Supplement of Adv. Stat. Clim. Meteorol. Oceanogr., 9, 83–102, 2023
https://doi.org/10.5194/ascmo-9-83-2023-supplement

Supplement of

Quantifying the statistical dependence of mid-latitude heatwave intensity and likelihood on prevalent physical drivers and climate change

Joel Zeder and Erich M. Fischer

Correspondence to: Joel Zeder (joel.zeder@env.ethz.ch)

Additional information on data, methods and results are provided in the supplementary Sect. S1 to S3, further standalone tables and figures are provided in the second part of this supplementary material document.

S1 Data

S1.1 Evaluation of the heatwave representation climate model simulations

Training a statistical model quantifying dependencies between physical observables based on climate model data implicitly requires that heatwaves and the related processes are accurately simulated in these datasets. Wehner et al. (2020) and Thorarinsdottir et al. (2020) evaluated both CESM version 1 and several of the CMIP6 models used in this study with respect to one-day annual maximum temperature (Tx1d) against different gridded observational products, and find that all models (also across model generations) show similar skill in reproducing extreme events. In terms of estimating 20-year return values of Tx1d correctly, CESM1-BGC even performs best among all tested CMIP5 models. This confirms the findings of Sillmann et al. (2013), who identify CESM1-BGC as one of the best-performing models in the representation of climate extremes. Overall, the multi-model CMIP6 ensemble only has minor biases concerning temperature extremes (Kim et al., 2020). CESM1 has further been used in mechanistic studies of heatwave storylines, as in Wehrli et al. (2019) or Gessner et al. (2021). An ensemble size of 83 members is also sufficiently large to reasonably estimate forced changes and characterise internal variability of mean climate (Suarez-Gutierrez et al., 2020) and extremes (Tebaldi et al., 2021). Given the performance evaluation, we assume the climate model data to contain the relevant process information necessary for the analysis.

S1.2 Pre-processing steps

Quantifying how process variables determine the intensity of heat extremes, comparing estimates across data sources and applying the respective statistical models to new datasets requires careful pre-processing of the data due to the inhomogeneous representation of process variables in different datasets (differences in climatological means and variability, diverse soil-layer definitions and spatial resolutions, etc.). A robust global warming signal is detectable in the large-scale thermal expansion of the lower troposphere (Christidis and Stott, 2015), whereas changes in soil moisture are more spatially heterogeneous since they are largely driven by changes in precipitation patterns (Seneviratne et al., 2010; Greve and Seneviratne, 2015), but forced trends in water availability have also been observed and attributed to anthropogenic climate change (Padrón et al., 2020). In order to decorrelate these variables from global mean temperature changes such that they only represent stationary year-to-year internal variability, we remove the long-term climate change-induced trend. Consequently, the dynamical predictor contains no signal of forced circulation change (as in Terray, 2021); thus, the statistical model is also not capable of representing thermodynamic changes in circulation (Vautard et al., 2016). The following section will outline the pre-processing steps, from retrieving the data from the original climate model and reanalysis dataset to detrending and scaling.

The variables daily mean near-surface air temperature (tas in CMIP6 nomenclature / t2m in ERA5 nomenclature) and geopotential height at 500 hPa (zg500 / z) are extracted from the respective simulations in the archive, and a seven-day running mean is calculated. Furthermore, total water content (mrsol / swvl) is integrated over the top soil layers down to roughly 20 cm (depending on the layer definition). Regarding the reanalysis data, global and upper-air variables GMST and Z500 are obtained from ERA5, and land surface variables t2m and SM are obtained from the ERA5-Land land-surface reanalysis product driven by the atmospheric component of ERA5. All variables are then mapped to a common 2.5°-by-2.5° regular grid (regridded elevation maps of location PNW are shown in supplementary Fig. S1), constituting the variables; seven-day running mean air temperature $x_{\mathrm{T}}^{\mathrm{abs}}$ [K] and geopotential height at 500 hPa $x_{\mathrm{Z}}^{\mathrm{abs}}$ [m], and monthly, vertically integrated soil moisture $x_{\mathrm{SM}}^{\mathrm{abs}}$ [$\mathrm{kg\,m^{-2}}$] (samples of the latter two are shown in Fig. 1). There is generally good agreement in the distribution and evolution of $x_{\mathrm{Z}}^{\mathrm{abs}}$ across climate model/reanalysis datasets. However, there are notable differences in the absolute SM amounts $x_{\mathrm{SM}}^{\mathrm{abs}}$ preceding heatwave events (supplementary Fig. S2b).

