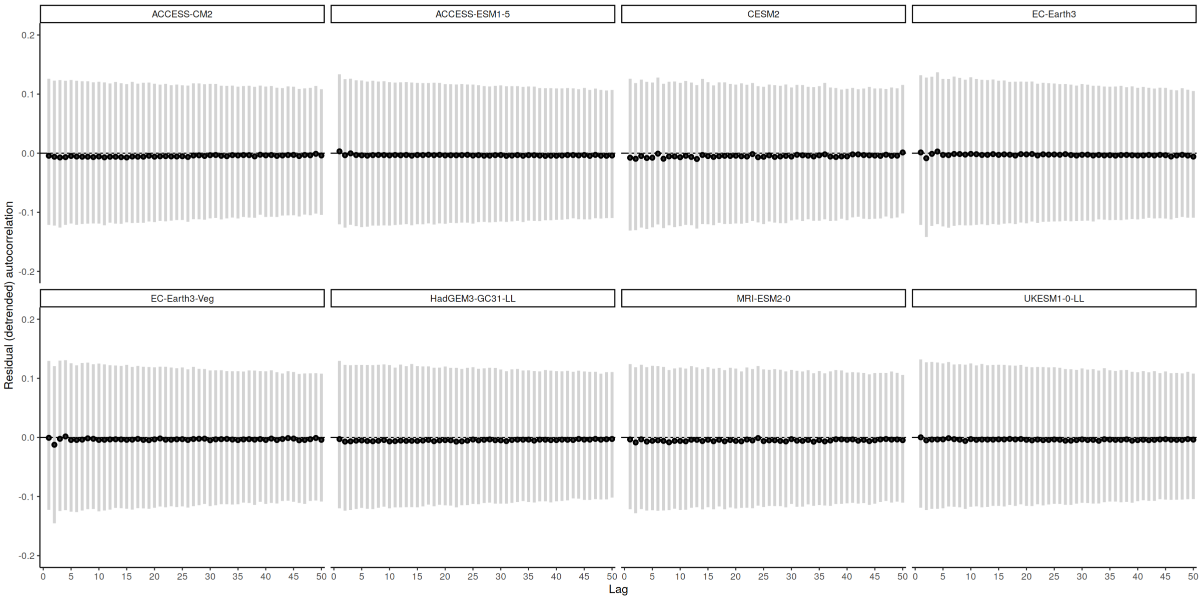
Supplementary Information

S1 Conditions for validity of method



SI Figure S1: Temporal autocorrelation of detrended residuals of Rx1d for lags from 1 to 50, for all eight CMIP6 models used. Detrending was done by subtracting the loess-filtered ensemble mean timeseries per gridcell.

As mentioned in the main text, the temporal independence of Rx1d values is a prerequisite for temporal independence of record breaking probabilities and validity of equations (1), (2) and (3). Fig. S1 shows that autocorrelations between Rx1d values at both high and low frequencies (short and long lags) are 0. The autocorrelation of the non-detrended values is positive and gradually decaying because there is a long term trend. This trend affects the record breaking probabilities over time, but does not violate temporal independence of Rx1d values.

