

Supporting Information for

Downscaling with AI reveals the large role of internal variability in fine-scale projections of climate extremes

Neelesh Rampal^{1,3,4}, Peter B. Gibson¹, Steven Sherwood^{3,4}, Laura Queen¹,
Hamish Lewis^{1,2}, Gab Abramowitz⁴

¹ National Institute of Water and Atmospheric Research (NIWA), New Zealand

² Te Aka Mātutua School of Science, University of Waikato, Hamilton, New Zealand

³ ARC Centre of Excellence for Weather of the 21st Century,
University of New South Wales, Sydney, Australia

⁴ Climate Change Research Centre, University of New South Wales, Sydney, Australia

Overview

This document contains the supporting information for the manuscript titled: *Downscaling with AI reveals large role of internal variability in fine-scale projections of climate extremes*. The contents (Figures S1–S44) are divided into two sections as outlined below with several subsections.

Contents

• Results

- Emulator performance in historical and future climates (Figures S1–S22)
 - * Historical Evaluation (Figures S1–S14)
 - * Climate Change Signal Evaluation (Figures S14–S22)
- A range of possible futures for temperature and precipitation extremes (Figures S23–S27)
- Uncertainty decomposition at fine scales (Figures S28–S34)

• Methods

- Comparison of the perfect vs imperfect framework (Figures S35–S38)
- Further sensitivity testing of the emulator (Figures S39–S40)
- Selection of GCMs (Figure S41)

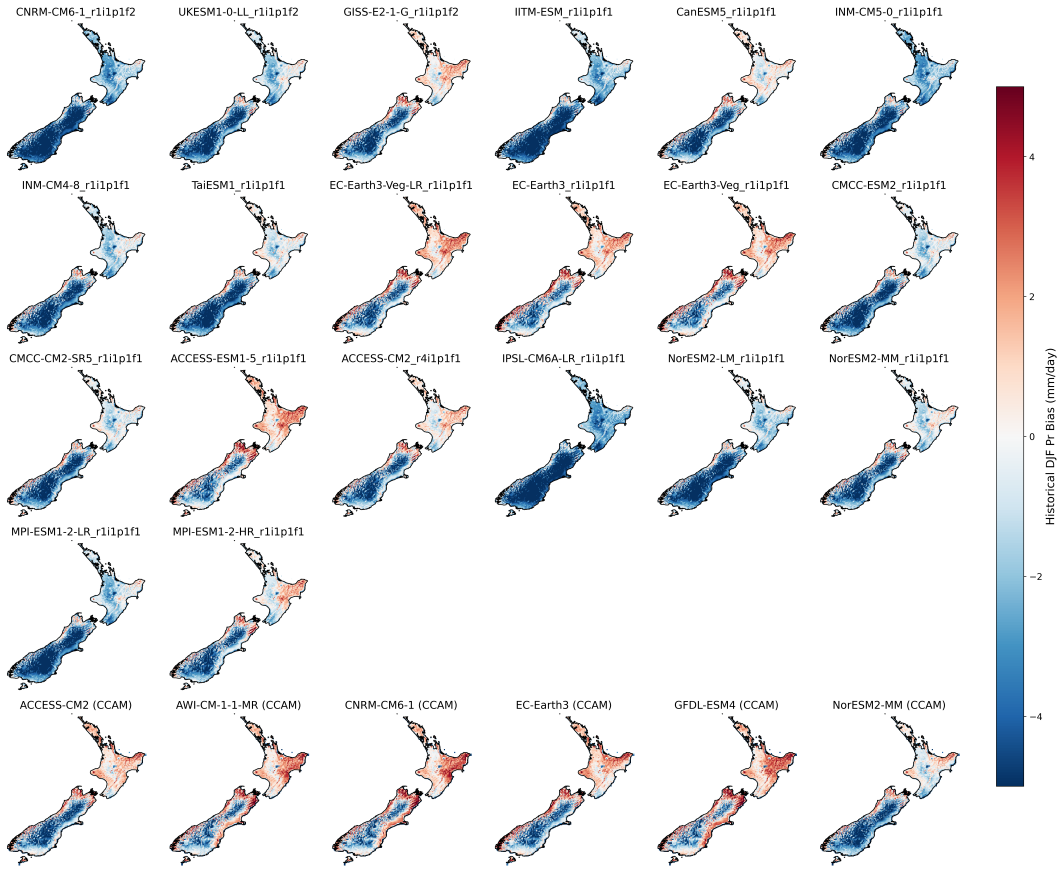
- Methods for uncertainty decomposition (Figures S42–S43)
- Methods for Generalized Extreme Value distribution Fits (Figures S44)

Results

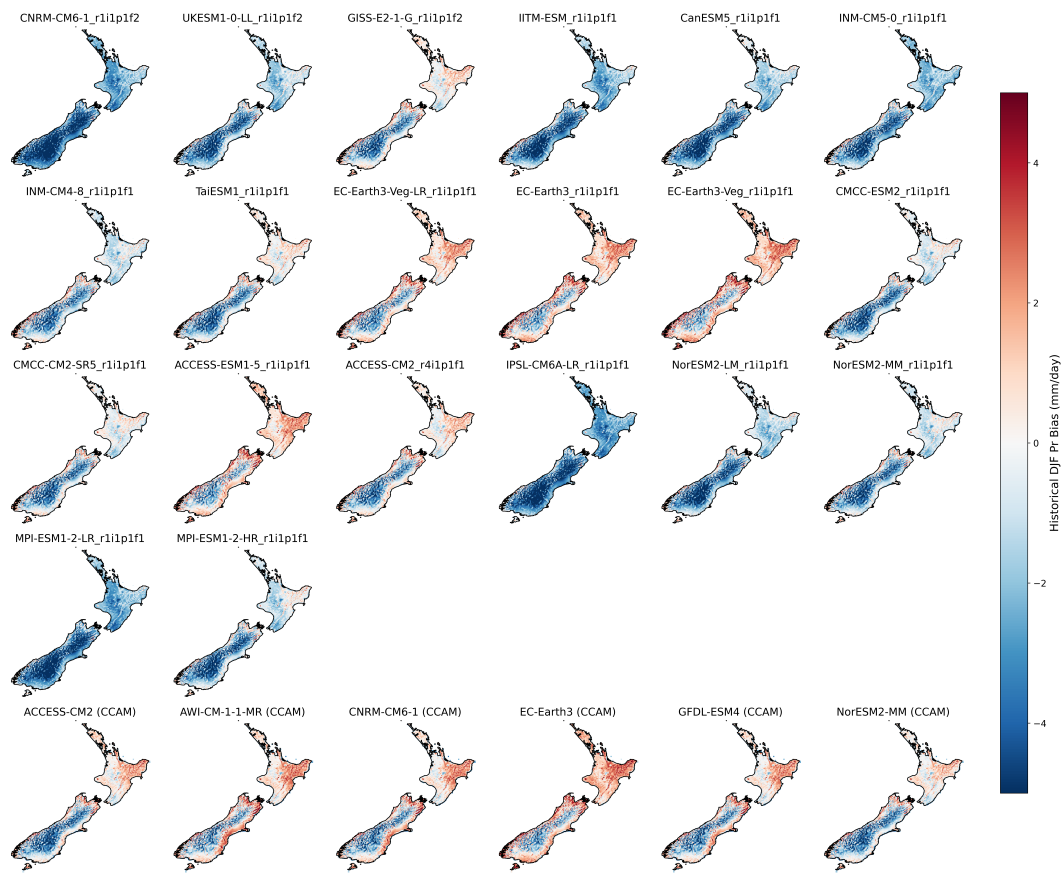
Emulator Performance in Historical and Future Climates

This section contains figures related to the above section in the main manuscript. We first begin by focusing on performance the historical period of simulation, where evaluation is performed against the VCSN observational dataset.

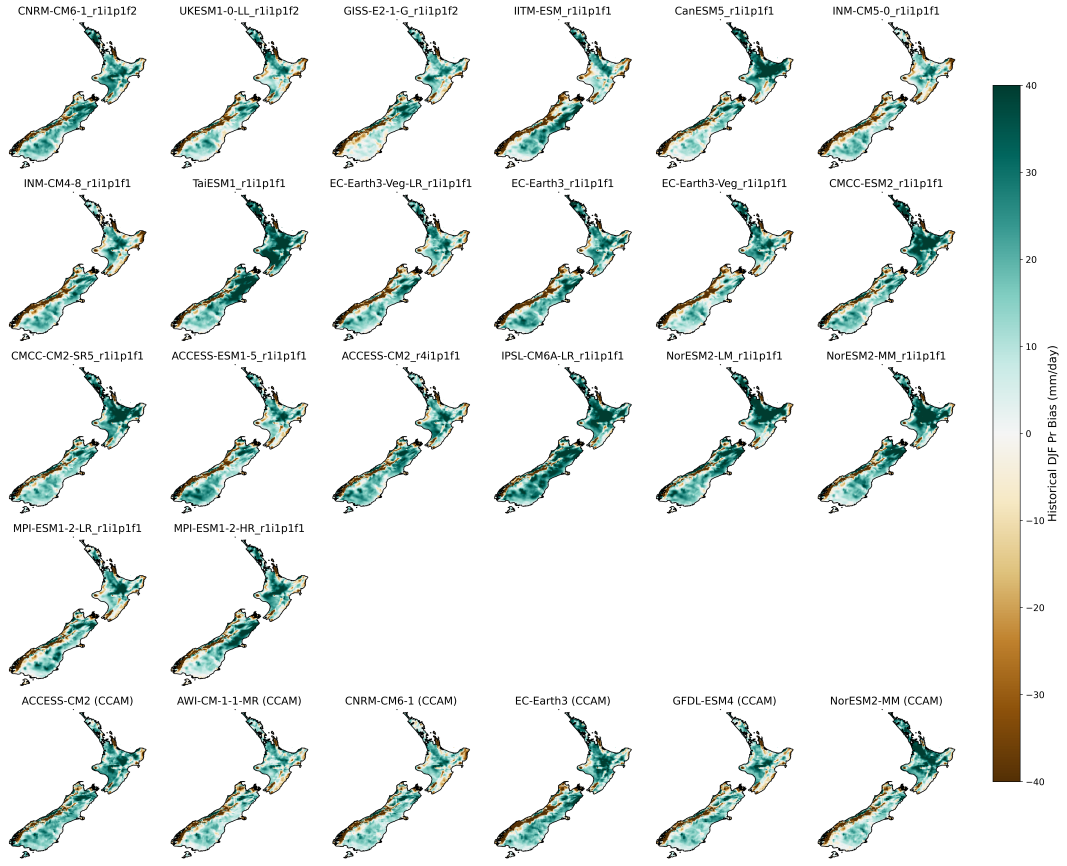
Historical Evaluation (Figures S1-S13)



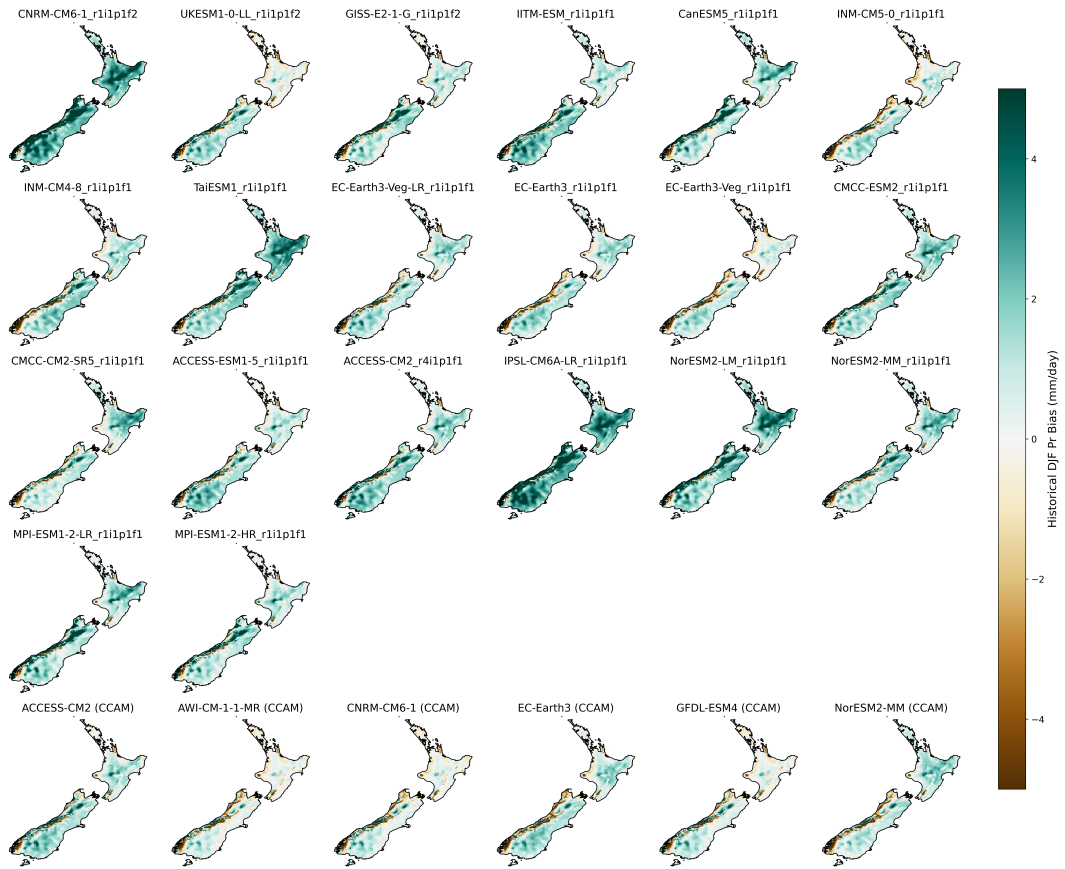
Supplementary Figure S1: Spatial patterns of historical climatological biases for TXx (1986-2005), comparing the emulator (top rows) and the ground truth RCM (bottom rows) relative to VCSN. Units are in °C.



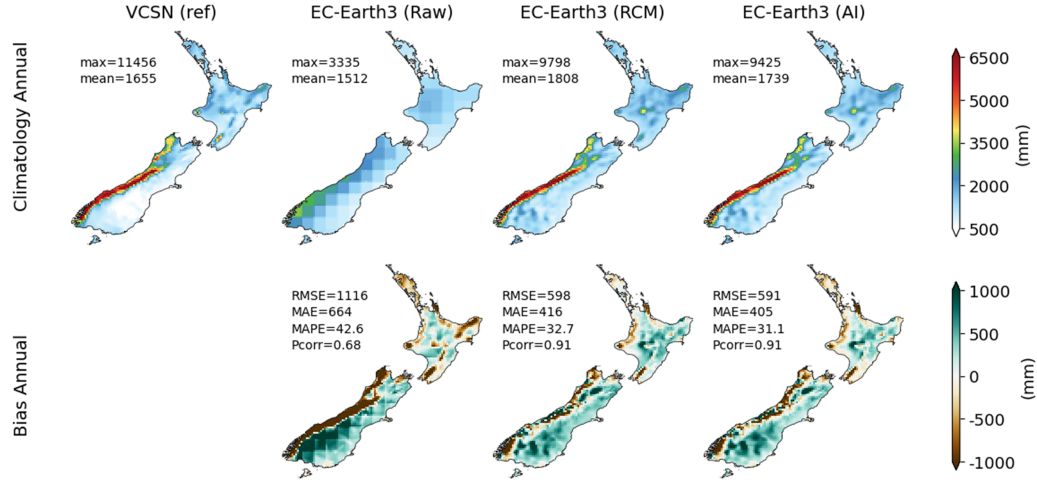
Supplementary Figure S2: Same as Figure S1, but for DJF Tasmax.



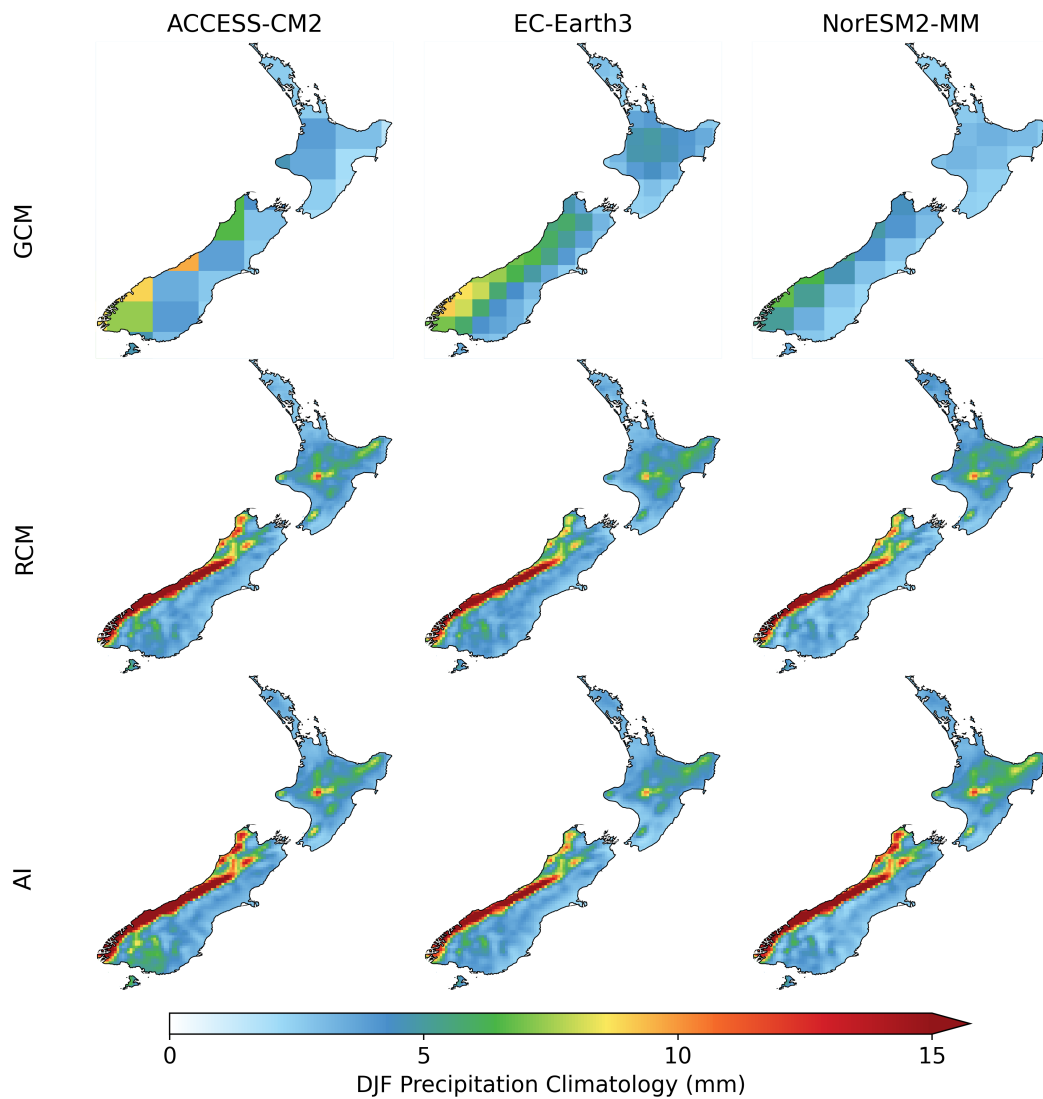
Supplementary Figure S3: Same as Figure S1, but for RX1Day. Errors are in units of mm/day.



