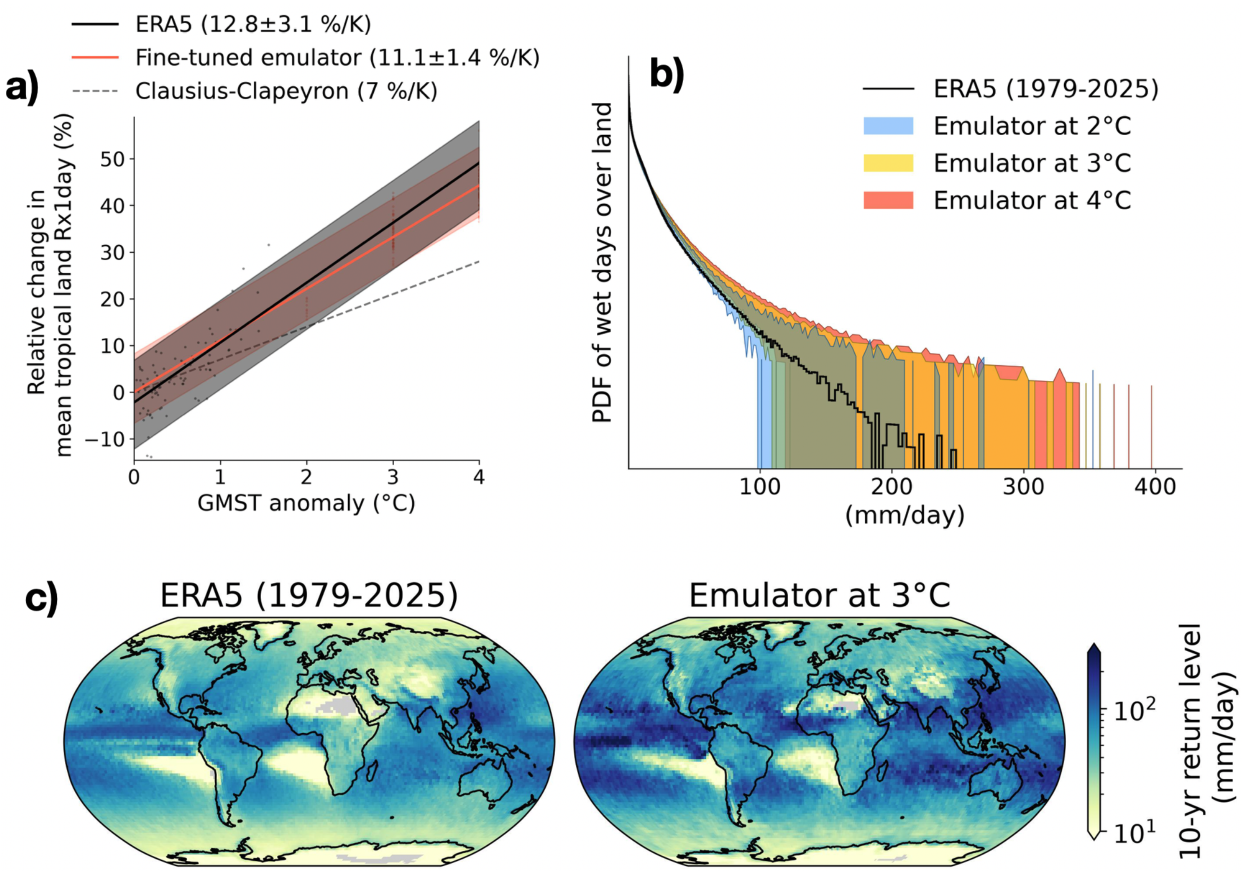

Figure 5: Projection of extreme precipitation under increasing global warming levels. (a) Trends in relative change of Rx1day averaged over 30°S-30°N function of global mean surface temperature, fitted for ERA5 and the fine-tuned emulator; shaded areas represent residuals 95% confidence range. (b) Distribution of wet days over land in ERA5 and for the fine-tuned emulator at 2, 3 and 4°C above pre-industrial levels; shaded areas represent 95% confidence range. (c) Grid-cell-level 20-year precipitation return levels estimated from ERA5 over 1979–2025 and from the emulator at 3°C of warming. Gray areas indicate areas with insufficient data to estimate return levels.