S2 Spatial pooling methods for observational/reanalysis GEV fits

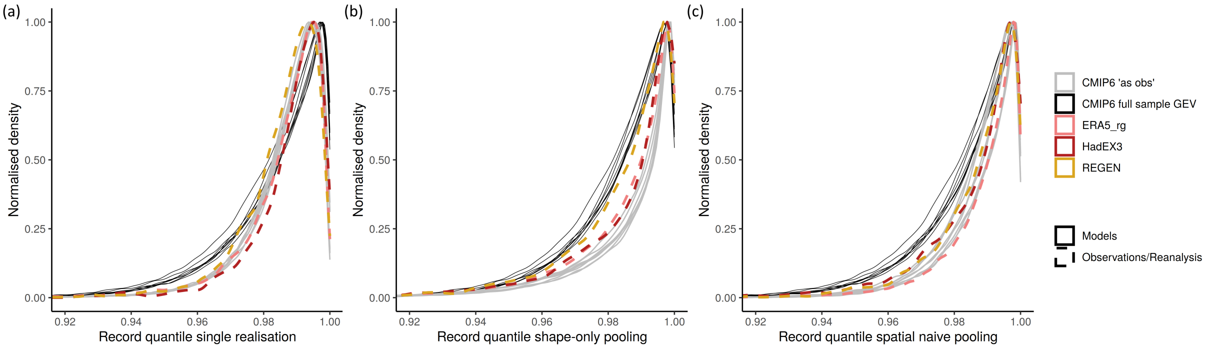
As mentioned in Sect. 4.2, we tested two different spatial pooling methods to improve the GEV fits to the short observational and reanalysis timeseries. Below we outline the effects of shape-only and naive spatial pooling.

For shape-only spatial pooling, strength is borrowed from neighbouring gridcells to improve the estimate of ξ only [88]. In practice, we pool Rx1d data within a spatial window of 5×5 gridcells to fit a GEV, determining unique μ and σ values for each gridcell in the window, but allowing only one universal ξ value for all gridcells in the window. This is achieved by defining spatial covariates for μ and σ that, as it were, 'turn on' the individual gridcells in the fitting procedure, while ξ has no spatial covariate. The resulting GEV parameters are assigned to the middle gridcell of the window. Naive spatial pooling, on the other hand, implies we simply fit one GEV to all the Rx1d data pooled in a spatial window, and assign the GEV parameters to the middle gridcell, as used by e.g. [89]. There is thus a degree of spatial smoothing for all three GEV parameters in this case, whereas the shape-only method aims at smoothing only ξ. [36] compares several GEV estimation methods including the two we test and finds that both are equally effective at improving accuracy of Rx1d return level estimates for return periods longer than the sample size.

Supplementary Fig. S2 shows the effect of the pooling schemes on the distribution of quantiles. Supplementary Fig. S2a-c show the normalised density of record quantiles, where the black lines show the model means determined based on full-ensemble GEV distributions, i.e. the "target", and grey lines show the model results when members are treated as HadEX3 observations. The coloured dashed lines correspond to record quantiles in the observational and reanalysis data. Supplementary Fig. S2a shows the quantiles obtained for GEV distributions fitted ‘as observations’ to timeseries without spatial pooling. The peak lies too low, and also low quantiles are underestimated: the record quantile distribution is too narrow. Below we discuss the effects of pooling and how the impact the record quantile distribution.

Shape-only pooling leads to smoother patterns and smaller magnitudes of ξ, which makes them more similar to the full-ensemble distribution’s ξ, see Supplementary Fig. S3b. However, also σ responds strongly to this pooling scheme, since the setup aims to fit a σ and μ that are specific to single gridcell data. One might say that shape-only pooling partly decouples the fit of μ and σ from the fit of ξ, where the μ and σ are fit with only part of the information (one gridcell). This leads to compensation effects in primarily the σ values, which strongly affect the estimated record quantiles. In part of the gridcells, the compensation leads to decreases in σ relative to the single-gridcell fits, which are associated with strong increases in quantiles estimated – the smaller the initial quantile, the stronger the increase. In another subset of gridcells, the compensation leads to increases in σ, which leads to decreasing quantile values, especially for already lower quantiles. The combination of these two σ-related quantile changes leads to clustering of most of the quantiles at very high values, and some being moved to very low values. Therefore we see a certain ”pulling apart” in the quantile distribution which leaves a gap in the middle ranges where quantiles between 0.95 and 0.99 should be found, Supplementary Fig. S2b. This effect is seen for the observational and reanalysis record quantiles as well. Most of the quantiles are overestimated in this approach, leading to a strong underestimation of record breaking probabilities. We tried different pooling window sizes of 3 × 3 and 5 × 5 on the HadEX3 grid, and see that the 5 × 5 window size results in smoother shape parameters, more similar to those of the full-ensemble fit, however, the discrepancy in the quantile distributions increases with increasing window size.