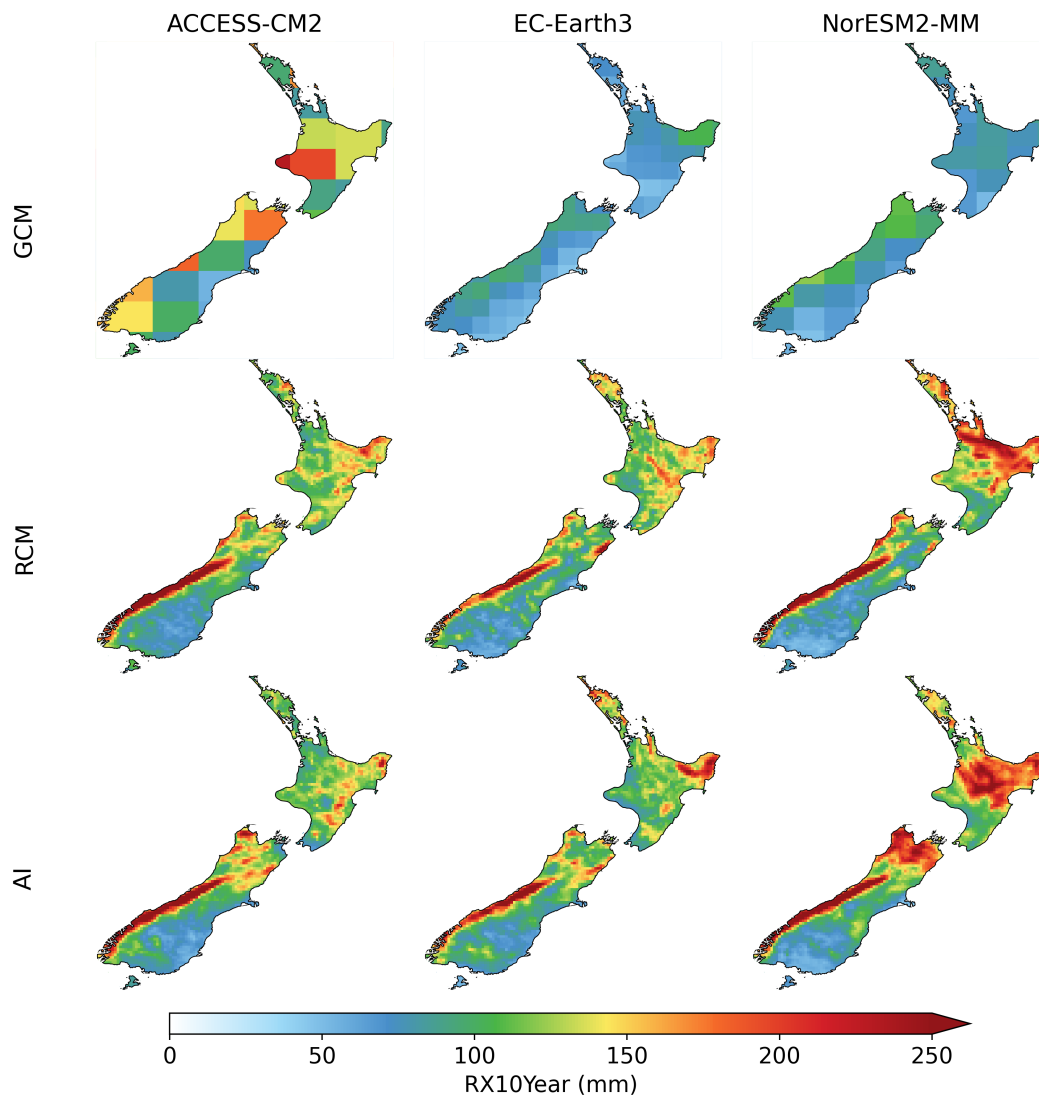
Supplementary Figure S4: Same as Figure S3, but for DJF seasonal mean precipitation.



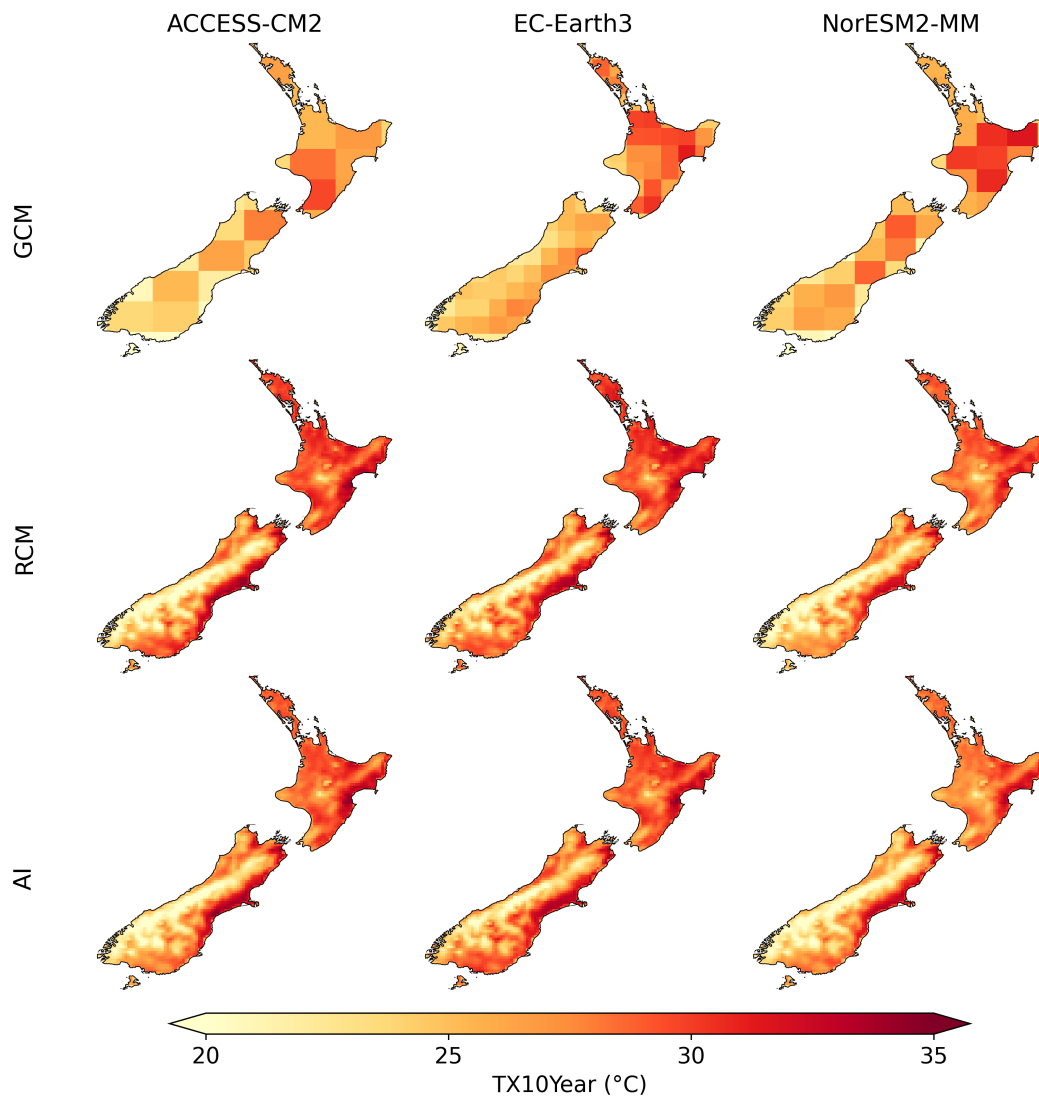
Supplementary Figure S5: An illustration of the improvement of downscaling (i.e. added value) for the annual mean precipitation climatolgy (1986-2005). The top row displays historical climatolgy from left to right: VCSN observations (reference), raw EC-Earth3 GCM, RCM, and the emulator (AI). The bottom row highlights biases against observations, illustrating that the emulator's (AI) biases closely resemble those of the RCM.



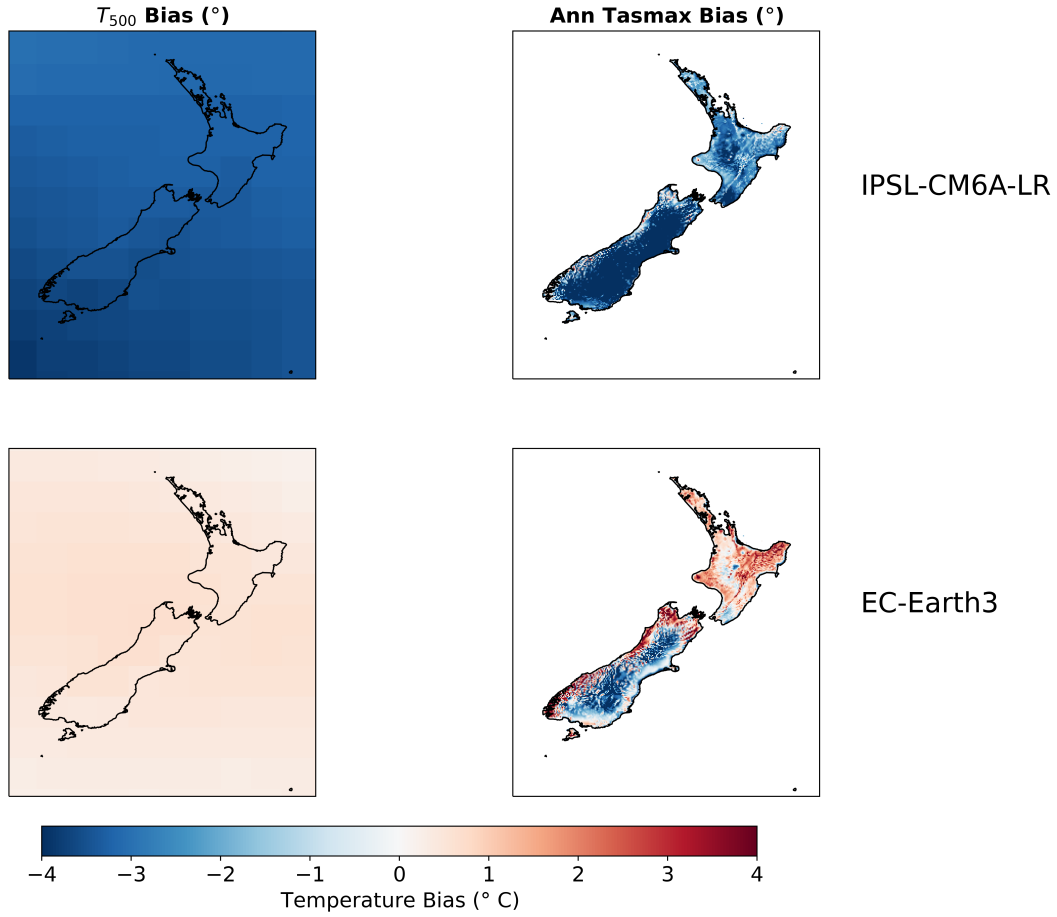
Supplementary Figure S6: Historical climatologies (1986-2005) of DJF precipitation climatology. The top row shows GCM fields, the second row displays the RCM, and the last row features emulator-downscaled climatologies. Note that while ACCESS-CM2 GCM was used for training, EC-Earth3 and NorESM2-MM simulations serve as out-of-sample evaluations. This figure shows that the RCM and emulator (AI) capture the climatologies with a similar level of detail.



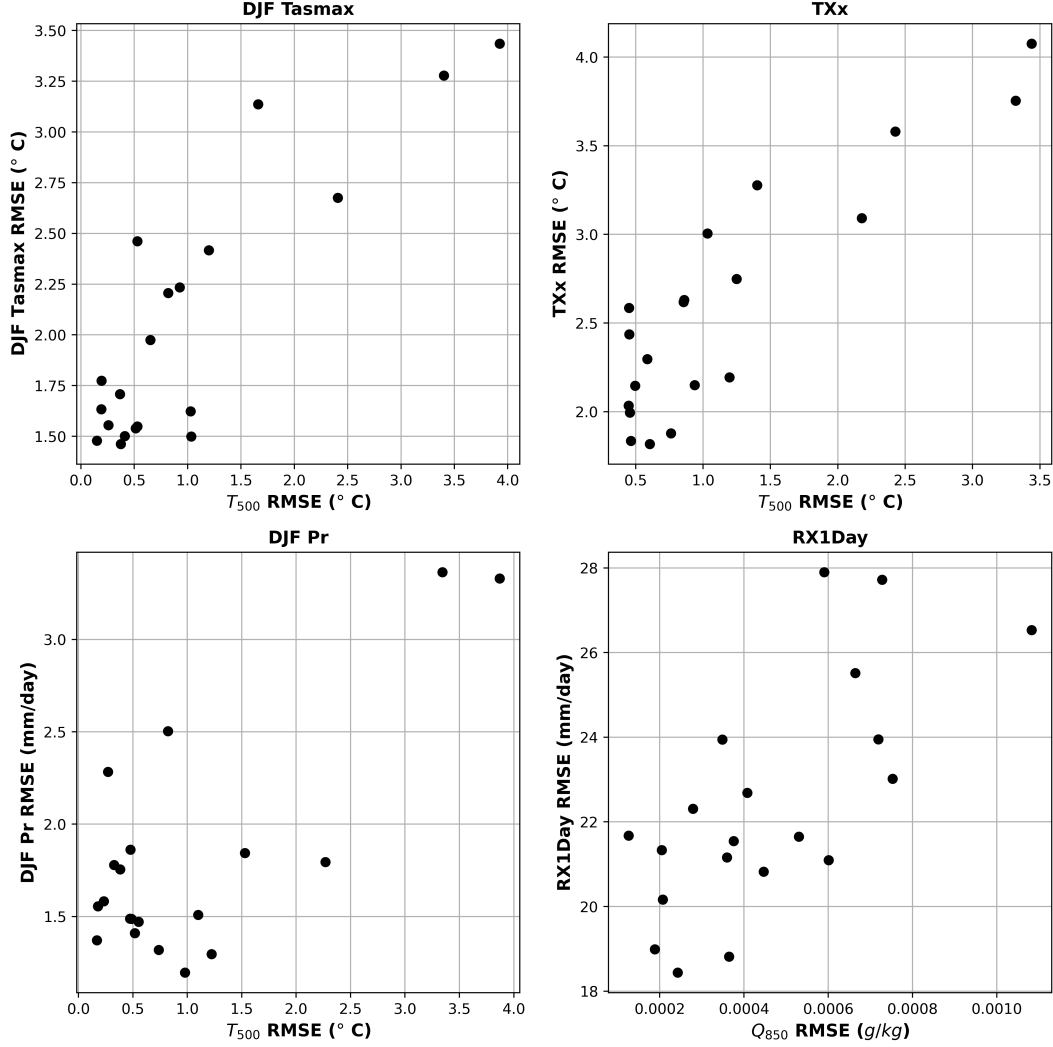
Supplementary Figure S7: Same as Figure S6, but for RX [10 year] (wettest-day-per-decade).



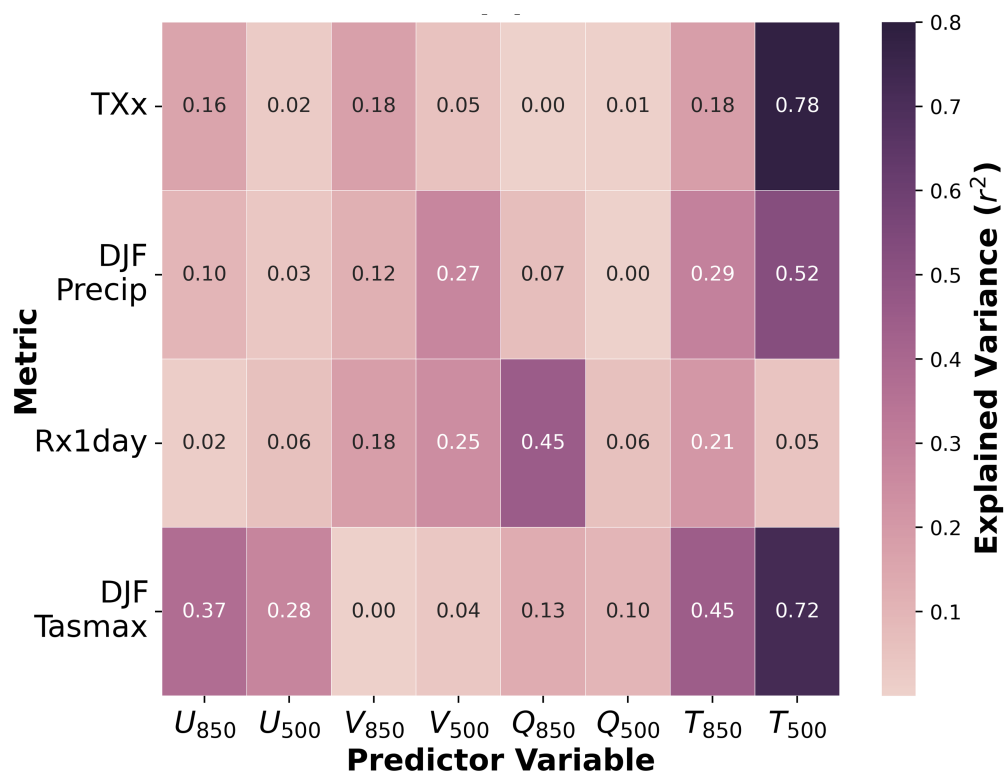
Supplementary Figure S8: Same as Figure S6, but for TX [10 year].



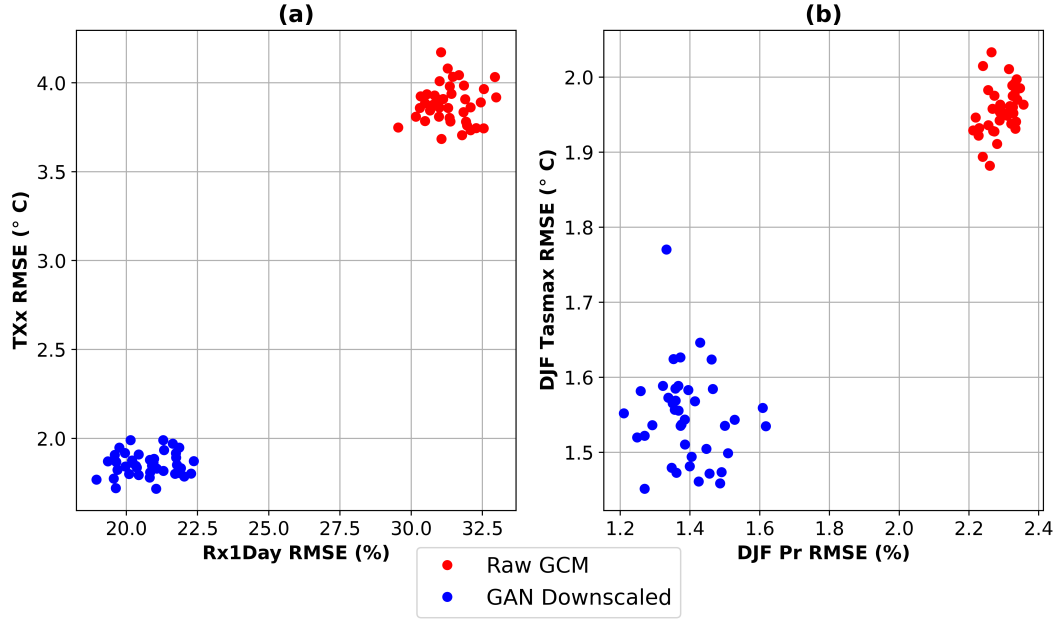
Supplementary Figure S9: Illustration of historical annual T_{500} biases (1986–2005) in the driving GCMs (left), computed relative to ERA5 regridded to 1.5° , and how these biases are inherited by the RCM emulator during downscaling for IPSL-CM6A-LR and EC-Earth3 (right), with RCM biases computed relative to VCSN tasmax observations over the same climatology period.