Following the evaluation against CESM-HR, we fine-tune the emulator on ERA5 reanalysis data. ERA5 tends to underrepresent the upper tail of local precipitation relative to station observations, particularly in the tropics (46). After spatial coarsening to a horizontal resolution of 2°, however, its precipitation tails over global land align well with satellite and gauge-based observations (Fig. 1b), making ERA5 a credible proxy for observed extreme precipitation at this resolution. We fine-tune the emulator on ERA5 data from 1950 to 2015, reserving 2021-2025 for hyperparameter tuning (see Methods), and use the resulting emulator to project global precipitation at warming levels of 2°C, 3°C, and 4°C above pre-industrial levels.

Historical precipitation records are too noisy to constrain trends with high confidence. Nevertheless, the tropical Rx1day sensitivity produced by the emulator aligns with the central estimate obtained from ERA5 (Fig. 5a). The emulator's sensitivity exceeds global observational estimates of approximately 6-10% $\cdot \mathrm{K}^{-1}$ (47, 48). However, it is consistent with evidence of higher sensitivity in the tropics; for example, O'Gorman (12) estimates a 6-14% $\cdot \mathrm{K}^{-1}$ sensitivity for the 99.9^{th} percentile of daily tropical precipitation.

This sensitivity is reflected in a progressive shift of the global land precipitation distribution towards heavier extremes with warming (Fig. 5b). Comparing 20-year return levels estimated from ERA5 with those projected by the emulator at 3°C shows increases in precipitation extremes across most land regions, although their magnitude varies considerably (Fig. 5c). The increases are particularly pronounced in Western Africa and South Asia, where estimated 20-year return levels rise from approximately 100 mm/day to around 400-500 mm/day. As shown in Section 2.1, however, the emulator can generate events more extreme than those occurring in the corresponding future climate simulation. The magnitude and frequency of the most exceptional projections (e.g. >500 mm/day), should therefore be interpreted with caution.

Greater confidence can be placed in the projected increase in the frequency of events between 100 and 400 mm/day, a range over which the emulator demonstrated skill in the CESM-HR evaluation. In ERA5 over 1979-2025, this range encompasses the estimated 20-year precipitation event across most regions of the world. We therefore use the emulator to examine how the frequency of these present-day 20-year events changes with warming (see Methods). Because small-scale artefacts limit the reliability of grid-cell-level projections, we conduct this analysis at the regional scale using AR6 regions (49).

Figure 6 shows that over global land, the annual frequency of a present-day 20-year event approximately doubles at 2°C of warming and triples at 4°C. These increases are consistent with CMIP6 projected ranges: the frequency of present-day 10-year events increases by factors of approximately 1.5-2.0 and 2.3-3.7 at 2°C and 4°C, respectively, while that of 50-year events increases by factors of 2.0-2.6 and 3.3-6 (50). Our uncertainty ranges are considerably narrower because the emulator is pre-trained on a single CMIP6 model and therefore does not represent inter-model variability. The ranges capture only finite-ensemble statistical estimation uncertainty arising from internal variability, which tends towards zero as the size of the emulated ensemble grows. The changes in annual frequency vary considerably among regions. In South-West South America, the frequency remains nearly unchanged, whereas in Western Africa it increases by a factor of seven at 4°C of warming. A small number of regions (e.g. West North-America) show a slight decline between 3°C and 4°C, but the frequency increases monotonically with warming across most regions.