In the naive pooling setting, a coherent scale and shape is fit to the pooled sample, leading to less local compensation effects of the scale parameter: we see much smaller changes relative to the single-gridcell fit, whereas the smoothing of the shape parameter is almost as effective as in the shape-only pooling setting. From Supplementary Fig. S2a to c we see improvement in the location of the peak and and higher quantiles brought about by spatial pooling, and minor improvement for the lower quantiles as well. Between quantiles 0.98 and 0.94 the offset remains however considerable, explaining the biases in the CCP values, shown below in the validation section. Also for the observations and reanalysis (coloured dashed lines) we see improvement in the peak location and a shape more similar to the black full-ensemble target lines. The comparison here is imperfect since we do not know the true quantile distribution of the observational and reanalysis record quantiles – the distribution can differ from the model distributions due to model and observational errors, and differences in coverage.



SI Figure S2: Normalised density plots of the historical record quantiles determined using different GEV fitting methods. In all plots, solid grey lines show the ‘true’ record quantiles determined from full-ensemble GEV fits, and solid black lines show the model-as-observations comparisons, where GEV distributions were fitted ‘as observations’ to the simulation data. Dashed coloured lines correspond to the observational and reanalysis datasets. (a) shows the results for GEV distributions fitted to single gridcell data, (b) shows the results for GEV distributions fitted using shape-only spatial pooling, and (c) shows the results for GEV distributions fitted using naive spatial pooling. See also Sect. 4.2.

We perform a few additional tests to confirm the seemingly better performance of naive spatial pooling. As the aim of our study is to estimate record breaking probabilities, we assess the skill of the probability estimate for each spatial pooling method using ranked probability skill scores (RPSS). These are determined by estimating cumulative record breaking probabilities treating model members as observations, and comparing the estimate to the actual future evolution of the model member; see Sect. 4.4 on validation for a full explanation. The RPSS represents the improvement of the estimation method in question relative to a benchmark. Table 3 in Sect. 4.4 shows RPSS values of the record breaking probability estimates corresponding to different GEV-methods used to determine the record quantiles. The first row refers to GEV distributions fitted to the 1950–2015 timeseries of single gridcells, where no optimisation has been done to reduce biases in the GEV fit due to the small sample size. We see a minor skill improvement of 5% over the benchmark. For shape-only spatial pooling, the improvement over the benchmark is in fact negative, i.e. the probability estimates are worse. We tested different spatial window sizes for the shape-only pooling, and see that 5 × 5 on the HadEX3 grid is the minimum size to achieve clearly smoother patterns of ξ. For larger windows, the results are similar but the prediction gets slightly worse as window size is increased. GEV fits using naive spatial pooling lead to more than twice as much skill gain as the single gridcell GEV. We tested larger window sizes and see that a window size of 3×3 on the HadEX3 grid is better than larger windows. The probability estimates based on full ensemble GEV distributions are 30% more accurate than the benchmark.

Lastly, we confirm that not just bulk properties, but also the spatial pattern of ξ and the probability metrics improve most when naive spatial pooling is used. Supplementary Fig. S3 shows the ξ (a-d), record quantile (e-g), state likelihood (i-l) and CCP (m-p) maps of one single member of ACCESS-ESM1-5 – the largest ensemble in our model selection. We show metrics based on single gridcell, shape-only pooled, naively pooled and full-ensemble GEV fits for visual comparison, and Supplementary table S4 provides the multi-model spatial correlations of these quantities obtained from the different ‘as observations’ GEV fits to those obtained using the full-ensemble GEV.

The maps and spatial correlations show a much stronger agreement of the naive pooling GEV based results with the full-ensemble GEV based results. For shape-only pooling, strong patterning appears that seems influenced by the climatology and leads to artificial regions of low state likelihood and high future record breaking probability.