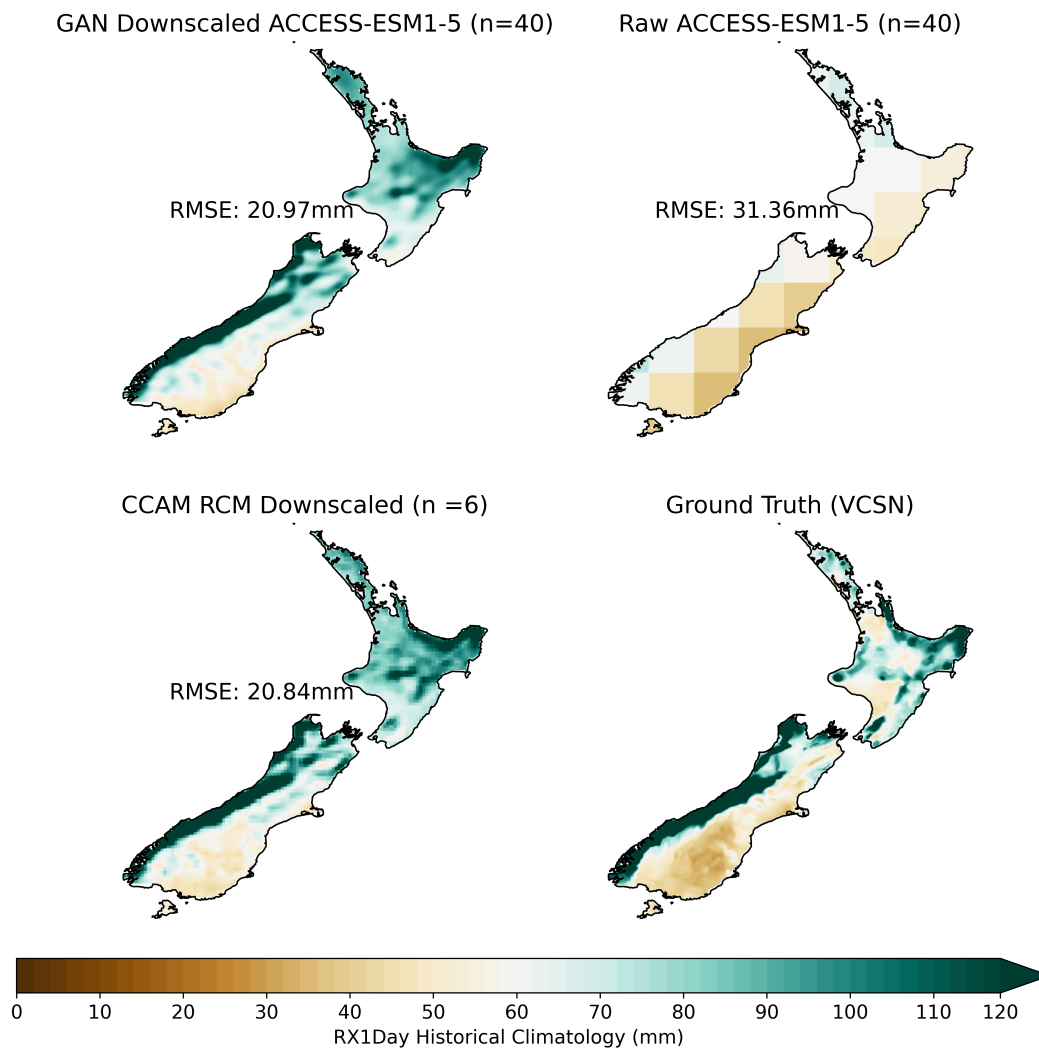
Supplementary Figure S10: Relationship between GCM ($n=20$) input biases (measured against ERA5 ground truth) and downscaled performance (measured against VCSN; observational reference), across four metrics (DJF tasmax, TXx, DJF pr, RX1Day). The x-axis displays the spatially averaged Root Mean Square Error (RMSE) of the historical annual climatology (1986-2005) for GCM inputs, relative to ERA5 ground truth (re-gridded to $1.5^\circ \times 1.5^\circ$). The x-axis shows the GCM input biases for T_{500} , except for RX1Day uses Q_{850} . The y-axis shows the RMSE of downscaled outputs (DJF tasmax, TXx, DJF PR, RX1Day) relative to gridded observations. These results suggest a link between errors in GCM inputs and the performance of downscaling efforts. Correlations between the metrics and GCM inputs are shown in Figure S11.



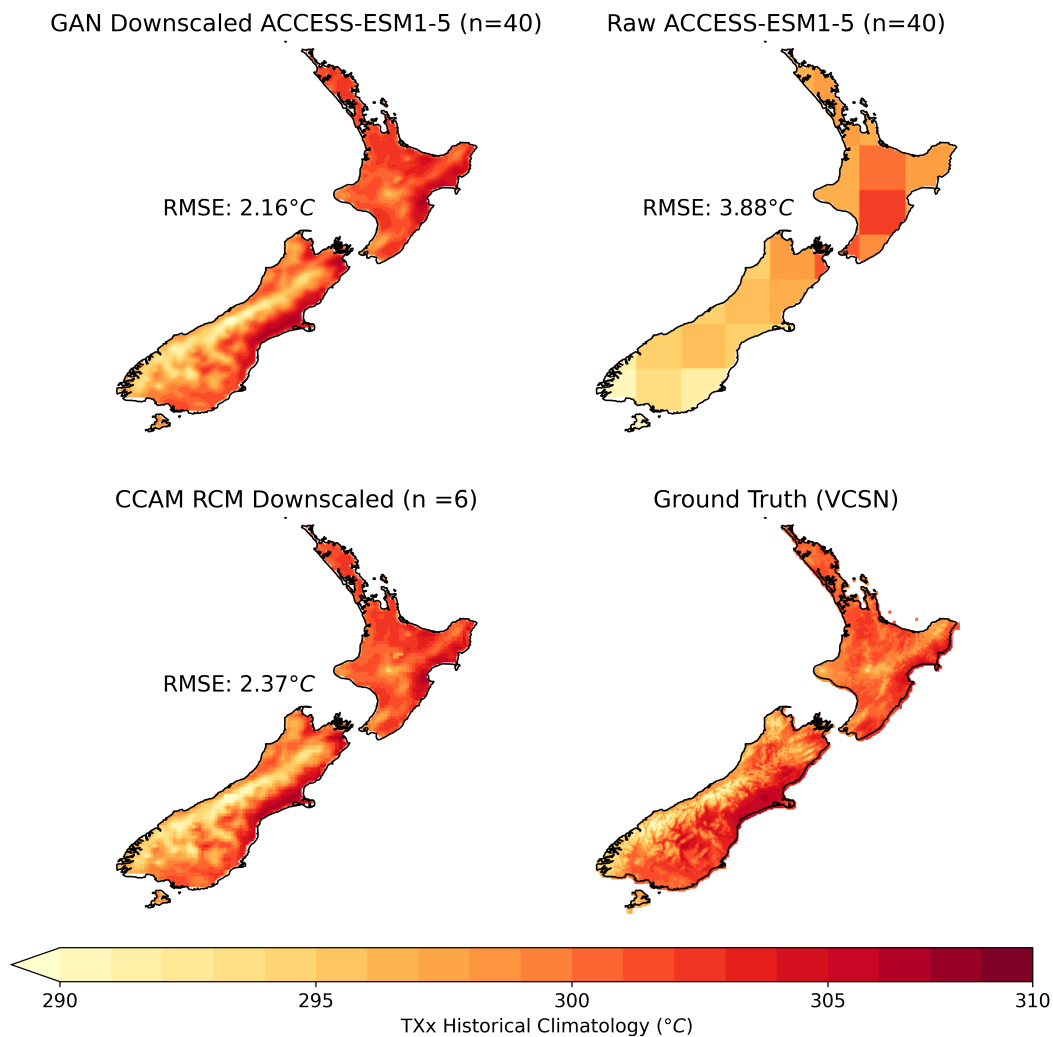
Supplementary Figure S11: Correlation between RMSE in GCM inputs (as in Figure S10) and downscaled biases relative to VCSN observations for each metric (y-axis). The strongest correlations are highlighted in Figure S10.



Supplementary Figure S12: This shows the Root Mean Square Error (RMSE) of downscaling versus no downscaling for the ACCESS-ESM1-5 ensemble ($n=40$). (a) TXx and RX1Day biases and (b) DJF tasmax and Pr biases are presented for both the emulator (GAN downscaled) and raw ACCESS-ESM1-5 fields. Each point denotes the bias of a single ensemble member.



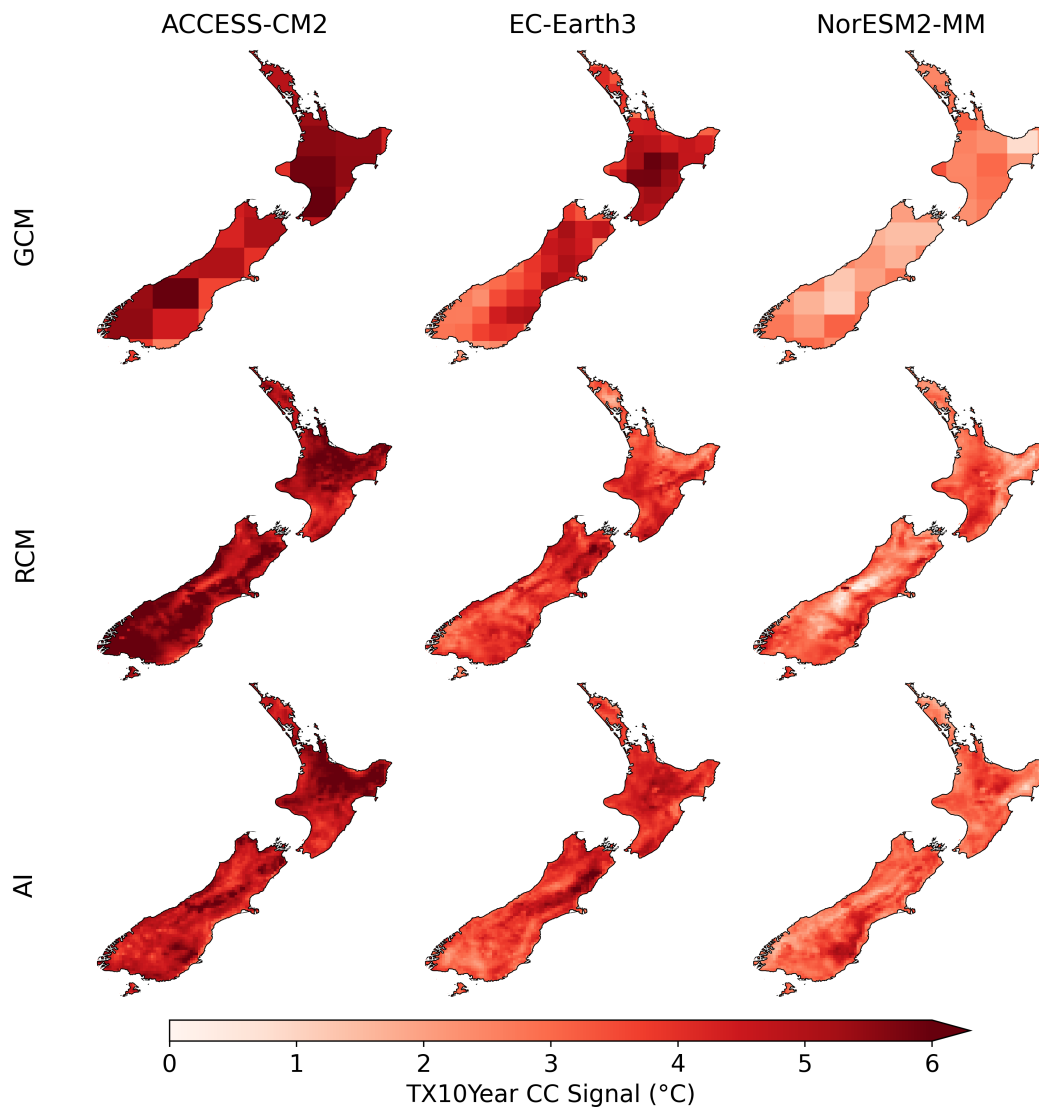
Supplementary Figure S13: This figure compares the historical climatology (1986-2005) of RX1Day from the ACCESS-ESM1-5 large ensemble (represented by its ensemble average) against other dynamical downscaled simulations and observations. It compares the GAN downscaled ensemble (top left), raw ACCESS-ESM1-5 (top right), and CCAM (bottom left) against the observed climatology (bottom right, ground truth). RMSE values are computed relative to the ground truth. CCAM RCM downscaled simulations (an average of six historical GCMs, not ACCESS-ESM1-5) provide a general benchmark for RCM skill.



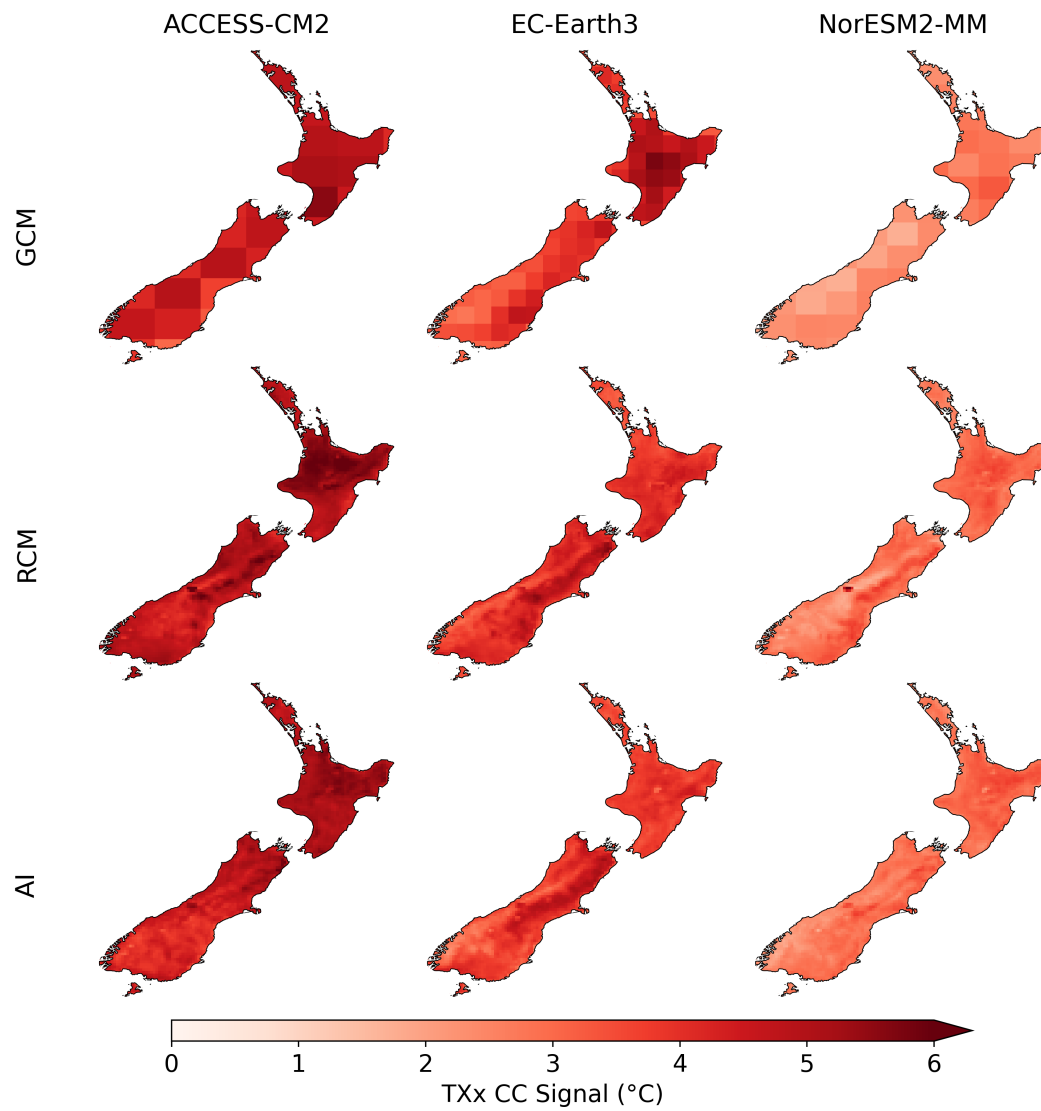
Supplementary Figure S14: Same as S13, but for TXx.