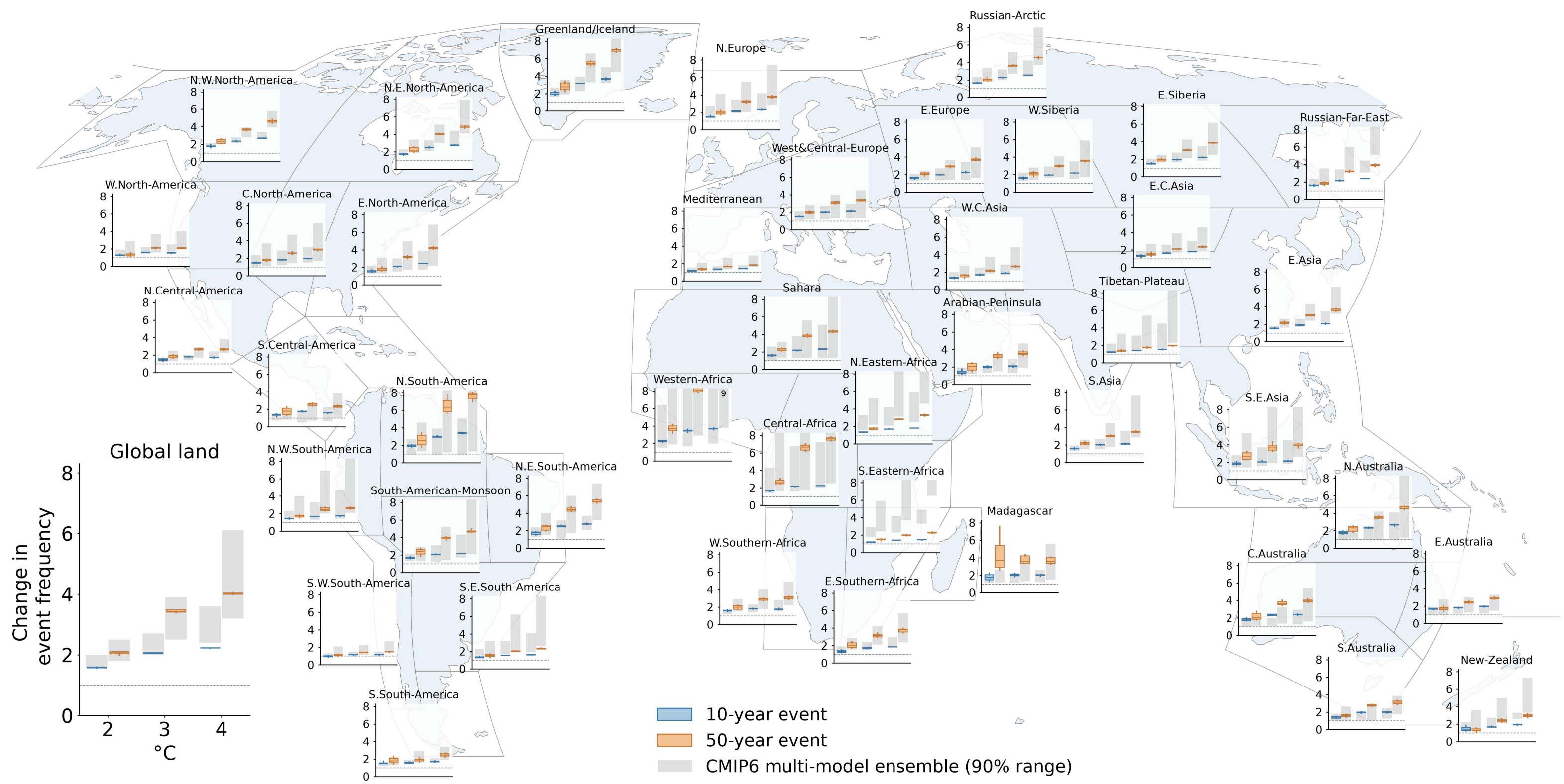

Figure 6: Projected changes in the frequency of a present-day 20-year return event under 2°C, 3°C, and 4°C of global warming levels relative to the pre-industrial levels. Present-day is taken as ERA5 data between 1979 and 2025. Results are shown for the global land area and the AR6 regions. For each box plot, the horizontal line and the box represent the median and central 66% uncertainty range, respectively, of the frequency changes across different realizations from the emulator, and the whiskers extend to the 90% uncertainty range. Adapted from Seneviratne et al. (50).

References

[1] Gregory M Flato. Earth system models: an overview. *Wiley Interdisciplinary Reviews: Climate Change*, 2(6):783–800, 2011. 1

[2] CD Jones. So what is in an earth system model? *Journal of Advances in Modeling Earth Systems*, 12(2):e2019MS001967, 2020. 1

[3] Valérie Masson-Delmotte, Panmao Zhai, Anna Pirani, Sarah L Connors, Clotilde Péan, Sophie Berger, Nada Caud, Yang Chen, Leah Goldfarb, Melissa I Gomis, et al. Climate change 2021: the physical science basis. *Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change*, 2(1):2391, 2021. 1

[4] Tiffany A Shaw, Paola A Arias, Mat Collins, Dim Coumou, Arona Diedhiou, Chaim I Garfinkel, Shipra Jain, Mathew Koll Roxy, Marlene Kretschmer, L Ruby Leung, et al. Regional climate change: consensus, discrepancies, and ways forward. *Frontiers in climate*, 6:1391634, 2024. 1