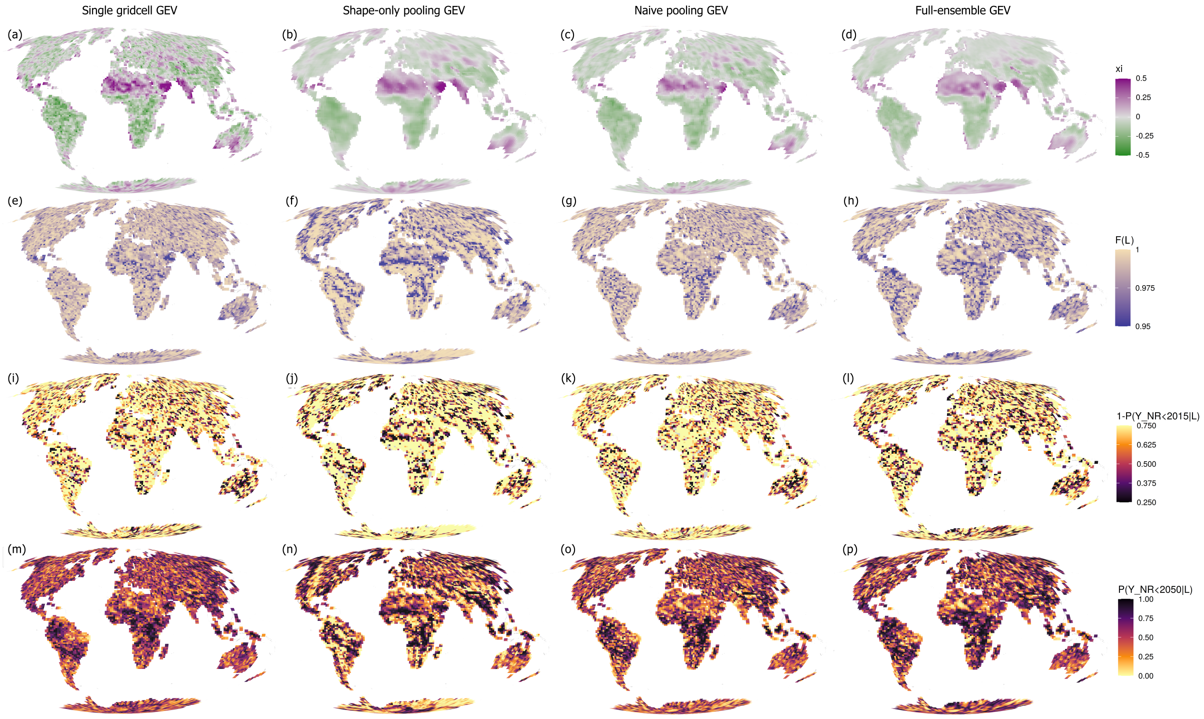
Given the clearly better performance (for our purposes) of naive spatial pooling in all tests performed, we employ naive spatial pooling for the observational/reanalysis GEV fits in our analysis.

SI Table S4: Spatial correlation coefficients of the variables listed in the top row; correlations of the result obtained using the GEV fitting method listed in the first column with the result obtained using the full-ensemble GEV. SL refers to state likelihood: $1 - P(T < Y_{\mathrm{NR}} \leq 2015|L)$ and CCP to $P(Y_{\mathrm{NR}} \leq 2050|L, T = 2015)$