Climate Change Signal Evaluation (Figures S15-S22)

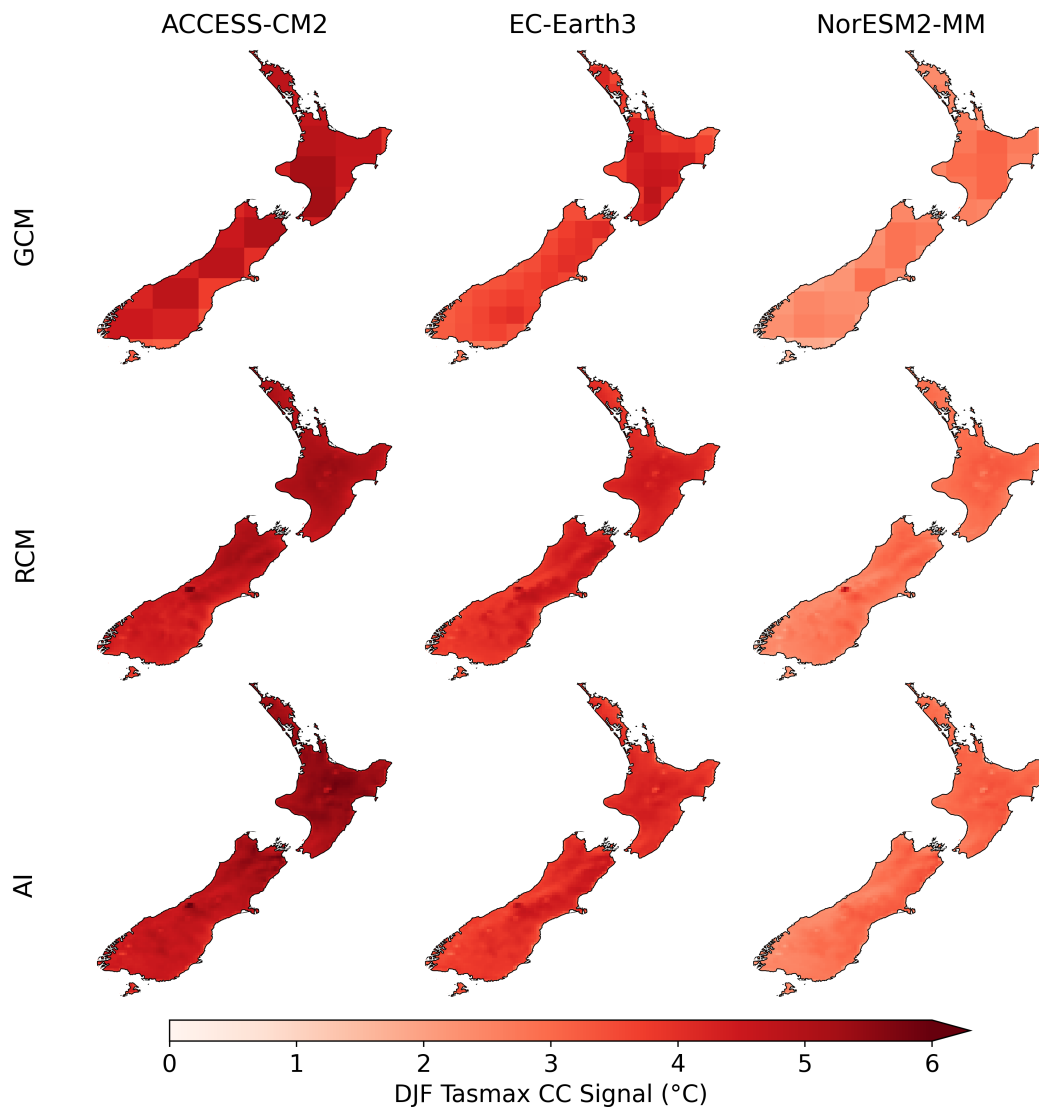
This section evaluates the emulator’s skill in capturing climate changes—defined as differences between future and historical climates—for seasonal means, annual extremes, and decadal extremes.



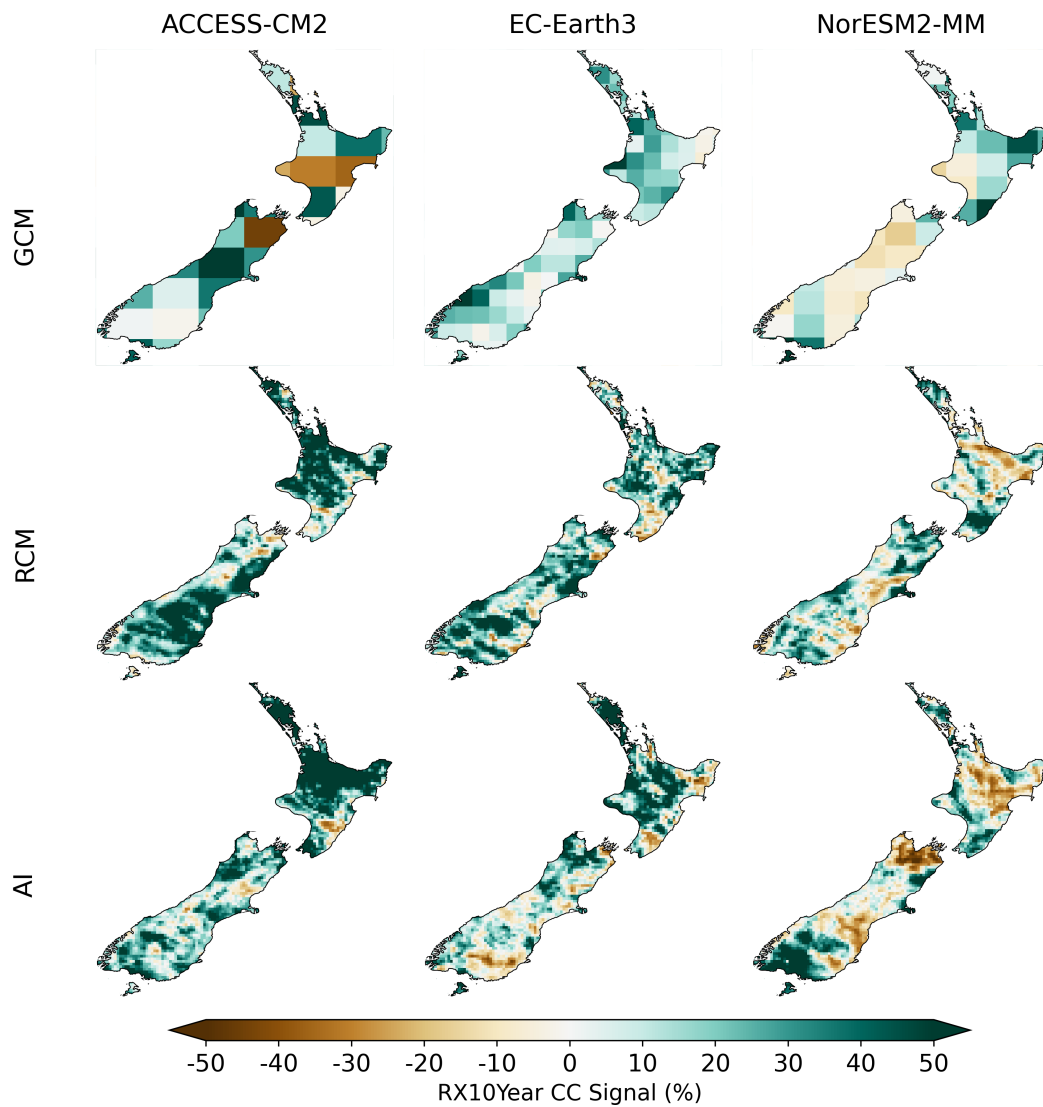
Supplementary Figure S15: Spatial patterns of the climate change signal in TX [10 year], defined as the difference between future (2080–2099) and historical (1986–2005) values under SSP3-7.0. The top row shows GCM outputs, the second row shows RCM results, and the third row shows emulator-downscaled signals. Results indicate that the emulator closely matches the RCMs for two out-of-sample GCMs (EC-Earth3 and NorESM2-MM), while ACCESS-CM2 was included in training.



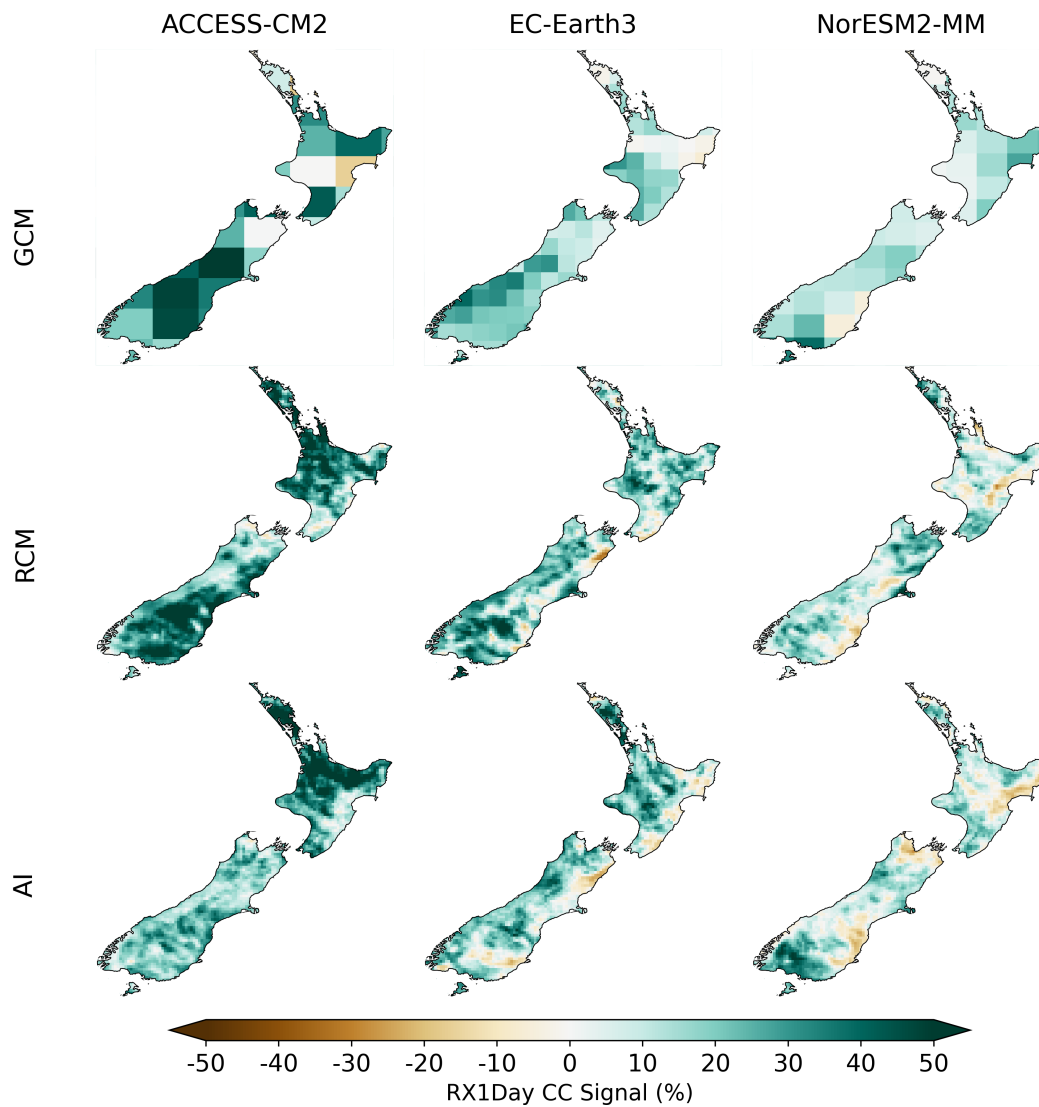
Supplementary Figure S16: Same as Figure S15, but for TXx.



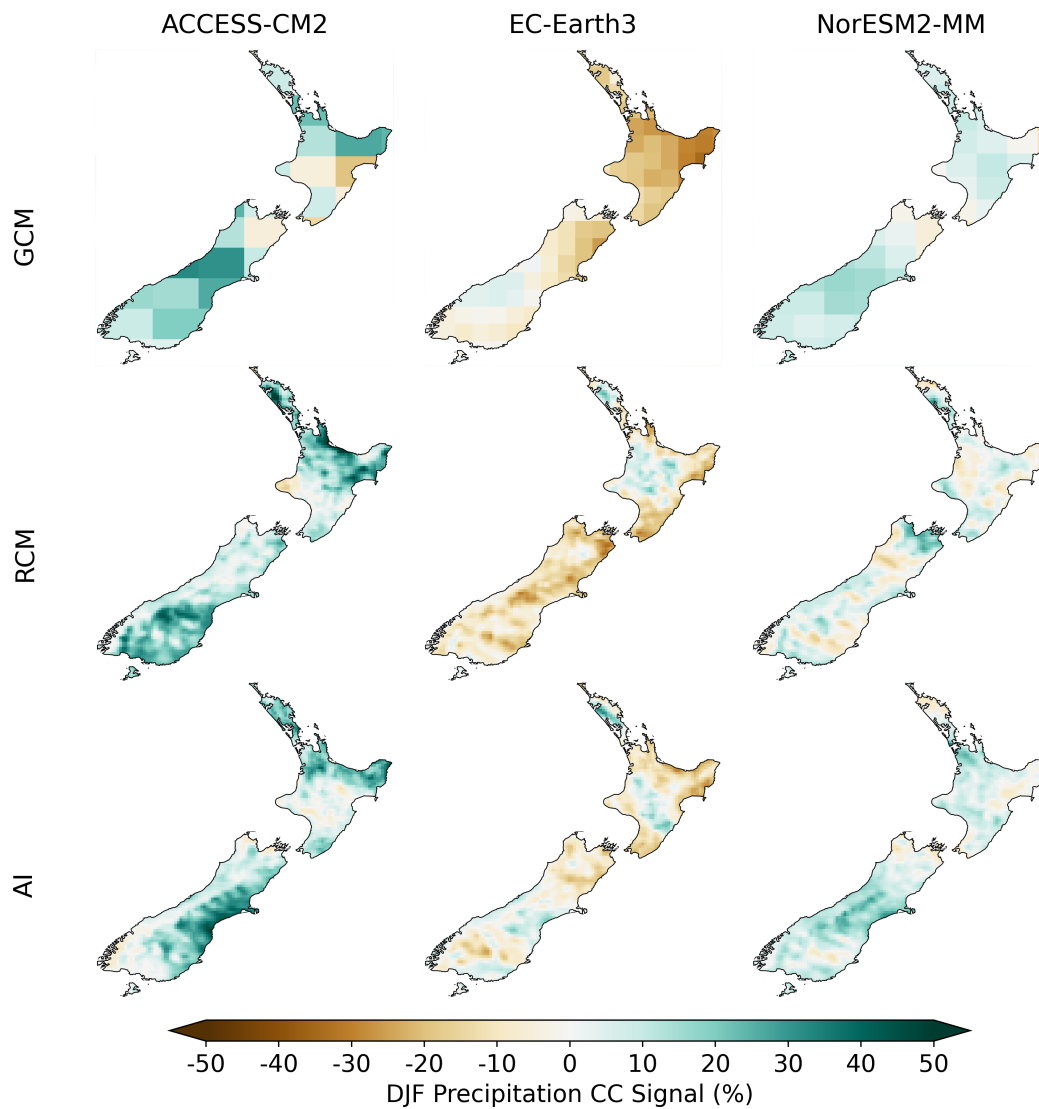
Supplementary Figure S17: Same as Figure S15, but for DJF Tasmax.



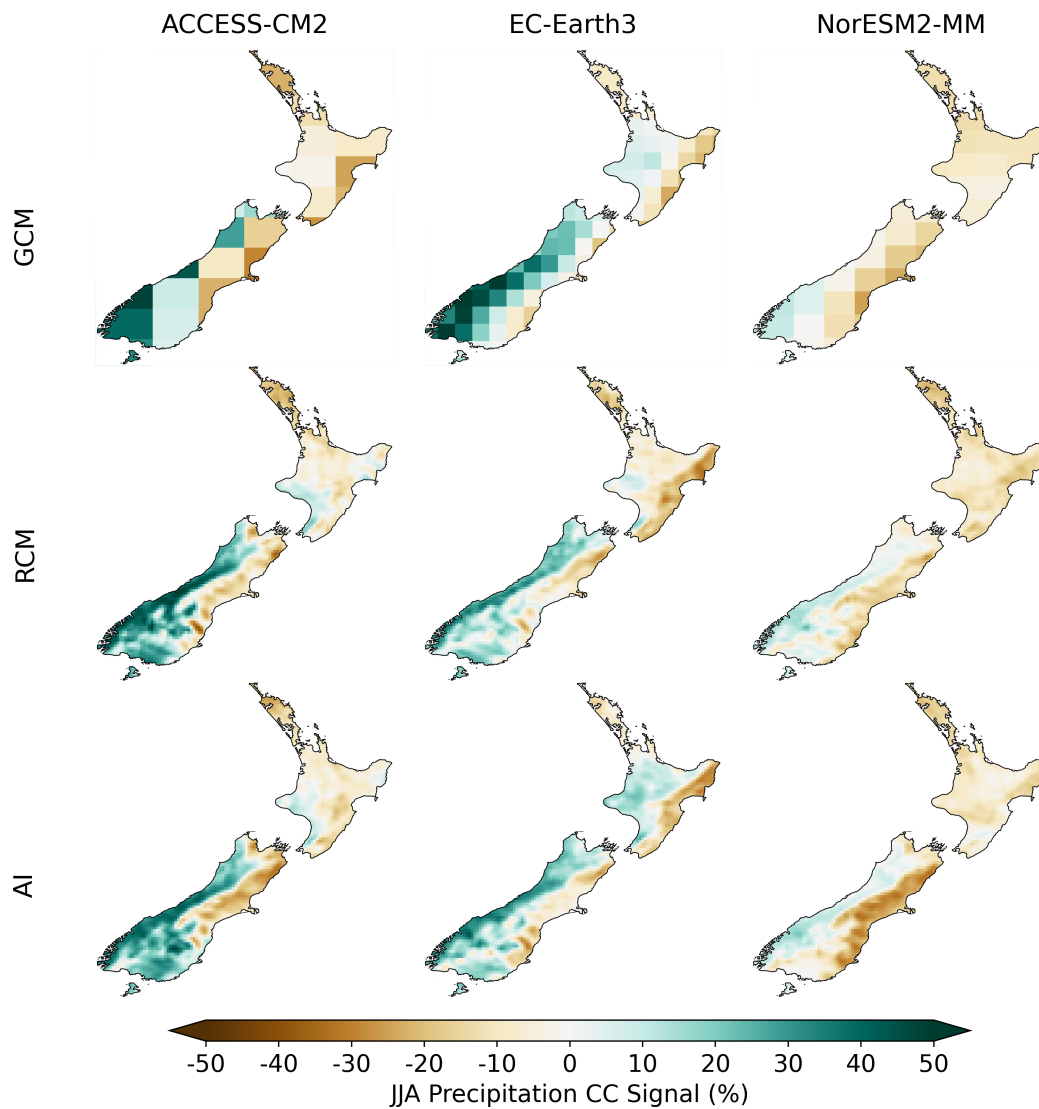
Supplementary Figure S18: Same as Figure S15, but for RX [10 year], and the climate change signal is the % difference between future and historical.