[5] Kevin E Trenberth. Changes in precipitation with climate change. *Climate research*, 47(1/2): 123–138, 2011. 1

[6] Michael Wehner, Peter Gleckler, and Jiwoo Lee. Characterization of long period return values of extreme daily temperature and precipitation in the cmip6 models: Part 1, model evaluation. *Weather and Climate Extremes*, 30:100283, 2020. 1

[7] Hebatallah Mohamed Abdelmoaty, Simon Michael Papalexiou, Chandra Rupa Rajulapati, and Amir AghaKouchak. Biases beyond the mean in cmip6 extreme precipitation: A global investigation. *Earth's Future*, 9(10):e2021EF002196, 2021. 1

[8] Amal John, Hervé Douville, Aurélien Ribes, and Pascal Yiou. Quantifying cmip6 model uncertainties in extreme precipitation projections. *Weather and Climate Extremes*, 36:100435, 2022. 1

[9] Claudia Tebaldi and Reto Knutti. The use of the multi-model ensemble in probabilistic climate projections. *Philosophical transactions of the royal society A: mathematical, physical and engineering sciences*, 365(1857):2053–2075, 2007. 1

[10] Shaw Chen Liu, Congbin Fu, Chein-Jung Shiu, Jen-Ping Chen, and Futing Wu. Temperature dependence of global precipitation extremes. *Geophysical Research Letters*, 36(17), 2009. 1

[11] Chein-Jung Shiu, Shaw Chen Liu, Congbin Fu, Aiguo Dai, and Ying Sun. How much do precipitation extremes change in a warming climate? *Geophysical Research Letters*, 39(17), 2012. 1

[12] Paul A O'Gorman. Sensitivity of tropical precipitation extremes to climate change. *Nature Geoscience*, 5(10):697–700, 2012. 1, 11

[13] Chad W Thackeray, Alex Hall, Jesse Norris, and Di Chen. Constraining the increased frequency of global precipitation extremes under warming. *Nature Climate Change*, 12(5):441–448, 2022. 1

[14] Wenxia Zhang, Kalli Furtado, Tianjun Zhou, Peili Wu, and Xiaolong Chen. Constraining extreme precipitation projections using past precipitation variability. *Nature Communications*, 13(1):6319, 2022. 1

[15] Maximilian Kotz, Stefan Lange, Leonie Wenz, and Anders Levermann. Constraining the pattern and magnitude of projected extreme precipitation change in a multimodel ensemble. *Journal of Climate*, 37(1):97–111, 2024. 1

[16] Alex J Cannon, Stephen R Sobie, and Trevor Q Murdock. Bias correction of gcm precipitation by quantile mapping: how well do methods preserve changes in quantiles and extremes? *Journal of climate*, 28(17):6938–6959, 2015. 1

[17] Douglas Maraun. Bias correcting climate change simulations-a critical review. *Current Climate Change Reports*, 2(4):211–220, 2016. 1

[18] Maike Holthuijzen, Brian Beckage, Patrick J Clemins, Dave Higdon, and Jonathan M Winter. Robust bias-correction of precipitation extremes using a novel hybrid empirical quantile-mapping method: Advantages of a linear correction for extremes. *Theoretical and Applied Climatology*, 149(1):863–882, 2022. 1

[19] Kyuhyun Byun and Alan F Hamlet. An improved empirical quantile mapping approach for bias correction of extreme values in climate model simulations. *Environmental Research Letters*, 20(1):014041, 2025. 1

[20] Bastien François, Soulivanh Thao, and Mathieu Vrac. Adjusting spatial dependence of climate model outputs with cycle-consistent adversarial networks. *Climate dynamics*, 57(11):3323–3353, 2021. 1

[21] Fang Wang and Di Tian. On deep learning-based bias correction and downscaling of multiple climate models simulations. *Climate dynamics*, 59(11):3451–3468, 2022. 1

[22] Philipp Hess, Stefan Lange, Christof Schötz, and Niklas Boers. Deep learning for bias-correcting cmip6-class earth system models. *Earth's Future*, 11(10):e2023EF004002, 2023. 1