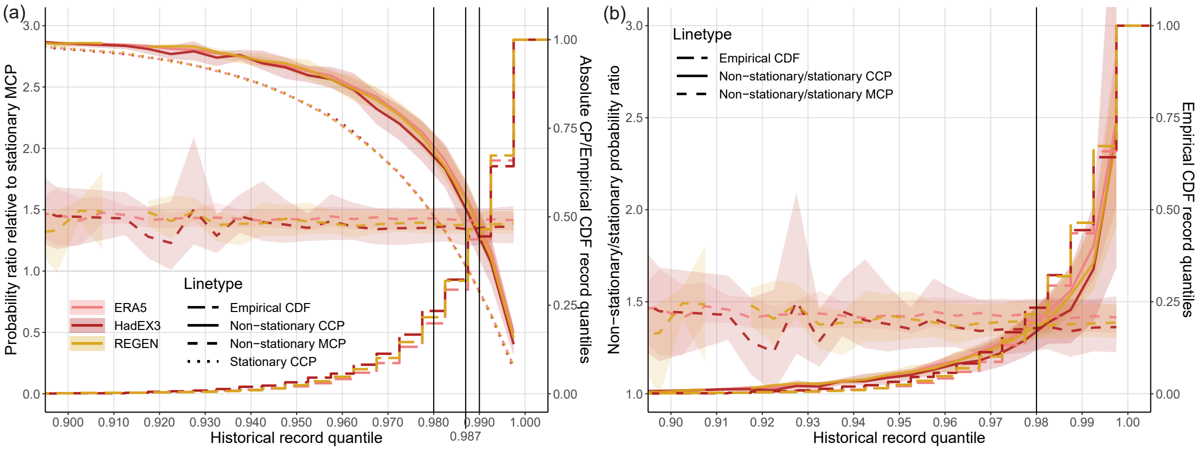
GEV fitting method	ξ	μ	σ	$F(L)$	SL	CCP
Single gridcell ‘as obs’ GEV	0.70	0.91	0.97	0.45	0.69	0.61
Shape-only spatial pooling GEV, 5 × 5 window	0.78	0.87	0.91	0.53	0.67	0.60
Naive spatial pooling GEV, 3 × 3 window	0.74	0.94	0.98	0.75	0.84	0.80

S3 Climate model results - comparison of GEV fitting methods

In Fig. S3 the sensitivity to the GEV fitting method is shown. In the first three columns the single model member is treated as an observational record. The single gridcell GEV leads to too much spatial variability, yet too little spread in the quantiles. The strong increase in spread but artificical pattern resulting from shape-only pooling (see Sect. 4) is very apparent in the second column. The moderate changes due to naive spatial pooling (third column) clearly do not reproduce the ‘true’ full-ensemble patterns (right column), but are closest in terms of magnitude, spread, and spatial pattern. The full-ensemble patterns, showing more distinct (random) regions of low record quantiles and state likelihood, and high future record breaking probabilities are indicative of what any true observational pattern could look like if we were able to determine the true underlying distribution. The pattern of record quantiles is random, and elevated or damped record breaking probabilities due to natural variability can occur anywhere.



SI Figure S3: Patterns of GEV shape parameter ξ (a-d), historical record quantile in the year of record occurrence (e-h), 2015 state likelihood (j-l), and PCC in 2050 (m-p). All patterns are from a single member randomly selected from the ACCESS-ESM1-5 ensemble. Shape parameter and historical record quantiles are computed based on GEV fits following the method indicated above the columns. GEV for future quantile evolution are in all cases based on the full ensemble, as in the main text, see 4 for details.



SI Figure S4: As main Fig. 3d-e, but with empirical (non-parametric) CDF per quantile level bin (on right y-axis), showing fraction of gridcells subject to the corresponding probability ratios

S4 Supplementary figure to main Fig. 3d-e

S5 Derivation quantile level of records

Intuitively, the expected quantile level of the maximum M_j of an i.i.d. sample $\{X_1, \dots, X_j\}$ of length j should correspond to $1 - \frac{1}{j+1}$. This follows from the fact that the marginal record breaking probability at timestep j is $\frac{1}{j}$. For example, for timestep $j = 3$, the average record breaking rate is $\frac{1}{3}$, meaning that the quantile level of the current record at time $j = 2$ is $1 - \frac{1}{3}$. This means that the record *set* at time $j = 1$ has an average quantile level of 0.5, hence, $1 - \frac{1}{j+1}$.