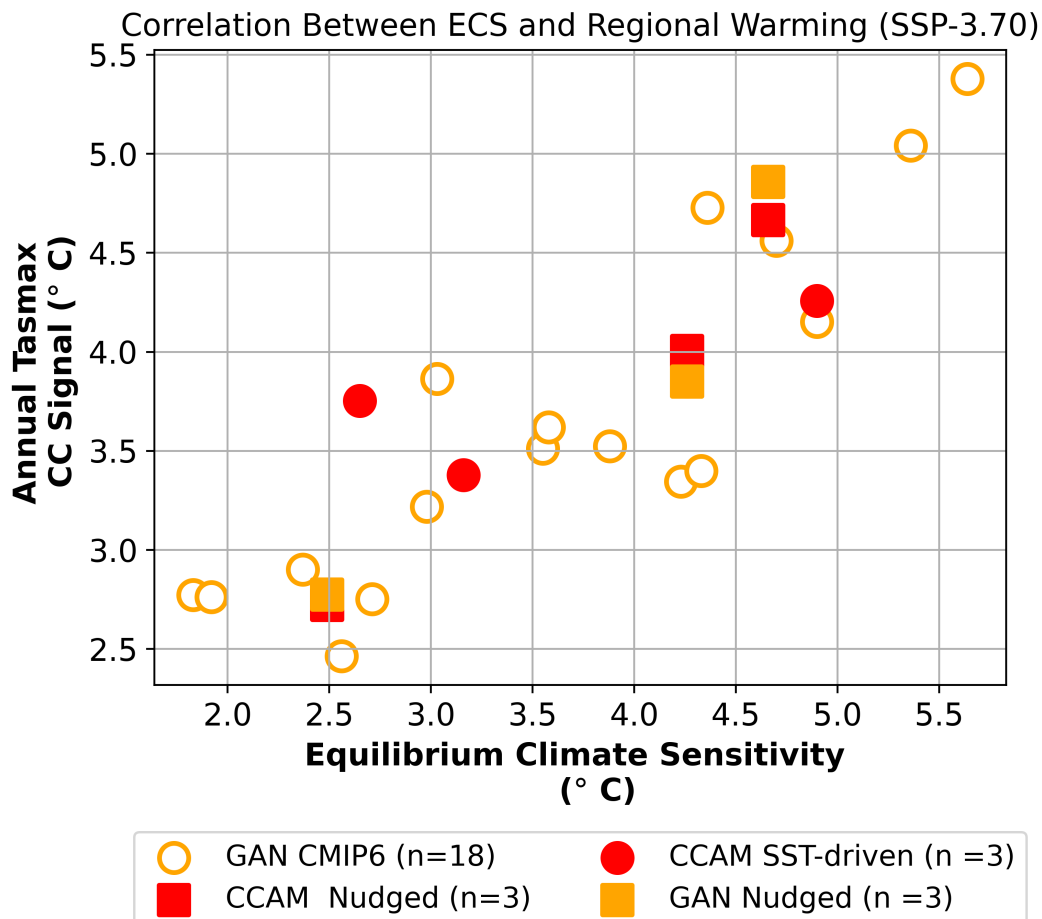
Supplementary Figure S19: Same as Figure S18, but for RX1Day.



Supplementary Figure S20: Same as Figure S18, but for DJF seasonal mean precipitation.



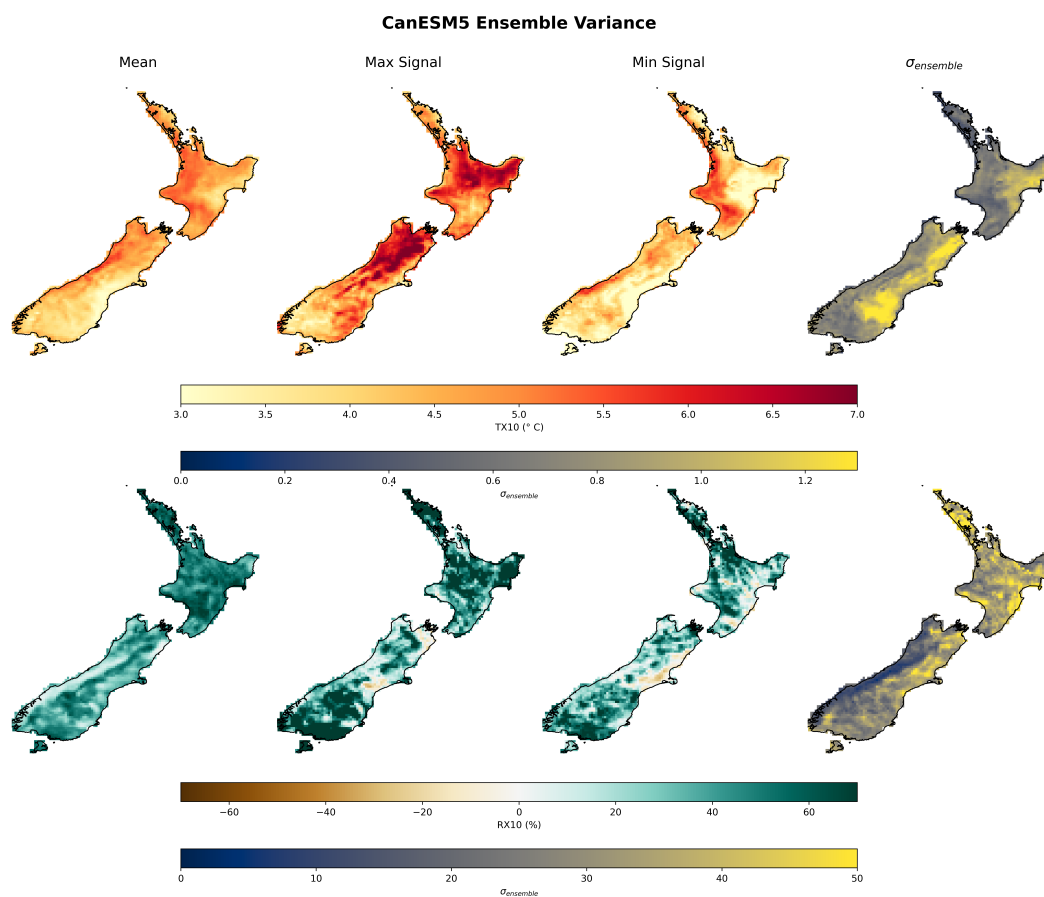
Supplementary Figure S21: Same as Figure S18, but for JJA seasonal mean precipitation.



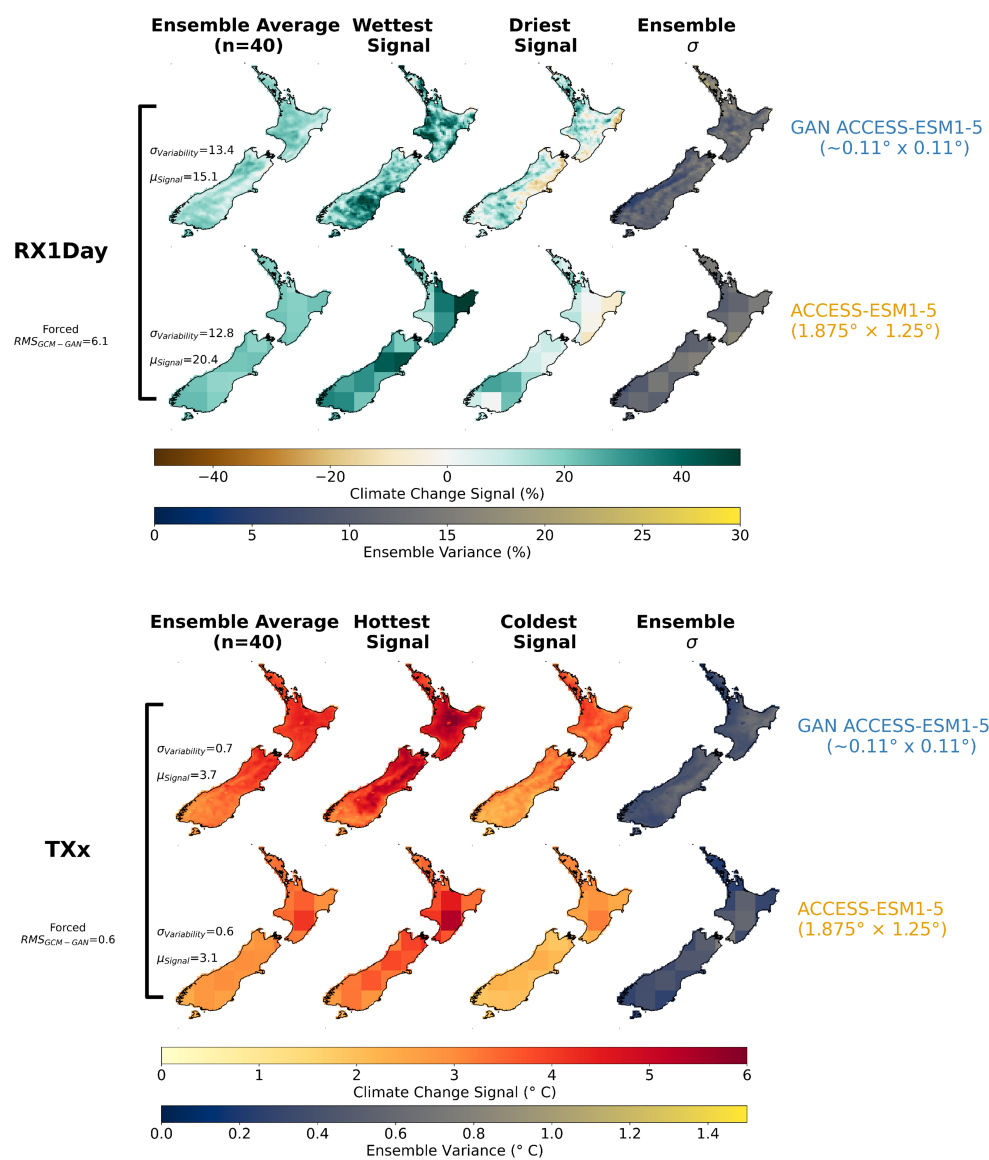
Supplementary Figure S22: Illustration of the relationship between GCM Equilibrium Climate Sensitivity (x-axis) and the emulator-downscaled annual temperature change (2080–2099 vs. 1986–2005) for the SSP3-7.0 scenario, averaged over land. Red squares show the three CCAM-nudged RCM simulations (ACCESS-CM2, EC-Earth3, NorESM2-MM), with corresponding GAN-downscaled signals (orange squares). Red circles represent SST-driven simulations not directly comparable to GAN outputs. Unfilled orange circles for the remaining 17 CMIP6 GCMs.

A range of possible futures for temperature and precipitation extremes (Figures S23-S27)

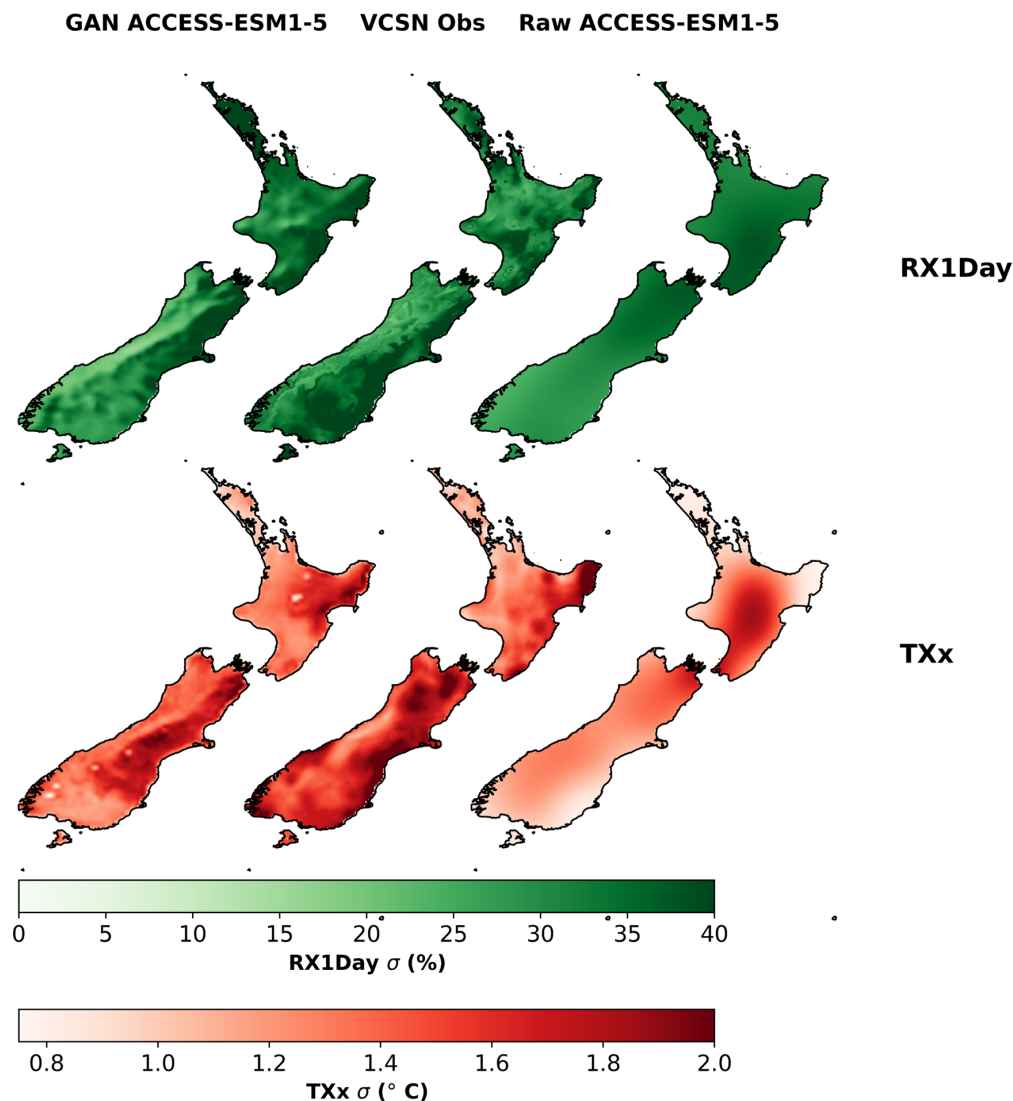
This section shows ensemble variance in the ACCESS-ESM1-5 and CanESM5 ensembles to complement the section “A range of possible futures for temperature and precipitation extremes”.



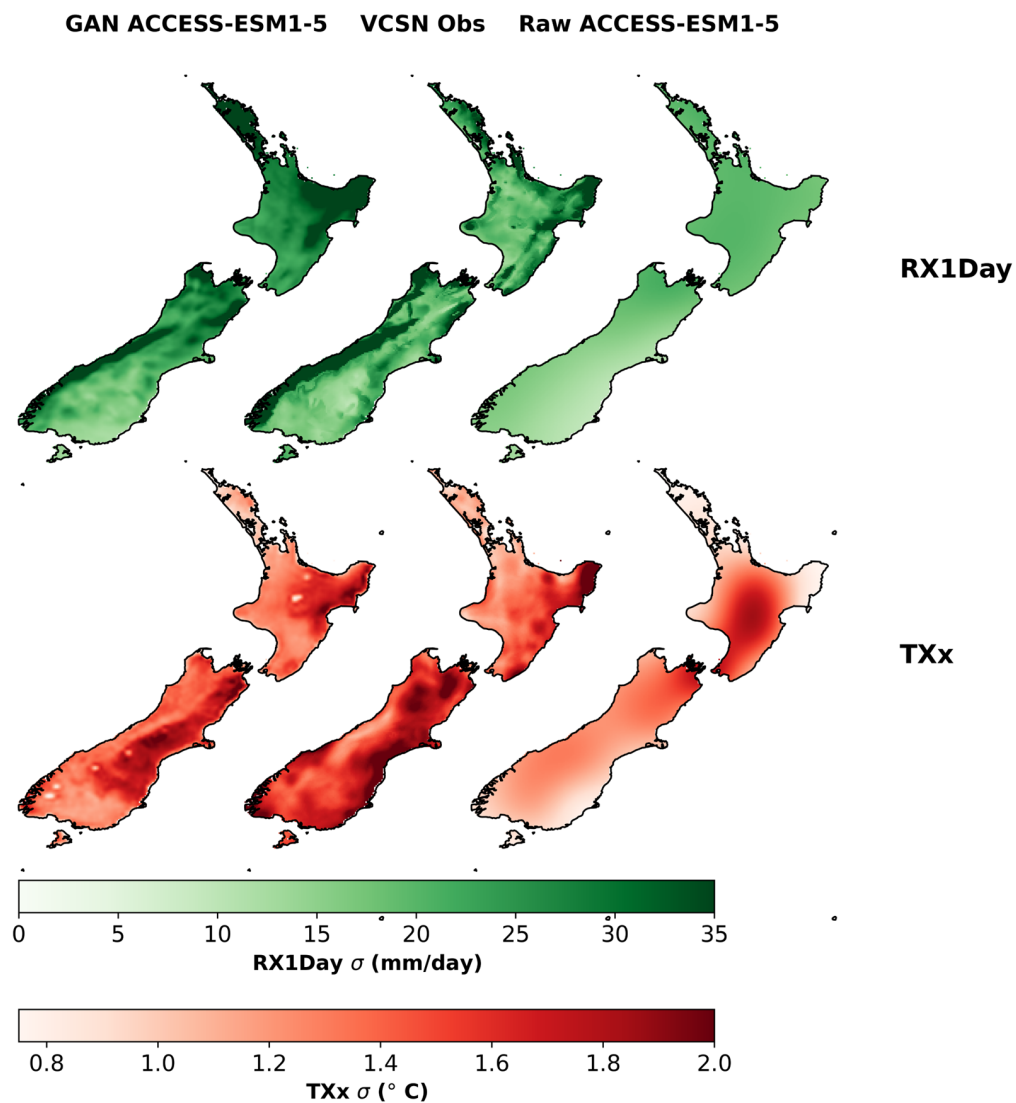
Supplementary Figure S23: Same as Figure 2 in the main text, but for the CanESM5 ensemble, showing only the emulator-downscaled outputs at 12 km resolution.



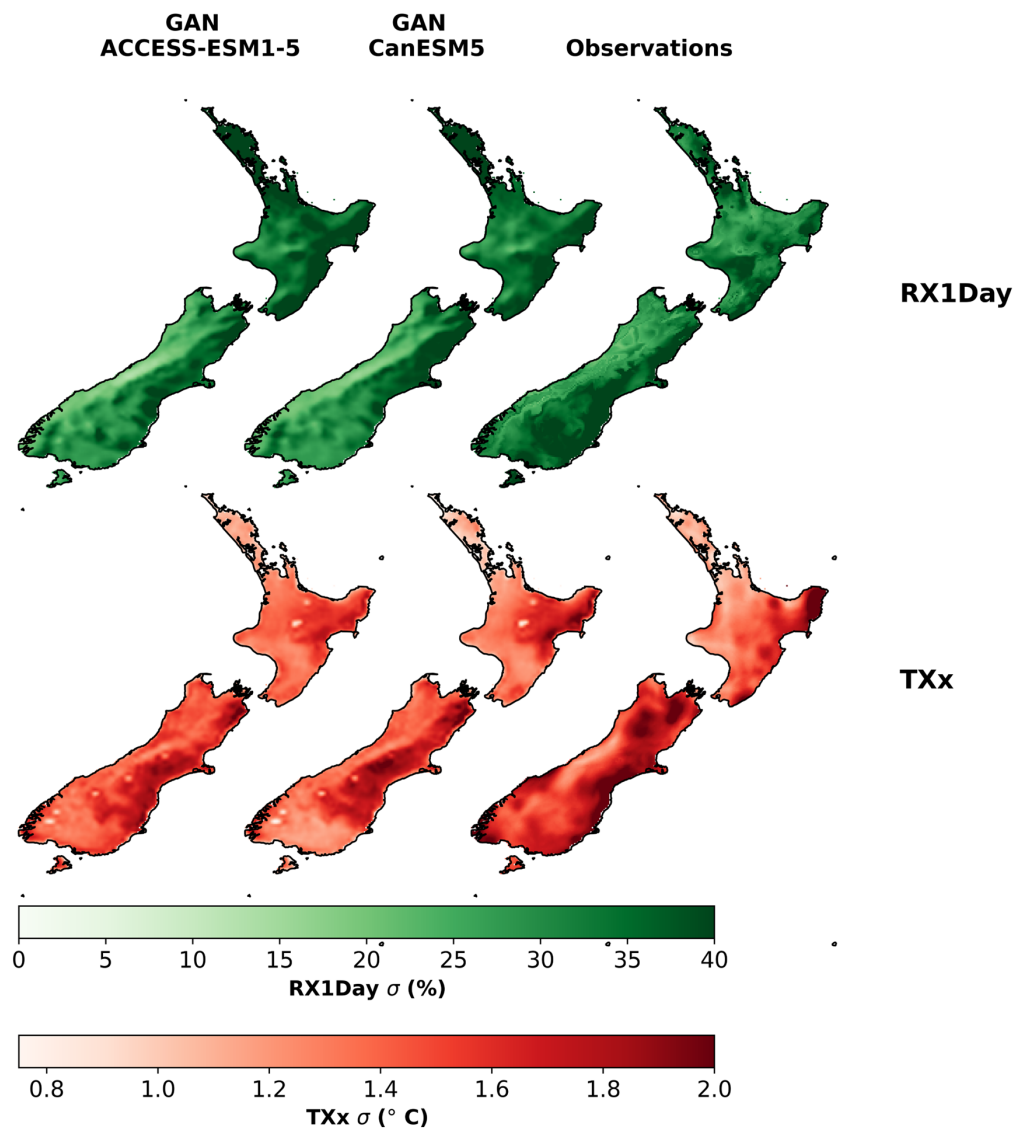
Supplementary Figure S24: Same as Figure 2 in the main text, but for annual extremes (RX1Day, TXx) in the ACCESS-ESM1-5 ensemble.