Global mean surface temperature x_{GMST} [°C] is obtained by smoothing GMST values of the pooled members of individual model/forcing ensembles, such that natural low-frequency variability is removed. Based thereon, Z500 and SM data is detrended at a grid point level, subtracting the respective summer seasonal mean expected at the respective warming level (lines in Fig. 1). In a second step, the detrended values were scaled using the estimated standard deviation of daily Z500 / monthly SM against the seasonal summer average. The resulting detrended and scaled variables $\tilde{\boldsymbol{x}}_{\mathrm{Z}}$ and x_{SM} are now dimensionless (in units of standard deviations) and should represent unforced dynamical and thermodynamical variability. Furthermore, the pre-processing accounts for systematic differences in mean and variability of $x_{\mathrm{Z}}^{\mathrm{abs}}$ and $x_{\mathrm{SM}}^{\mathrm{abs}}$ across models (supplementary Fig S2), which should render the estimates of statistical models derived from different climate model data more comparable.

The following pre-processing steps are applied to the climate model input data globally in order to retrieve the relevant predictor variables $\boldsymbol{x} = (x_{\mathrm{GMST}}, x_{\mathrm{SM}}, \tilde{\boldsymbol{x}}_{\mathrm{Z}}^{\intercal})^{\intercal}$:

1. Retrieve global fields of daily near-surface air temperature and geopotential height at 500 hPa, as well as monthly fields of water content for various soil layers from climate model or reanalysis datasets.

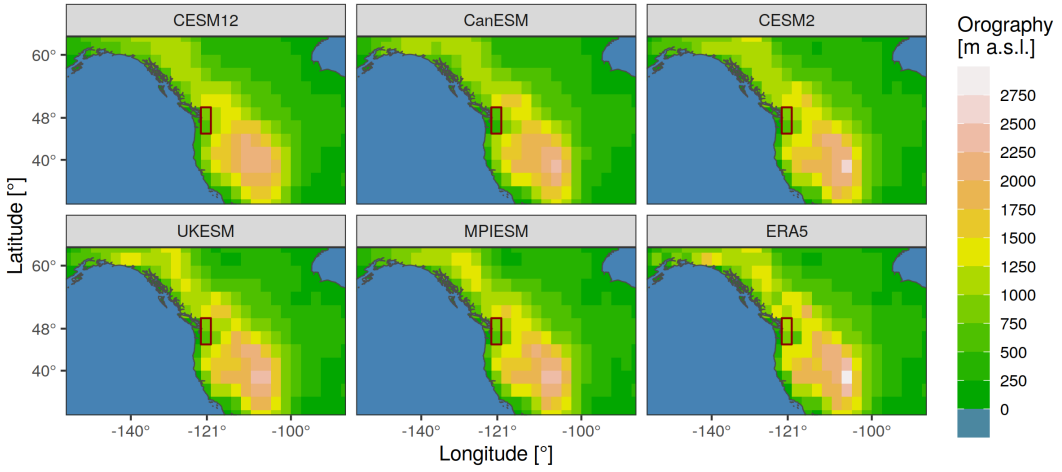

Figure S1. Land orography at location PNW interpolated to 2.5°-by-2.5° resolution in different datasets.