Formally, we derive this result as follows. We are looking for the quantile level of the maximum

M_j, i.e. $F_X(M_j)$, where F_X is the CDF of each of the i.i.d. data points X_i and M_j is the maximum of $X_1, \ldots, X_j$. In order to find the expected value of $F_X(M_j)$, we need an expression for the PDF of $F_X(M_j)$, which we find by determining the CDF of $F_X(M_j)$ in equation (11) and taking the derivative of that expression in equation (12).

$$\begin{multline} F(F_X(M_j)) = P(F_X(M_j) \leq F_X(m)) = \\ P(M_j \leq m) = P(\max(X_1, \ldots, X_j) \leq m) = P(X_1 \leq m) \cdot \cdots \cdot P(X_j \leq m) = F_X(m)^j \quad (11) \end{multline}$$

In the equation above we use the property that $P(F_X(M_j) \leq F_X(m)) = P(M_j \leq m)$. To find the PDF for $F_X(M_j)$, we take the derivative of the previous expression:

$$f(F_X(M_j)) = jF_X(m)^{j-1} f_X(m) \quad (12)$$

Now we can determine the expected value of $F_X(M_j)$ as follows:

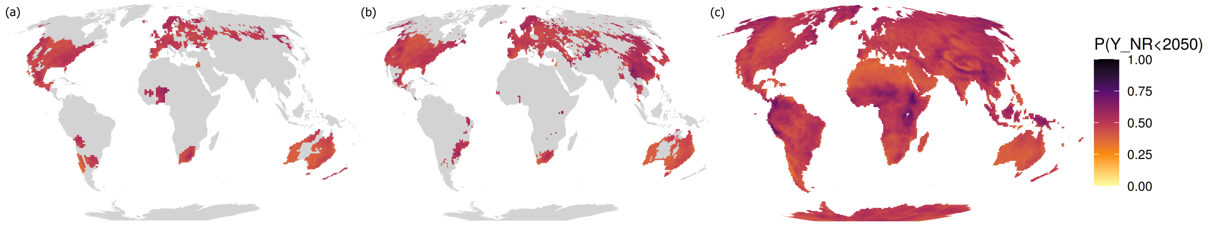
$$\mathbb{E}[F_X(M_j)] = \int F_X(m) f(F_X(M_j)) dm = \int F_X(m) jF_X(m)^{j-1} f_X(m) dm = j \int F_X(x)^j f_X(x) dx \quad (13)$$

We can substitute $u = F_X(x)$ in the previous integral and instead of integrating $f_X(x)dx$ we then integrate over du from 0 to 1, leading to the final result:

$$\mathbb{E}[F_X(M_j)] = j \int_0^1 u^j du = j \cdot \left[\frac{1}{j+1} u^{j+1} \right]_0^1 = \frac{j}{j+1} = 1 - \frac{1}{j+1} \quad (14)$$

Note: the above holds if $\{X_1, ..., X_j\}$ is a stationary, i.i.d. sample. This property manifests in the stationary CCP/MCP ratios in Fig. 3d-e.

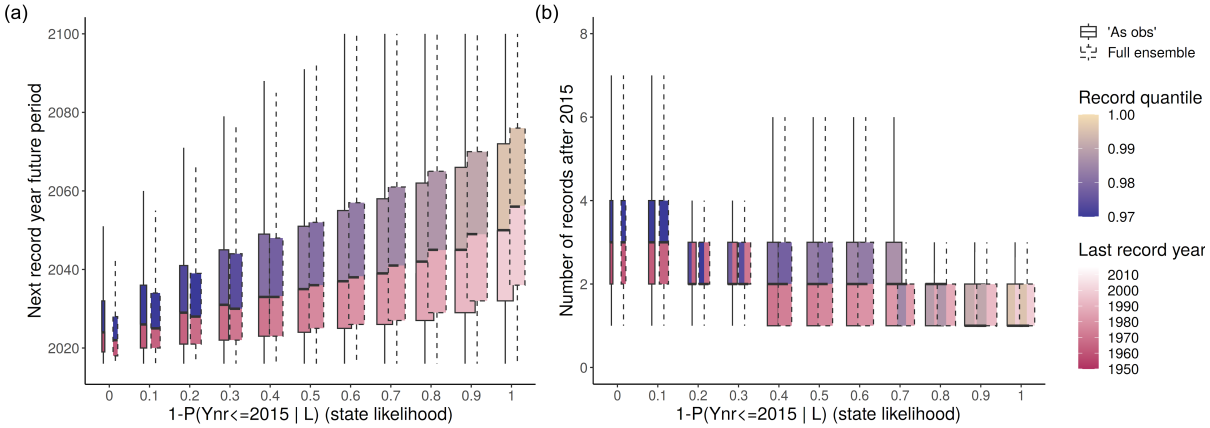
S6 Marginal cumulative record breaking probability patterns