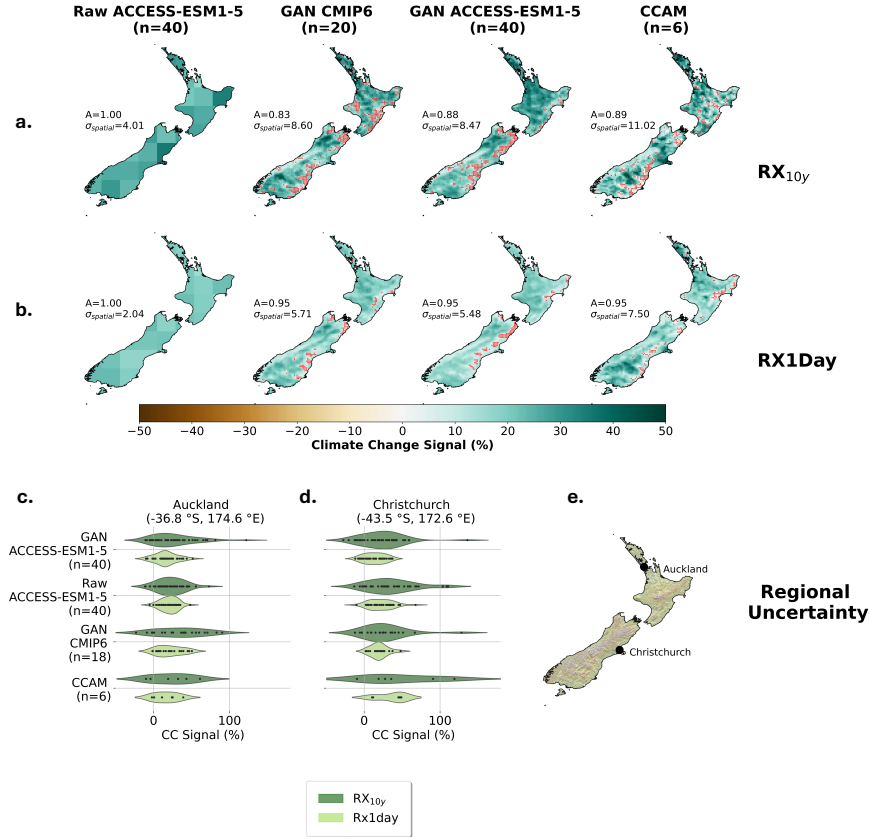
Supplementary Figure S25: Historical interannual variability (standard deviation over 1973–2014) in downscaled annual extremes from ACCESS-ESM1-5 (averaged across all members), compared with VCSN observations and raw ACCESS-ESM1-5. Variability is computed after removing the mean over this period, using absolute values for temperature and percentage anomalies for precipitation. This Figure is directly comparable to the ensemble variance in Figure 2.



Supplementary Figure S26: Same as Figure S25 but using absolute anomalies for RX1Day as a reference.

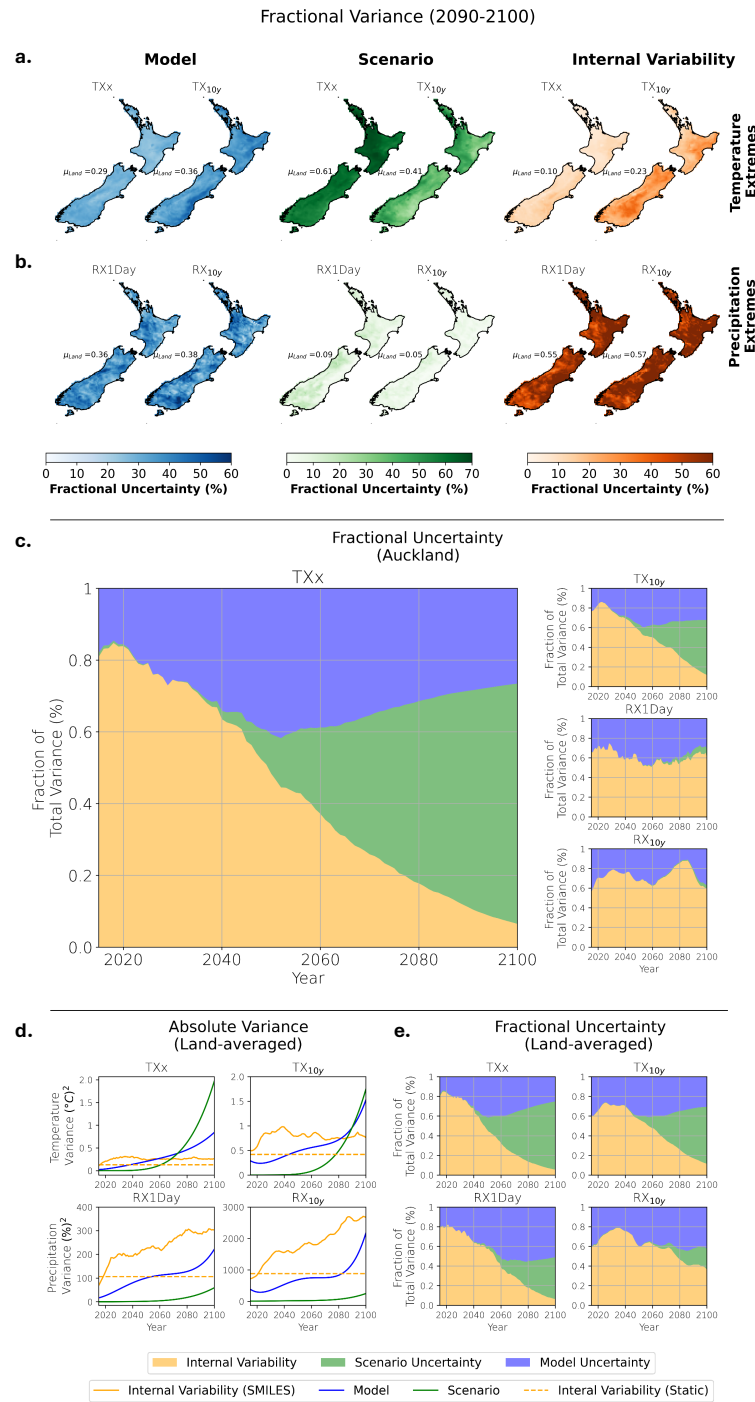


Supplementary Figure S27: Same as Figure S25 but including CanESM5.

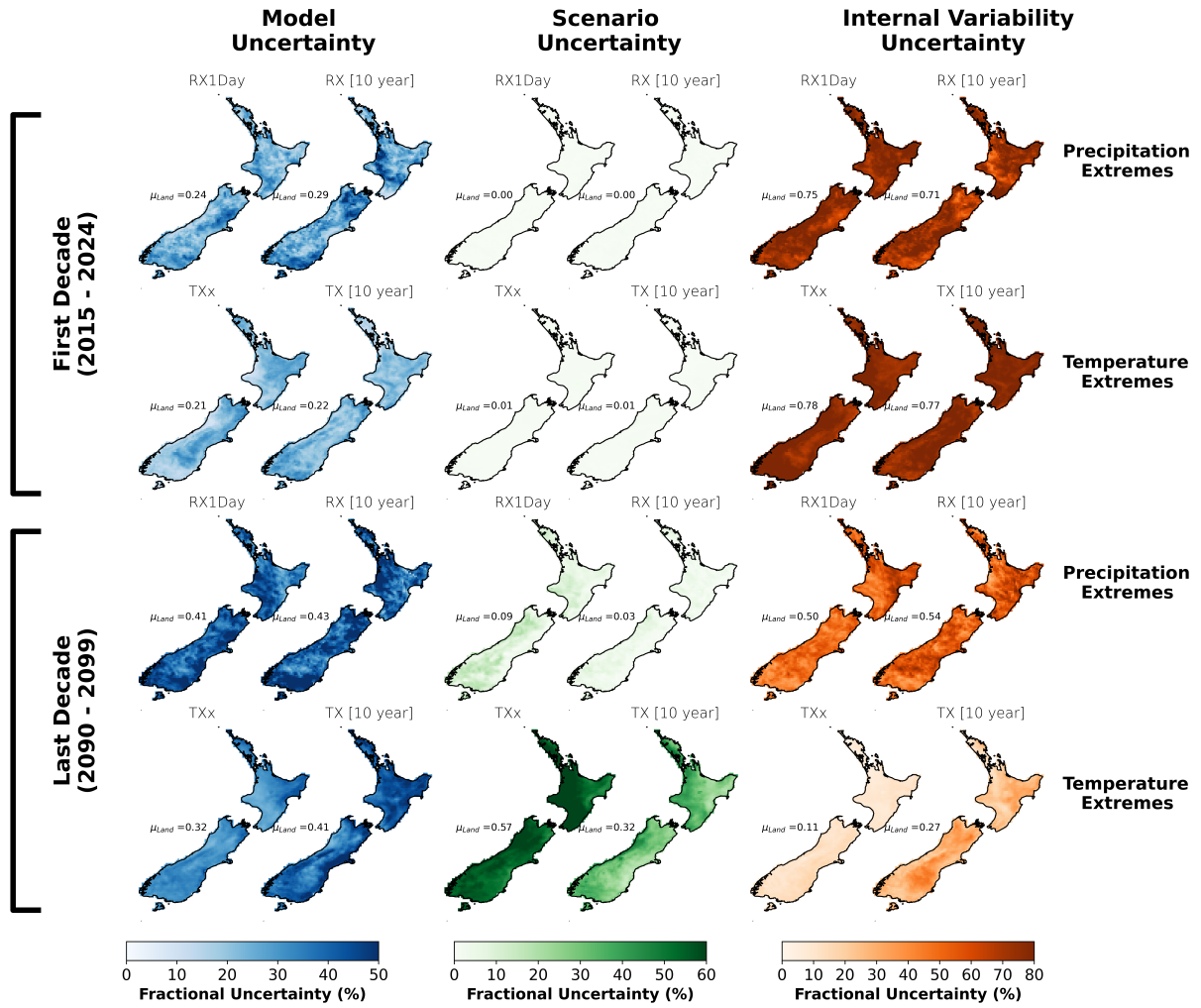


Supplementary Figure S28: Spread in fine-scale extreme precipitation projections from a large ensemble: a comparison of GCM, RCM, and emulator. End of century (2080–2099) climate change signals (%) for the SSP3-7.0 scenario relative to the historical climatology. (a–b): Ensemble-mean climate change signal in RX_{10y} (a) and RX_{1Day} (b) across four simulations—raw ACCESS-ESM1-5, GAN downscaled CMIP6 (20 members), GAN downscaled ACCESS-ESM1-5 (40 members), and CCAM (6 members). Red stippling is shown where fewer than 4/6 (66.6%) members agree on the sign of the change. σ_{spatial} shows the spatial standard deviation of the ensemble averages shown in (a) and (b), and A indicates the percentage of the area with sign agreement (regions without red stippling). Panels (c) and (d) display the spread of RX_{10y} (dark green) and Rx_{1Day} (light green) climate change signals across individual ensemble members for Auckland (c) and Christchurch (d), with locations shown in (e).

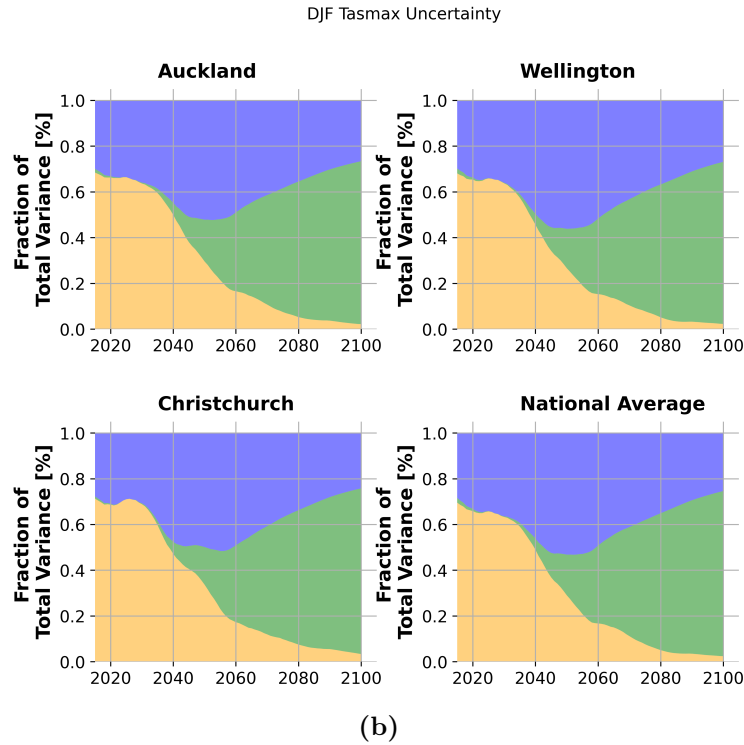
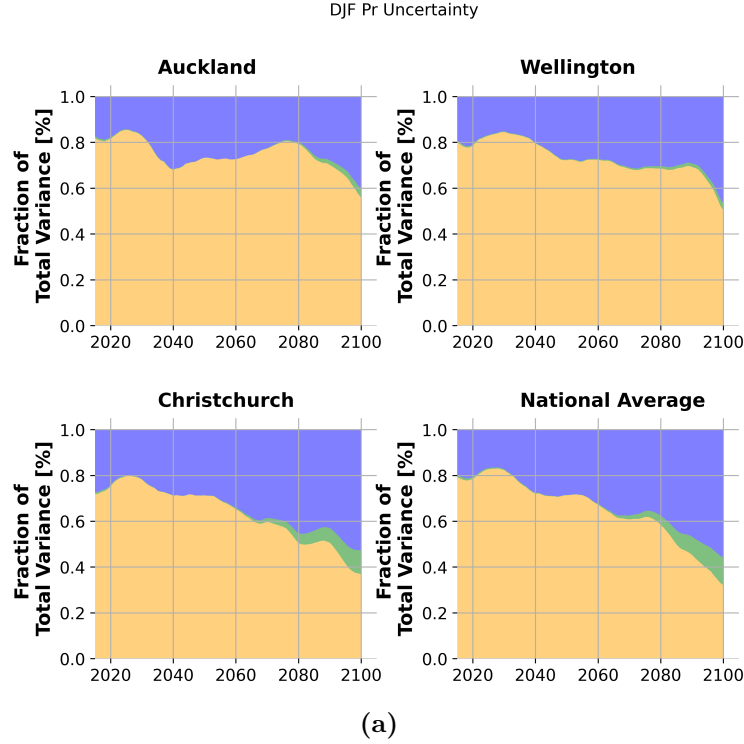
Uncertainty decomposition at fine scales (Figures S29-S34)



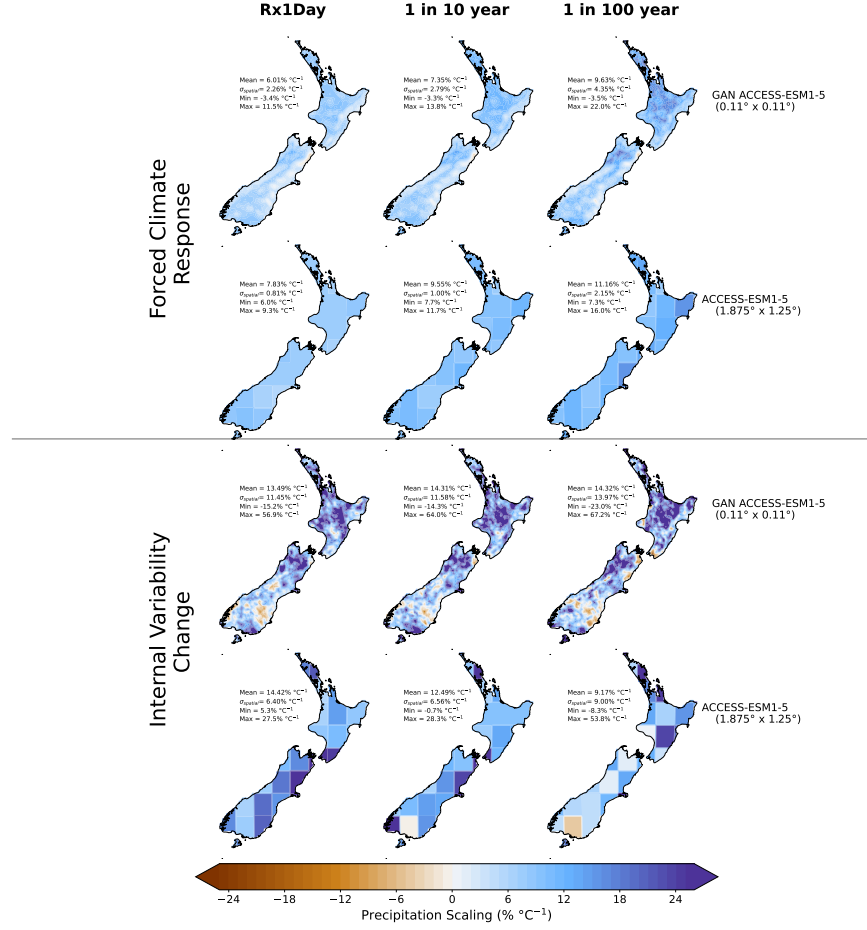
Supplementary Figure S29: Same as in Fig 3. in the main text, but using percentage anomalies instead



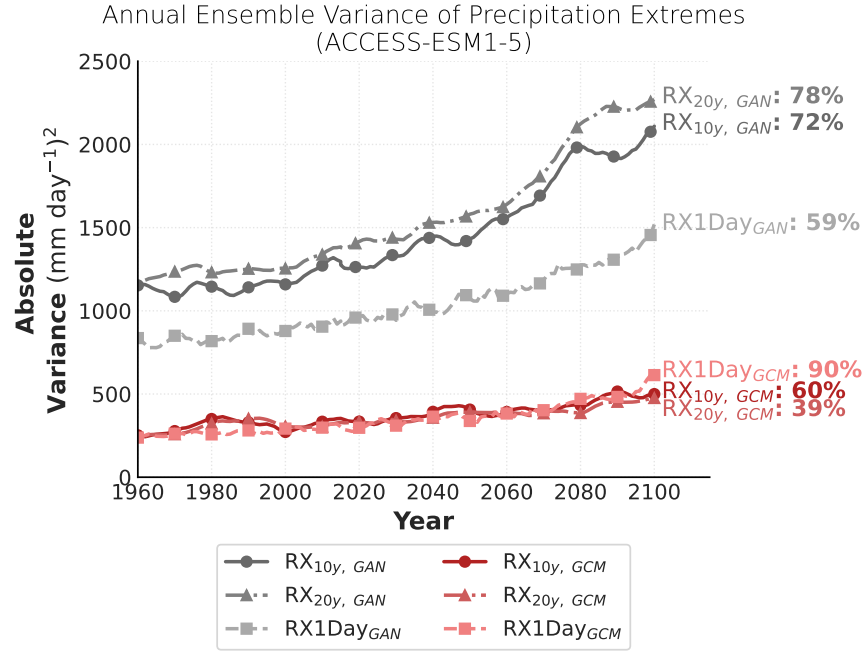
Supplementary Figure S30: Same as in Figure 3 in the main text (top row), but for the first decade of the projection period (2015-2024).