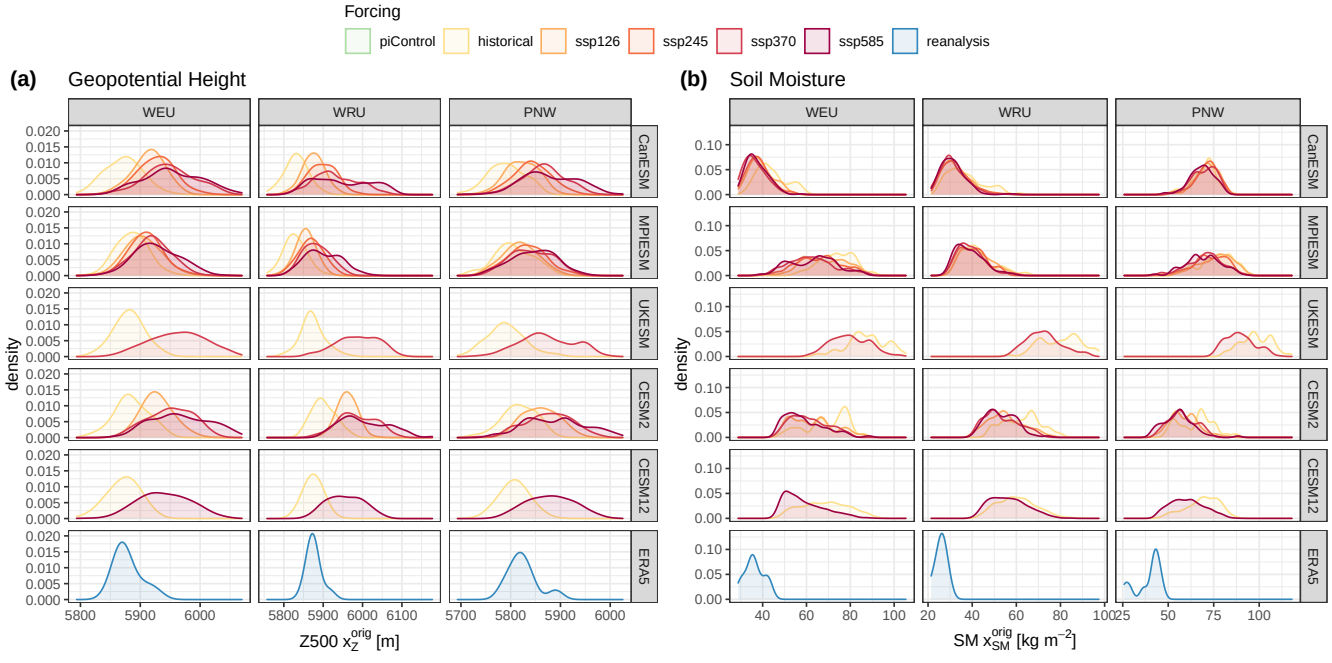

Figure S2. Densities of five-year block maximum heatwave event (a) absolute geopotential height x_Z^{abs} (over the area of interest), and (b) soil moisture x_{SM}^{abs} for locations WEU, WRU and PNW (columns) and different datasets (rows).

2. Integrate total water content from the surface to roughly 20 – 22 cm soil depth. For CanESM and CESM2, the lowest layer considered (10 – 35 cm and 16 – 26 cm) was only partially integrated, assuming a homogeneous distribution of water within the soil layer.

3. Calculate seven-day running means of daily near-surface temperature and geopotential height.

4. Remap all variables to a regular 2.5°-by-2.5° latitude-longitude grid using second-order conservative interpolation (Jones, 1999); seven-day running mean air temperature $x_{\mathrm{T}}^{\mathrm{abs}}$ [K] and geopotential height at 500 hPa $x_{\mathrm{Z}}^{\mathrm{abs}}$ [m], and monthly, vertically integrated soil moisture $x_{\mathrm{SM}}^{\mathrm{abs}}$ [kg m^{-2}]