[23] Zhong Yi Wan, Ignacio Lopez-Gomez, Robert Carver, Tapio Schneider, John Anderson, Fei Sha, and Leonardo Zepeda-Núñez. Regional climate risk assessment from climate models using probabilistic machine learning. *arXiv preprint arXiv:2412.08079*, 2024. 1

[24] Philipp Hess, Michael Aich, Baoxiang Pan, and Niklas Boers. Fast, scale-adaptive and uncertainty-aware downscaling of earth system model fields with generative machine learning. *Nature Machine Intelligence*, 7(3):363–373, 2025. 1

[25] Michael Aich, Philipp Hess, Baoxiang Pan, Sebastian Bathiany, Yu Huang, and Niklas Boers. Conditional diffusion models for downscaling and bias correction of earth system model precipitation. *Geoscientific Model Development*, 19(4):1791–1808, 2026. 1

[26] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022. 2

[27] Oliver Watt-Meyer, Brian Henn, Jeremy McGibbon, Spencer K Clark, Anna Kwa, W Andre Perkins, Elynn Wu, Lucas Harris, and Christopher S Bretherton. Ace2: accurately learning subseasonal to decadal atmospheric variability and forced responses. *npj Climate and Atmospheric Science*, 8(1):205, 2025. 2

[28] Haiwen Guan, Troy Arcomano, Ashesh Chattopadhyay, and Romit Maulik. Lucie: A lightweight uncoupled climate emulator with long-term stability and physical consistency. *Journal of Advances in Modeling Earth Systems*, 17(11):e2025MS005152, 2025. 2

[29] Janni Yuval, Ian Langmore, Dmitrii Kochkov, and Stephan Hoyer. Neural general circulation models for modeling precipitation. *Science Advances*, 12(2):eadv6891, 2026. 2

[30] Jacob B Landsberg and Elizabeth A Barnes. Forecasting the future with yesterday's climate: Temperature bias in ai weather and climate models. *Geophysical Research Letters*, 53(6): e2025GL119740, 2026. 2

[31] Bosong Zhang and Timothy M Merlis. The equilibrium response of atmospheric machine-learning models to uniform sea surface temperature warming. *Journal of Advances in Modeling Earth Systems*, 18(9):e2026MS005815, 2026. 2

[32] Gosha Geogdzhayev, Andre N Souza, Glenn R Flierl, and Raffaele Ferrari. An eof-based emulator of means and covariances of monthly climate fields. *EGUsphere*, 2025:1–37, 2025. 2

[33] Noah D Brenowitz, Tao Ge, Akshay Subramaniam, Peter Manshausen, Aayush Gupta, David M Hall, Morteza Mardani, Arash Vahdat, Karthik Kashinath, and Michael S Pritchard. Climate in a bottle: Towards a generative foundation model for the kilometer-scale global atmosphere. *arXiv preprint arXiv:2505.06474*, 2025. 2

[34] Shahine Bouabid, Andre Nogueira Souza, and Raffaele Ferrari. Score-based generative emulation of impact-relevant earth system model outputs. *Journal of Advances in Modeling Earth Systems*, 18(3):e2025MS005558, 2026. 2, 3

[35] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014. 2

[36] Yoo-Geun Ham, Jeong-Hwan Kim, and Jing-Jia Luo. Deep learning for multi-year enso forecasts. *Nature*, 573(7775):568–572, 2019. 2

[37] Francesco Immorlano, Veronika Eyring, Thomas le Monnier de Gouville, Gabriele Accarino, Donatello Elia, Stephan Mandt, Giovanni Aloisio, and Pierre Gentine. Transferring climate change physical knowledge. *Proceedings of the National Academy of Sciences*, 122(15): e2413503122, 2025. 2

[38] Y Qiang Sun, Pedram Hassanzadeh, Mohsen Zand, Ashesh Chattopadhyay, Jonathan Weare, and Dorian S Abbot. Can ai weather models predict out-of-distribution gray swan tropical cyclones? *Proceedings of the National Academy of Sciences*, 122(21):e2420914122, 2025. 2

[39] Y Qiang Sun, Pedram Hassanzadeh, Tiffany Shaw, and Hamid A Pahlavan. Predicting beyond training data via extrapolation versus translocation: Ai weather models and dubai's unprecedented 2024 rainfall. *arXiv preprint arXiv:2505.10241*, 2025. 2