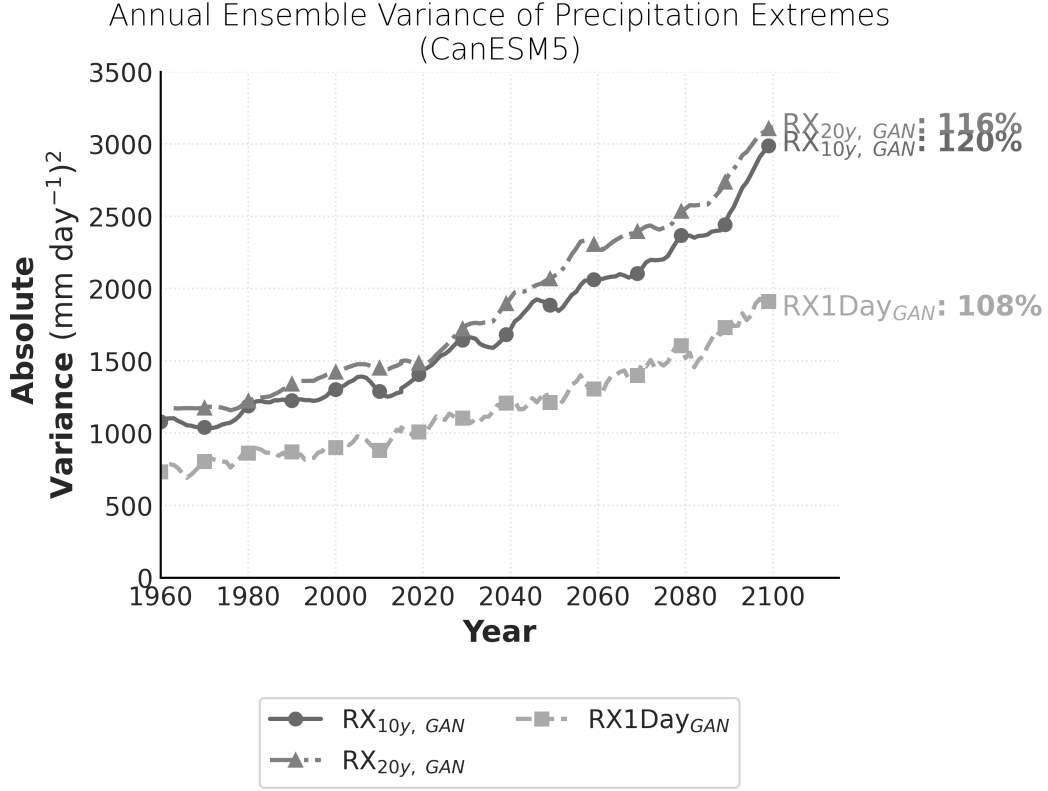
Supplementary Figure S31: Same as Figure S29 (percentage anomalies), but for seasonal mean DJF precipitation (top two rows) and tasmax (bottom two rows) for individual cities across New Zealand. As described in the main text, national average fractional uncertainty is calculated by first averaging trends across all land grid points before applying the fractional decomposition



Supplementary Figure S32: Spatial patterns of projected end-of-century changes in the mean forced climate response and internal variability (ensemble standard deviation) are shown for Rx1day (left), 10% annual exceedance probability (AEP; middle), and 1% AEP events (right) from the ACCESS-ESM1-5 ensemble. The top two rows show changes in the mean forced response, and the bottom two rows show changes in internal variability; in each case, the GAN is shown in the top row and the GCM in the bottom row. Changes are expressed as percentage differences between the historical (1950–2000) and future (2050–2100) periods, scaled by the respective warming rates in the GCM and GAN. Violin plots summarize the spatial distribution across all New Zealand grid points, with annotations indicating the national spatial mean.



Supplementary Figure 33: Temporal evolution of internal variability variance for precipitation extremes (RX10y, RX20y, RX1Day) in the raw vs GAN-Downscaled ACCESS-ESM1-5 ensemble. Here, internal variability variance is computed annually (no decadal smoothing as in Figure 3 of the main manuscript), comparing GAN (grey) and raw GCM (red) precipitation. Variance increases are shown in text for the future period (2080–2100) relative to the historical baseline (1986–2005)



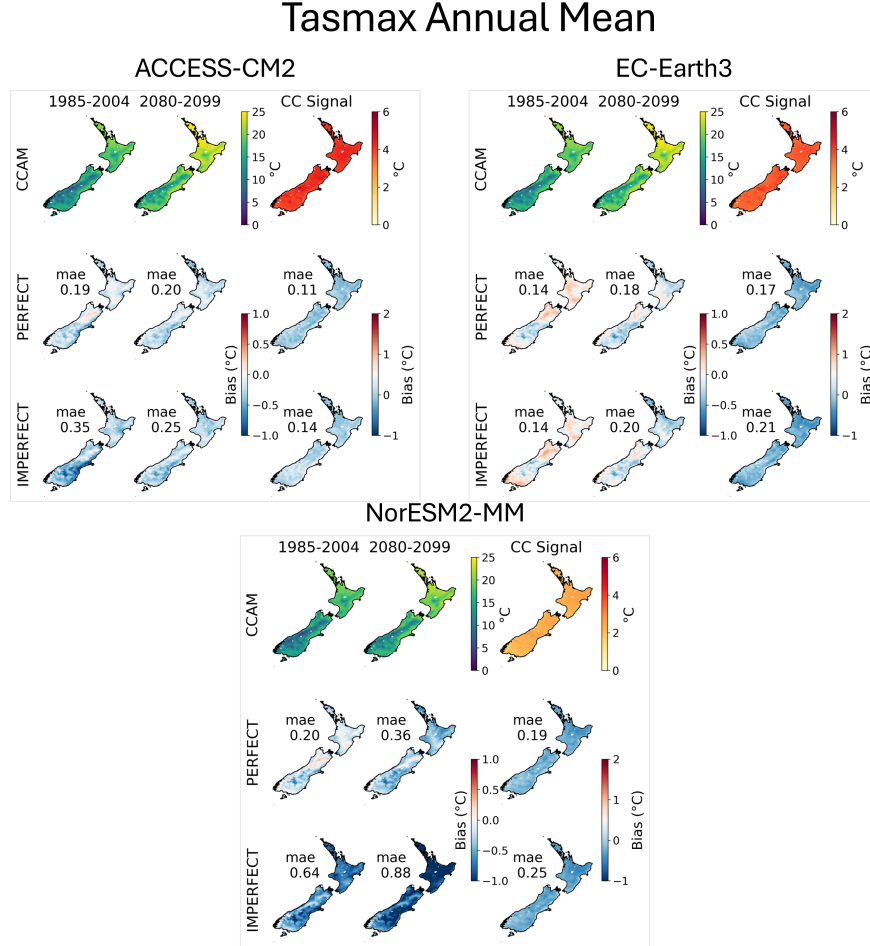
Supplementary Figure 34: Same as in Figure S33 but for CanESM5.

Online Methods

A comparison of the perfect vs imperfect framework (Figures S35-S38)

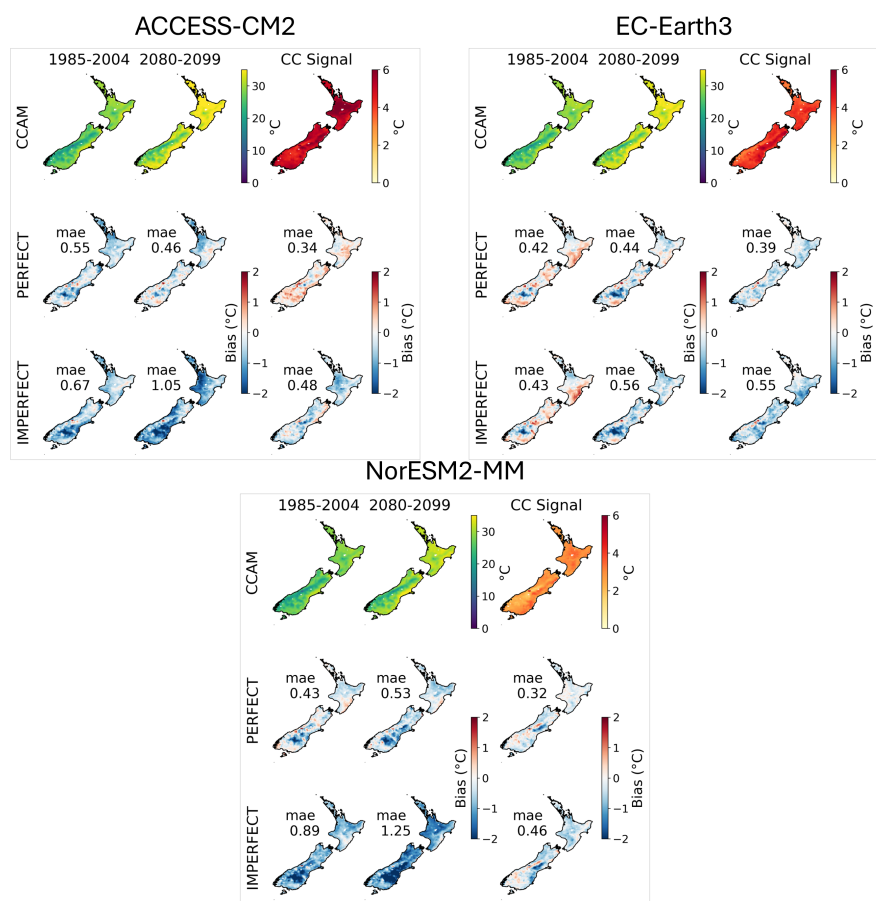
This section provides supplementary material for the Online Methods section. We’ve previously shown that our emulator generates plausible climate change signals consistent with RCMs. Here, we aim to demonstrate how biases change when the emulator is applied to coarsened-RCM fields (perfect application) versus directly to GCM inputs (imperfect application). We also assess if these differences impact historical climatology and the emulator’s ability to produce climate change signals. We evaluate the emulator’s out-of-sample performance using three GCMs: ACCESS-CM2 (used in training), and the unseen EC-Earth3 and NorESM2-MM, for which ground truth RCM simulations are available. Figures S35-S38 illustrate the differences between perfect and imperfect biases relative to ground truth RCMs across these three GCMs. This comparison covers annual maximum surface air temperature (tasmax), precipitation, and extreme indices like RX1Day (maximum 1-day precipitation) and TXx (annual maximum daily maximum temperature). Generally, biases are slightly higher in the imperfect framework, as indicated by the Root Mean Square Error (RMSE) values on the plots. However, signal errors are very similar. This suggests that while subtle mismatches in GCM inputs can affect the circulation (which RCMs are known to modify),

and, consequently, temperature and precipitation climatology, the climate change signal itself is well-preserved. We observe broadly similar bias patterns in the emulator, though they do vary between GCMs.



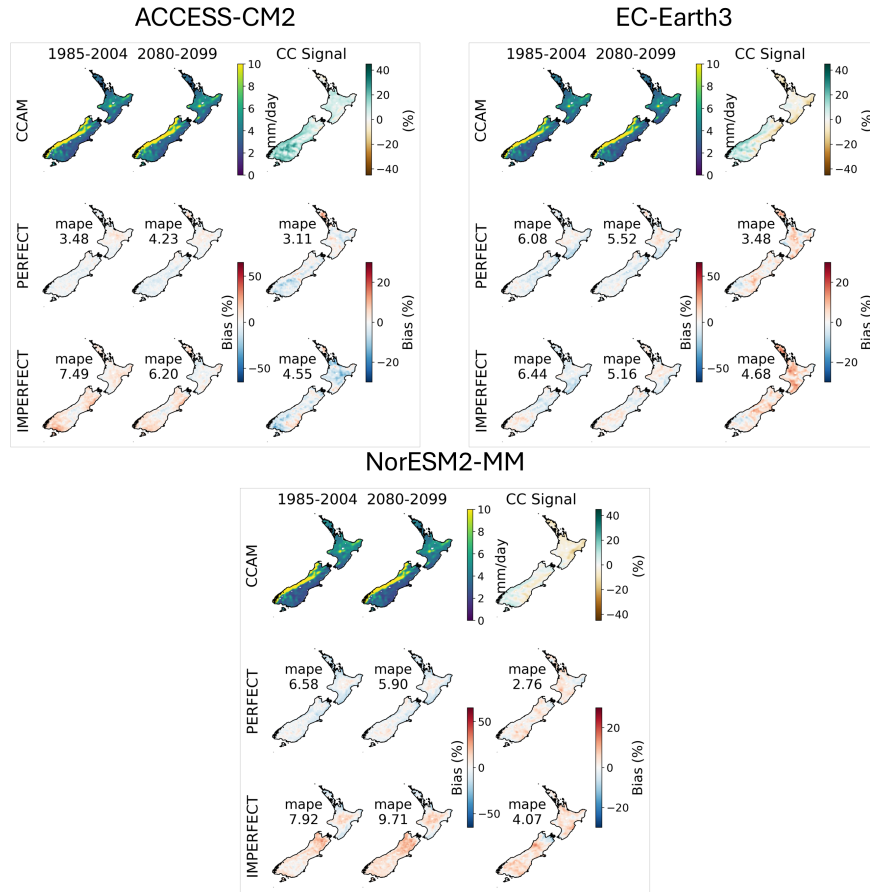
Supplementary Figure S35: Comparison of biases (difference between emulator and ground truth) in the historical climatology (1985–2004), future climatology (2080–2099), and the climate change signal (future minus historical) for three GCMs: ACCESS-CM2, EC-Earth3, and NorESM2-MM. Each plot shows: (top) the RCM ground truth (CCAM), (middle) emulator biases in the perfect framework (relative to CCAM), and (bottom) biases in the imperfect framework (emulator applied directly to GCM fields). The RMSE is shown in text in each of the subplots.

Tasmax TXx



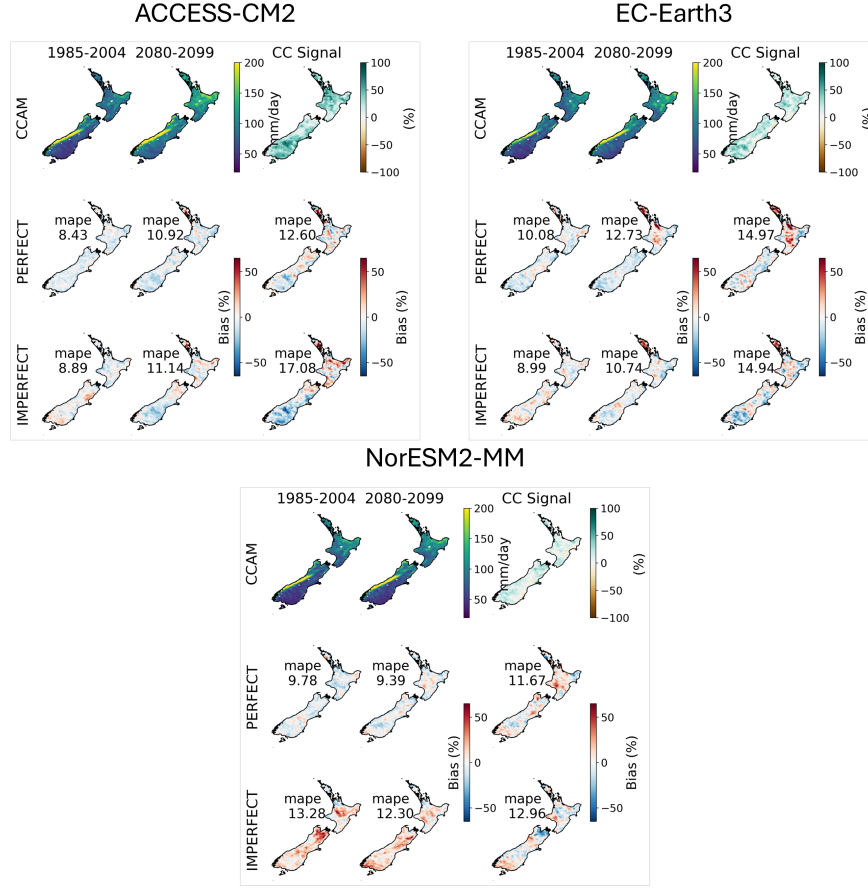
Supplementary Figure S36: Same as Figure S35 but for TXx (hottest day of the year).

Pr Annual Mean



Supplementary Figure S37: Same as Figure S35, but for annual mean precipitation. Note, the climate change signal is computed as a % difference for precipitation.

Pr rx1d



Supplementary Figure S38: Same as Figure S37, but for RX1Day (wettest-day-of-the-year).

Further Sensitivity Testing of the Emulator (Figures S39-S40)

To better understand how the selection and quantity of GCM/RCM pairs used for training impact emulator performance, we conducted several sensitivity experiments. We used a simpler emulator (1.5 million parameters vs. 3.5 million in the main study) to enable faster training, with these sensitivity tests playing a key role in selecting the final model architecture and offering a more efficient alternative to full-scale testing.

These experiments aimed to:

- Confirm that our main findings were robust and were not due to a specific training setup or random chance.
- Determine whether training on multiple GCMs yields better performance than training on individual GCMs.