5. Obtain smooth GMST predictor $x_{\mathrm{GMST}}\,[^\circ\mathrm{C}] = \tilde{x}_{\mathrm{GMST}}(t) - \bar{x}_{\mathrm{GMST}}^{\mathrm{orig,pictl}}$, where $\tilde{x}_{\mathrm{GMST}}(t)$ is the temporally smoothed trend term of (ensemble mean) GMST values based on the local polynomial regression fitting method "loess" (Cleveland et al., 1990; Hastie et al., 2009), and $\bar{x}_{\mathrm{GMST}}^{\mathrm{orig,pictl}}$ is the respective pre-industrial mean smoothed GMST value. The loess span parameter was adjusted so that the resulting annual GMST residuals have approximately the same variance as those under transient forcing. For ERA5, a span parameter $\alpha = 1/3$ instead of $1/5$ is selected to account for the larger variability (as no ensemble mean GMST is provided).

6. Calculate linearly forced trends in seasonal summer mean soil moisture $\bar{x}_{\mathrm{SM}}^{\mathrm{seas}}$ and geopotential height $\bar{x}_{\mathrm{Z}}^{\mathrm{seas}}$ as function of smoothed GMST x_{GMST}

$$\bar{x}_{\mathrm{SM}}^{\mathrm{seas}} = \widehat{\beta}_{0,\mathrm{SM}} + \widehat{\beta}_{1,\mathrm{SM}} x_{\mathrm{GMST}} \tag{S1}$$

$$\bar{x}_{\mathrm{Z}}^{\mathrm{seas}} = \widehat{\beta}_{0,\mathrm{Z}} + \widehat{\beta}_{1,\mathrm{Z}} x_{\mathrm{GMST}} \tag{S2}$$

where the respective slope and intercept coefficients are estimated with a linear least-squares model, regressing seasonal summer (ensemble) mean soil moisture and geopotential height values against smoothed GMST. Supplementary Fig. S3 shows maps of $\widehat{\beta}_{0,\mathrm{Z}}$ and $\widehat{\beta}_{1,\mathrm{Z}}$ for the CESM12 model. Model fits were inspected visually, as in supplementary Fig. S4.

7. Estimate standard deviation of detrended monthly soil moisture s_{SM} and daily geopotential height s_{Z} values:

$$s_{\mathrm{SM}} = \mathrm{sd}(x_{\mathrm{SM}}^{\mathrm{abs}} - \bar{x}_{\mathrm{SM}}^{\mathrm{seas}}) \tag{S3}$$

$$s_{\mathrm{Z}} = \mathrm{sd}(x_{\mathrm{Z}}^{\mathrm{abs}} - \bar{x}_{\mathrm{Z}}^{\mathrm{seas}}) \tag{S4}$$

8. Obtain detrended and scaled variables x_{Z} and x_{SM} by removing the GMST forced trend in summer seasonal means and dividing with the estimated standard deviation:

$$x_{\mathrm{SM}} = \frac{x_{\mathrm{SM}}^{\mathrm{abs}} - \bar{x}_{\mathrm{SM}}^{\mathrm{seas}}}{s_{\mathrm{SM}}} \tag{S5}$$

$$x_{\mathrm{Z}} = \frac{x_{\mathrm{Z}}^{\mathrm{abs}} - \bar{x}_{\mathrm{Z}}^{\mathrm{seas}}}{s_{\mathrm{Z}}} \tag{S6}$$