SI Figure S5: Marginal cumulative record breaking probability (MCP) by 2050 as defined in the main text for HadEX3 (a), REGEN (b) and ERA5 (c).

Figure S5 shows the MCP by 2050 for the observational and reanalysis datasets. The MCP includes only the effect of climate change on record breaking probabilities, and is independent of the historical record level to be exceeded. The difference in magnitude between HadEX3 and REGEN on the one hand and ERA5 on the other hand is due to the start year; for HadEX3 and REGEN the cumulative summation starts in 2016, for ERA5 in 2024.

S7 Additional validation state likelihood

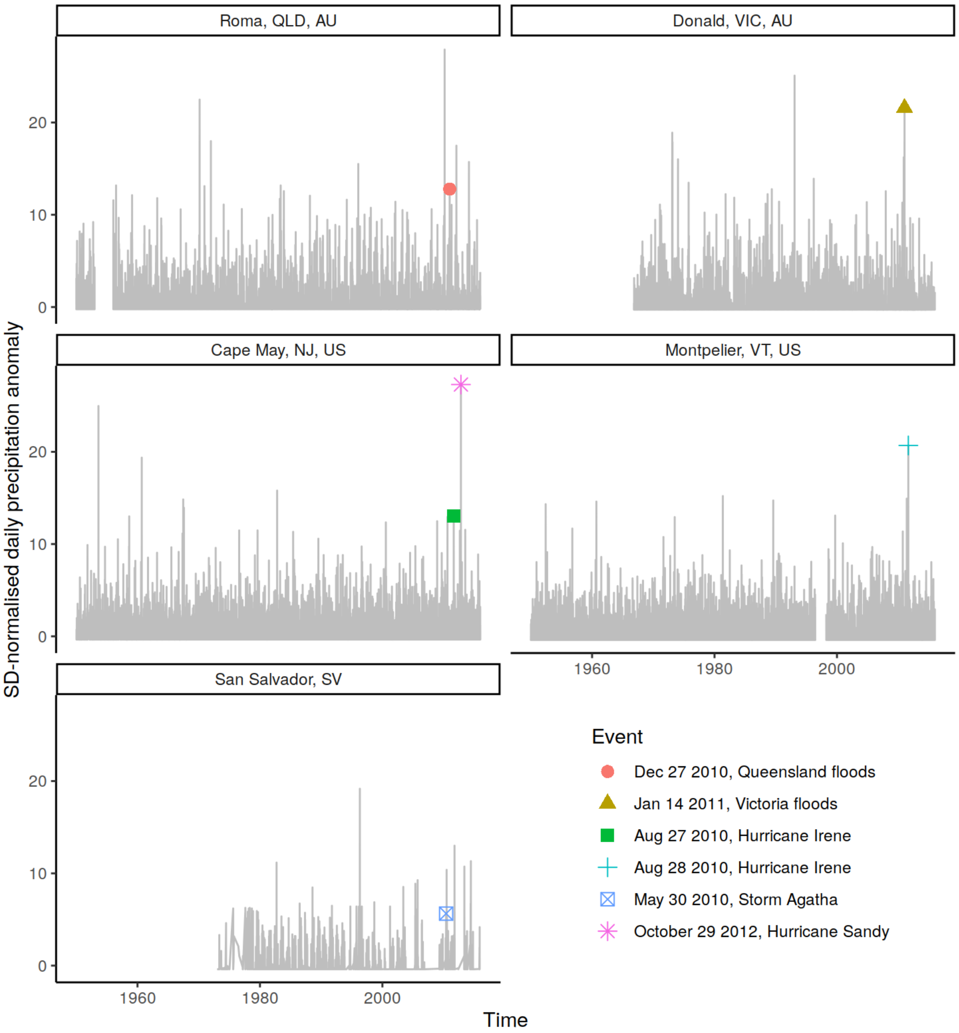


SI Figure S6: Correlation of the state likelihood (binned) in 2015 with the year of next record occurrence (a) and with the number of records in the period 2016–2100 (B), evaluated in all CMIP6 models. The colour shading of the bars show the associated bin-means of record quantile level and record-setting year T, with a clear gradient towards lower record quantile levels (less extreme) and longer-ago years as the state likelihood decreases.

Fig. S6 shows the correlation between state likelihood and indicators of future record breaking in CMIP6 models, in part validating its use as an indicator of disaster potential.

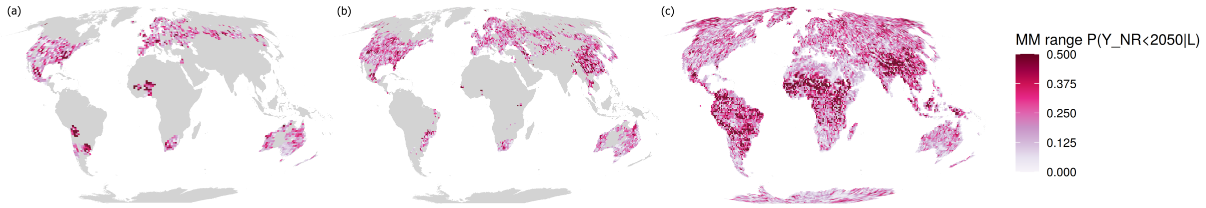
Fig. S7 shows that the selected single station data in the regions discussed in the main text shows a clear signature corresponding to the events we associated with the record breaking. These stations were selected based on their location being in the gridcells of interest; gridcells with 2009 state likelihood ≤ 0.25 and record breaking in the years 2010–2015, and reported in the disaster reports of the events in question [57–60]. Not all stations show the maximum daily event, which is expected as HadEX3 aggregates multiple stations in their Rx1d product, but all stations show daily precipitation values in the uppermost quantiles of the full sample. Storm Agatha (San Salvador) is least well represented, also in other stations in the region that we assessed. The lower data quality in Central America plays a big part (exemplified by the sparsity of the record for San Salvador): this impairs confidence in both the accuracy of records in HadEX3 as well in the verification data itself.

For reference, we added a marker for Hurricane Sandy in the timeseries for Cape May, which evidently led to record-breaking precipitation in New Jersey, where it made landfall. This event of October 2012 is not clearly visible as record breaking cluster in HadEX3, which we hypothesise could be associated with the single landfall of Sandy, as opposed to the repeated ‘bouncing’ landfall of Irene.



SI Figure S7: Daily precipitation observations from single stations from the GHCN-Daily network [90–93] and the Australian Bureau of Meteorology [94, 95]. Events mentioned in the main text are indicated with markers [57–60].

S8 Model uncertainty in probability estimates



SI Figure S8: Intermodel range (difference between maximum and minimum projected value) for CCP 2050. The spread is due to model uncertainty in the temporal evolution of the record quantile, thus a combined measure of differences in climate sensitivity and local patterns of Rx1d changes

Fig. S8 shows the model uncertainty in record breaking projections – the range of the model-mean ensemble of CCP 2050 projections is shown. The overarching pattern is the well-known uncertainty pattern where precipitation and/or Rx1d changes in the tropics and monsoon regions are most uncertain and feature highest intermodel differences. These are also the regions that are projected to see the largest changes in both absolute and relative magnitude of extreme precipitation [52], and are also most vulnerable. The pattern is modified by the local record quantiles; where CCP values are largest due to the combination of low record quantiles and strong climate change, the uncertainties are largest too.