[40] Laura A Mansfield, Peer J Nowack, Matt Kasoar, Richard G Everitt, William J Collins, and Apostolos Voulgarakis. Predicting global patterns of long-term climate change from short-term simulations using machine learning. *npj Climate and Atmospheric Science*, 3(1):44, 2020. 2

[41] Ping Chang, Dan Fu, Xue Liu, Frederic S Castruccio, Andreas F Prein, Gokhan Danabasoglu, Xiaoqi Wang, Julio Bacmeister, Qiuying Zhang, Nan Rosenbloom, et al. Future extreme precipitation amplified by intensified mesoscale moisture convergence. *Nature Geoscience*, 19 (1):33–41, 2026. 2

[42] Ping Chang, Shaoqing Zhang, Gokhan Danabasoglu, Stephen G Yeager, Haohuan Fu, Hong Wang, Frederic S Castruccio, Yuhu Chen, James Edwards, Dan Fu, et al. An unprecedented set of high-resolution earth system simulations for understanding multiscale interactions in climate variability and change. *Journal of Advances in Modeling Earth Systems*, 12(12): e2020MS002298, 2020. 2, 3

[43] Dirk Olonscheck, Laura Suarez-Gutierrez, Sebastian Milinski, Goratz Beobide-Arsuaga, Johanna Baehr, Friederike Fröb, Tatiana Ilyina, Christopher Kadow, Daniel Krieger, Hongmei Li, et al. The new max planck institute grand ensemble with cmip6 forcing and high-frequency model output. *Journal of Advances in Modeling Earth Systems*, 15(10):e2023MS003790, 2023. 3

[44] Kimmo Ruosteenoja, Kirsti Jylhä, Jouni Räisänen, and Antti Mäkelä. Surface air relative humidities spuriously exceeding 100% in CMIP5 model output and their impact on future projections. *Journal of Geophysical Research: Atmospheres*, 122(18):9557–9568, 2017. 5

[45] Poulomi Ganguli and Bruno Merz. Increasing probability of extreme rainfall preconditioned by humid heatwaves in global coastal megacities. *npj Climate and Atmospheric Science*, 8(1): 144, 2025. 7

[46] David A Lavers, Adrian Simmons, Freja Vamborg, and Mark J Rodwell. An evaluation of era5 precipitation for climate monitoring. *Quarterly Journal of the Royal Meteorological Society*, 148(748):3152–3165, 2022. 10

[47] Qiaohong Sun, Xuebin Zhang, Francis Zwiers, Seth Westra, and Lisa V Alexander. A global, continental, and regional analysis of changes in extreme precipitation. *Journal of Climate*, 34 (1):243–258, 2021. 11

[48] Seth Westra, Lisa V Alexander, and Francis W Zwiers. Global increasing trends in annual maximum daily precipitation. *Journal of climate*, 26(11):3904–3918, 2013. 11

[49] Maialen Iturbide, José M Gutiérrez, Lincoln M Alves, Joaquín Bedia, Ruth Cerezo-Mota, Ezequiel Cimadevilla, Antonio S Cofiño, Alejandro Di Luca, Sergio Henrique Faria, Irina V Gorodetskaya, et al. An update of IPCC climate reference regions for subcontinental analysis of climate model data: definition and aggregated datasets. *Earth System Science Data*, 12(4): 2959–2970, 2020. 11

[50] Sonia I. Seneviratne, Xuebin Zhang, Muhammad Adnan, Wafae Badi, Claudine Dereczynski, Alejandro Di Luca, Subimal Ghosh, Ismail Iskandar, James Kossin, Sophie Lewis, Friederike Otto, Izidine Pinto, Masaki Satoh, Sergio M. Vicente-Serrano, Michael Wehner, and Botao Zhou. Weather and climate extreme events in a changing climate. In Valérie Masson-Delmotte, Panmao Zhai, Anna Pirani, Sarah L. Connors, Clotilde Péan, Sophie Berger, Nada Caud, Yang Chen, Leah Goldfarb, Melissa I. Gomis, Mengtian Huang, Katherine Leitzell, Elisabeth Lonnoy, J. B. Robin Matthews, Tom K. Maycock, Tim Waterfield, Özge Yelekçi, Rong Yu, and Botao Zhou, editors, *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 1513–1766. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021. doi: 10.1017/9781009157896.013. URL https://doi.org/10.1017/9781009157896.013. 11, 12