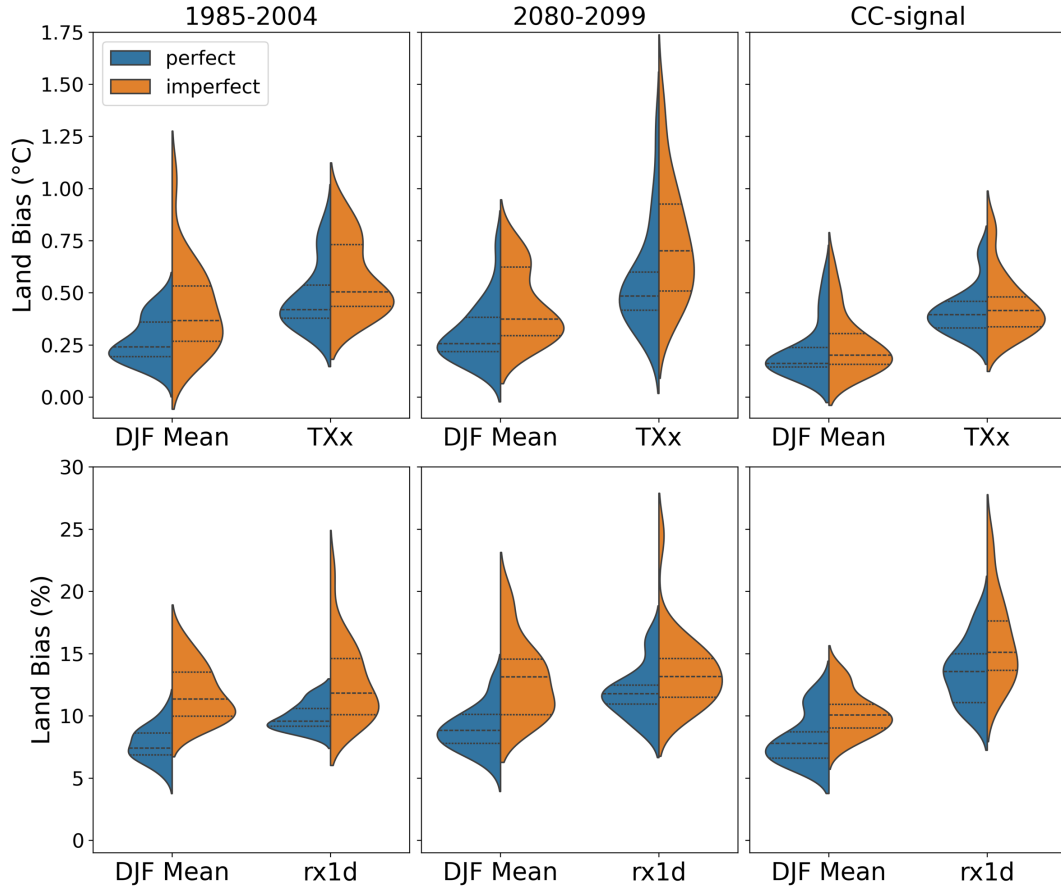
Specifically, we trained emulators on:

- Single GCMs (ACCESS-CM2, EC-Earth3, or NorESM2-MM individually) – 3 experiments for each variable.
- Two pairs of GCM (for example, ACCESS-CM2 and NorESM2-MM) – 3 experiments for each variable to cover all unique permutations of the GCM training pairs.
- Five or six GCMs – 1 experiment.

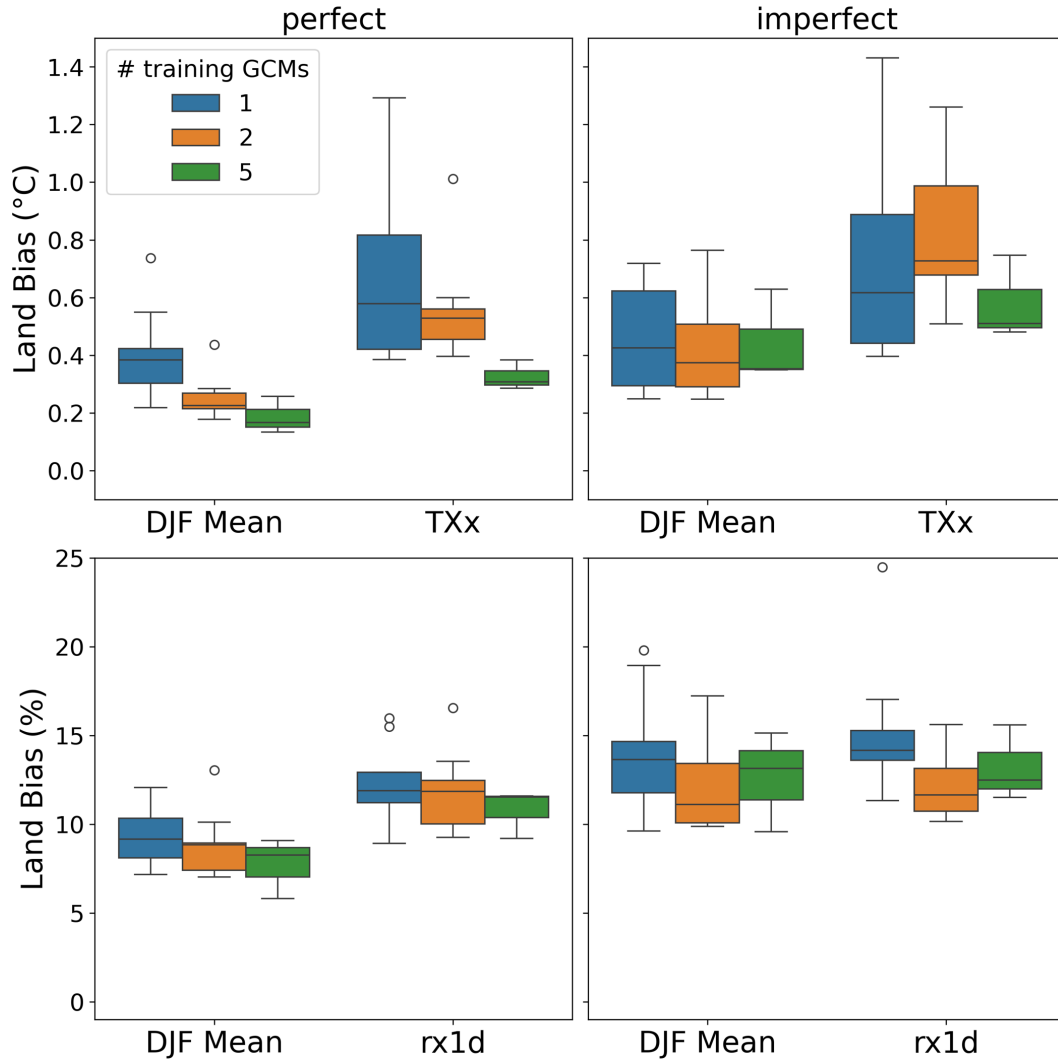
All emulators were trained for 200 epochs. The approximate training times using two A100 GPUs were: 15 hours for a single GCM, 30 hours for two GCM, and 75 hours for five GCMs. For multi-GCM training, each batch was stratified to include an equal number of training instances from each GCM.

For each experiment and variable, performance was assessed in-sample (on the training GCMs) and out-of-sample (on unseen GCMs), using both the perfect and imperfect frameworks. The training procedure for both precipitation and temperature followed the same approach described for the temperature model in the main manuscript, consistent with Doury et al. (2022).

To assess the differences between perfect and imperfect training performance, we summarize their differences across all experiments, spanning various GCMs, training setups, and metrics. Each point in the violin plots represents one emulator evaluated on three GCMs in both frameworks. Overall, imperfect performance is only slightly worse than perfect, with minimal differences for signal metrics, indicating that the emulator largely preserves the climate change signal (Figure S39). As shown in Figure S40, training on five GCMs did not improve future performance (2080–2099 climatology) compared to training in a single GCM, in the perfect or imperfect framework.

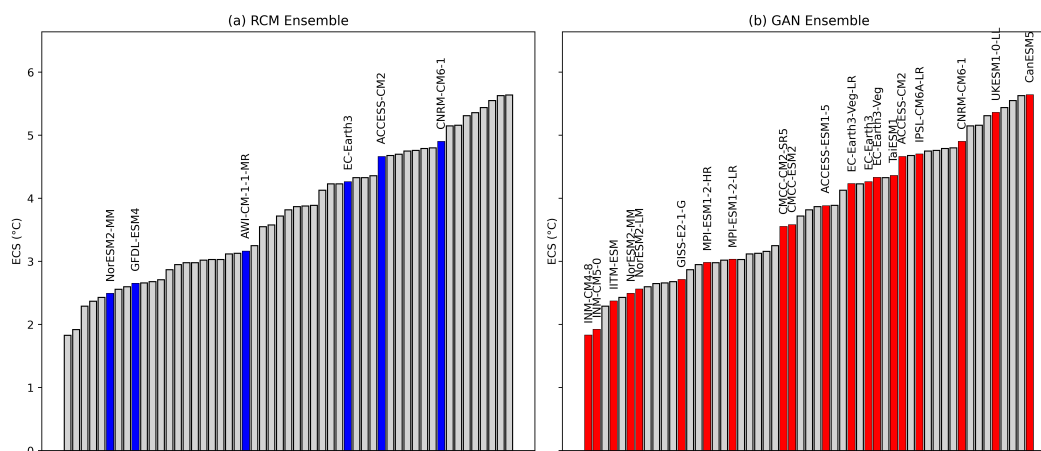


Supplementary Figure S39: Area-averaged RMSE for temperature (top row) and MAPE for precipitation (bottom row) for the historical period (first column), future period (second column), and climate change signal (third column). Results are shown for both the perfect (blue) and imperfect (orange) frameworks. The spread in the violin plots reflects performance across different GCMs and training setups (e.g., training on EC-Earth3 vs. ACCESS-CM2).



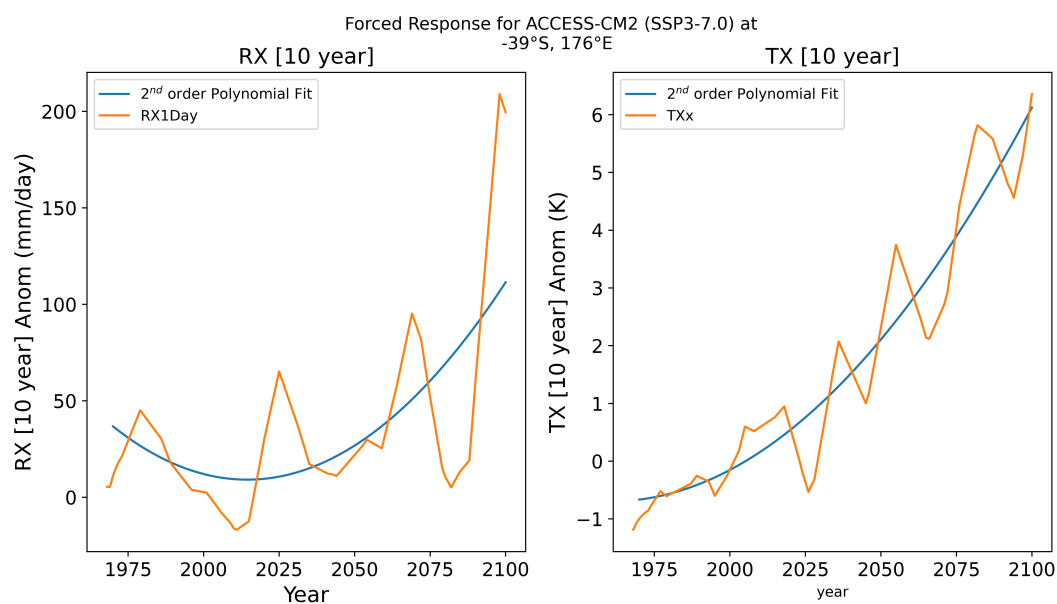
Supplementary Figure S40: Area-averaged RMSE for temperature (top row) and MAPE for precipitation (bottom row) over the future period (2080–2099), comparing emulator performance when trained on 1, 2, or 5 GCMs. The spread reflects variation across GCMs and experimental setups.

Selection of GCMs (Figure 41)

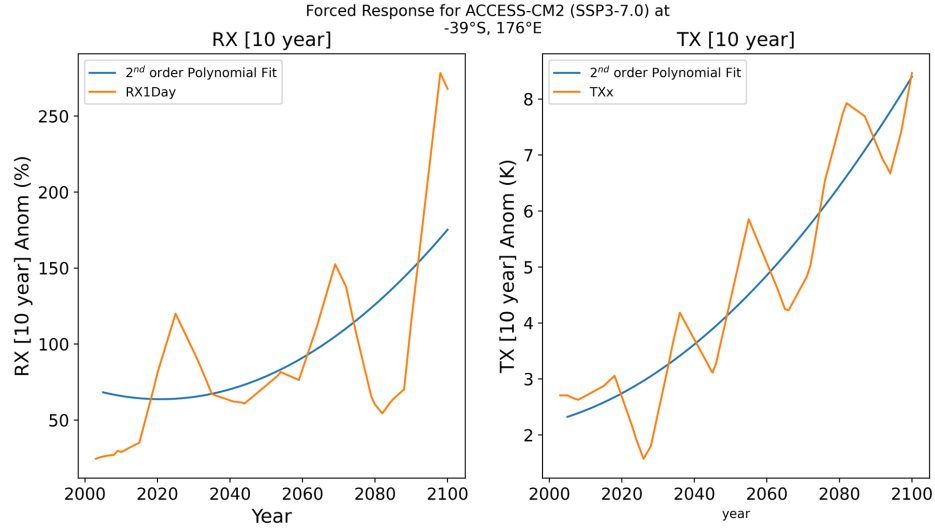


Supplementary Figure S41: Selected GCMs for the RCM ensemble (a) and the emulator (GAN) ensemble (b), shown with their equilibrium climate sensitivity, based on the 53 CMIP6 GCMs.

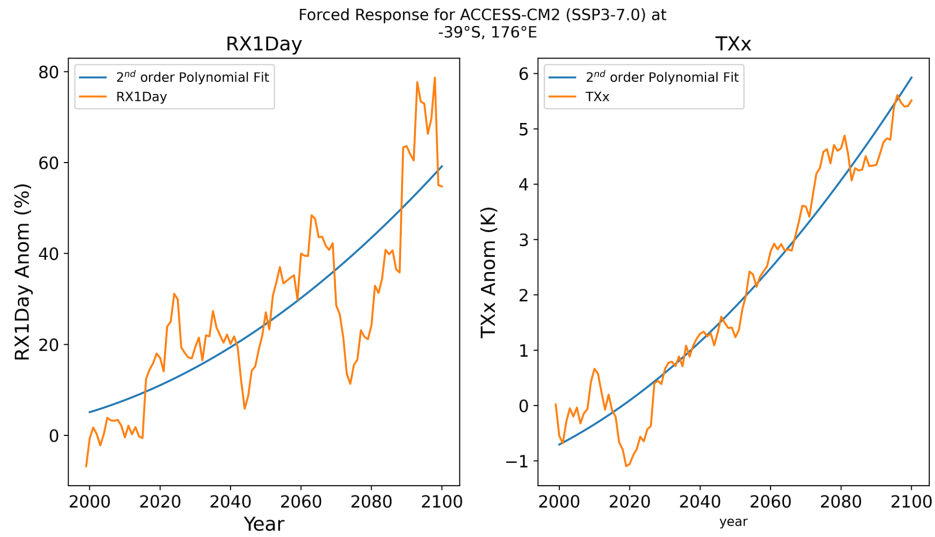
Methods concerning uncertainty decomposition (Figures S42-S43)



Supplementary Figure S42: Polynomial fits (2nd order) for RX [10 year] and TX [10 year].



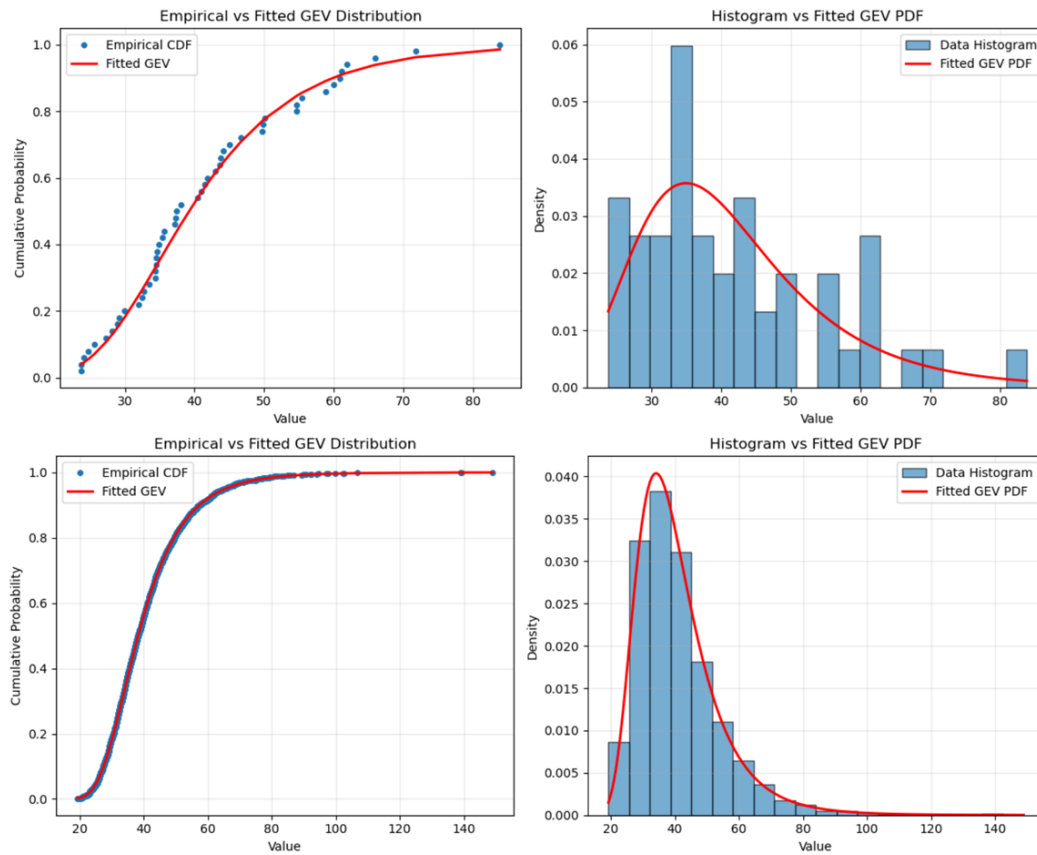
(a)



(b)

Supplementary Figure S43: Same as S42 but using percentage anomalies for precipitation (Figure S29). RX10yr and TX10yr appear smoother due to the use of rolling decadal maxima followed by a rolling average. RX1Day is also smoothed using a 10-year rolling average in both the top and bottom rows.

Methods GEV Distribution fits (Figures S44)



Supplementary Figure S44: Example GEV fits at a single grid point. Top panels: fit using one ensemble member over 50 years ($n=50$ samples), showing (left) empirical versus fitted cumulative distribution and (right) fitted distribution with empirical histogram. Bottom panels: fit using all 40 ensemble members pooled together ($n=2000$ samples). The fits are shown for RX1day (annual maxima of precipitation; units of mm/day).