All pre-processing steps so far have been conducted globally at the grid point level. For the following, the data is further extracted for specific areas of interest listed in Table S1. The sample of heatwave events analysed in the following studies is all five-year maxima in seven-day running mean near-surface temperature $y_{\mathrm{Tx7d}}(t,s,M)$ [°C] at dates t and locations $s \in \{\mathrm{WEU}, \mathrm{WRU}, \mathrm{PNW}\}$ for a distinct climate model/reanalysis dataset M. The respective predictor variables are therefore $\boldsymbol{x}(t,s,M) = (x_{\mathrm{GMST}}(t,M), x_{\mathrm{SM}}(t,s,M), \tilde{\boldsymbol{x}}_{\mathrm{Z}}(t,s,M))^{\intercal}$. The predictor $x_{\mathrm{GMST}}(t,M) \in \mathbb{R}^{+}$ [K] corresponds to the value of smoothed GMST x_{GMST} in the year of BM, $x_{\mathrm{SM}}(t,s,M) \in \mathbb{R}$ [sd] is a weighted average of the previous and current monthly value of detrended and scaled soil moisture (weighted with respect to the date of the extreme within the current month), and a vector $\tilde{\boldsymbol{x}}_{\mathrm{Z}}(t,s,M) \in \mathbb{R}^{512}$ detrended and scaled Z500 values at the ~ 528 grid points in a region of $\pm 40^\circ$ lon and $\pm 20^\circ$ lat from the centre of location s. As the extent of the region determines the number of coefficients to be estimated, there are limitations to its size. Jézéquel et al. (2018) argue for smaller domain sizes, and the longitudinal extent of 80° (roughly 6000 km at 45° N) covers the approximate synoptic scale (roughly 1000 km) of extra-tropical cyclones (Jézéquel et al., 2018). Thus there is prior evidence of the selected region size to suffice for the intentions of this study. In the last step, the values of $\tilde{\boldsymbol{x}}_{\mathrm{Z}}(t,s,M)$ are standardised by subtracting the multi-model temporal mean and dividing by the respective standard deviation in order to make the inference procedure more stable.

Whereas the former steps have been conducted globally at the grid point level, the following ones will be specific for a location s (c.f. Table S1, and, as before, specific for a model M).

9. Five-year BM $y_{\mathrm{Tx7d}}(t,s,M)$ of annual maximum seven-day running mean air temperature are retrieved by first spatially averaging $x_{\mathrm{T}}^{\mathrm{abs}}$ within the area of interest, then selecting maxima of five-year consecutive blocks in pre-industrial control runs, or common annual blocks of five ensembles in transient simulations. The mid-date t of the seven-day running mean period determines the date of the corresponding y_{Tx7d} heatwave event. t_{year} denotes the respective year of the date, t_{month} the month, and t_{day} the day-of-month index.

10. The predictor value $x_{\mathrm{GMST}}(t_{\mathrm{year}}, M)$ corresponds to the smoothed GMST value in the year of the respective BM $y_{\mathrm{Tx7d}}(t,s,M)$.

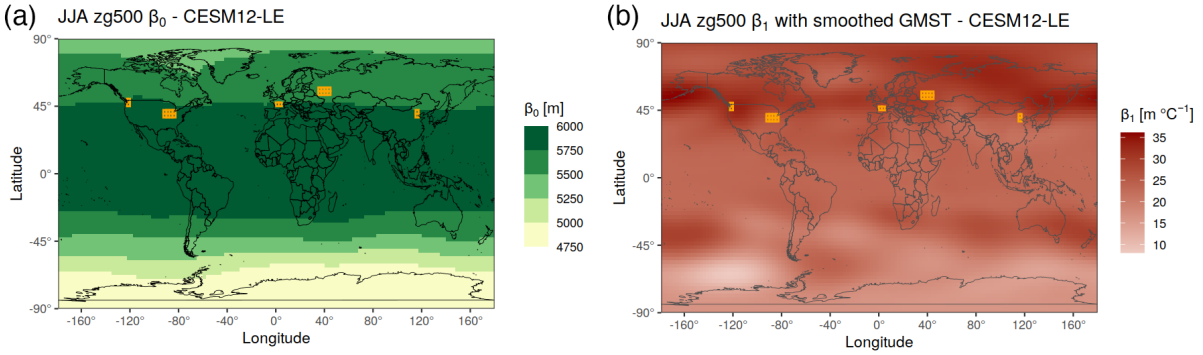

Figure S3. Estimated least-squares coefficients $\widehat{\beta}_{0,Z}$ (a) and $\widehat{\beta}_{1,Z}$ (b) of CESM12 JJA mean geopotential height regressed against smoothed GMST.

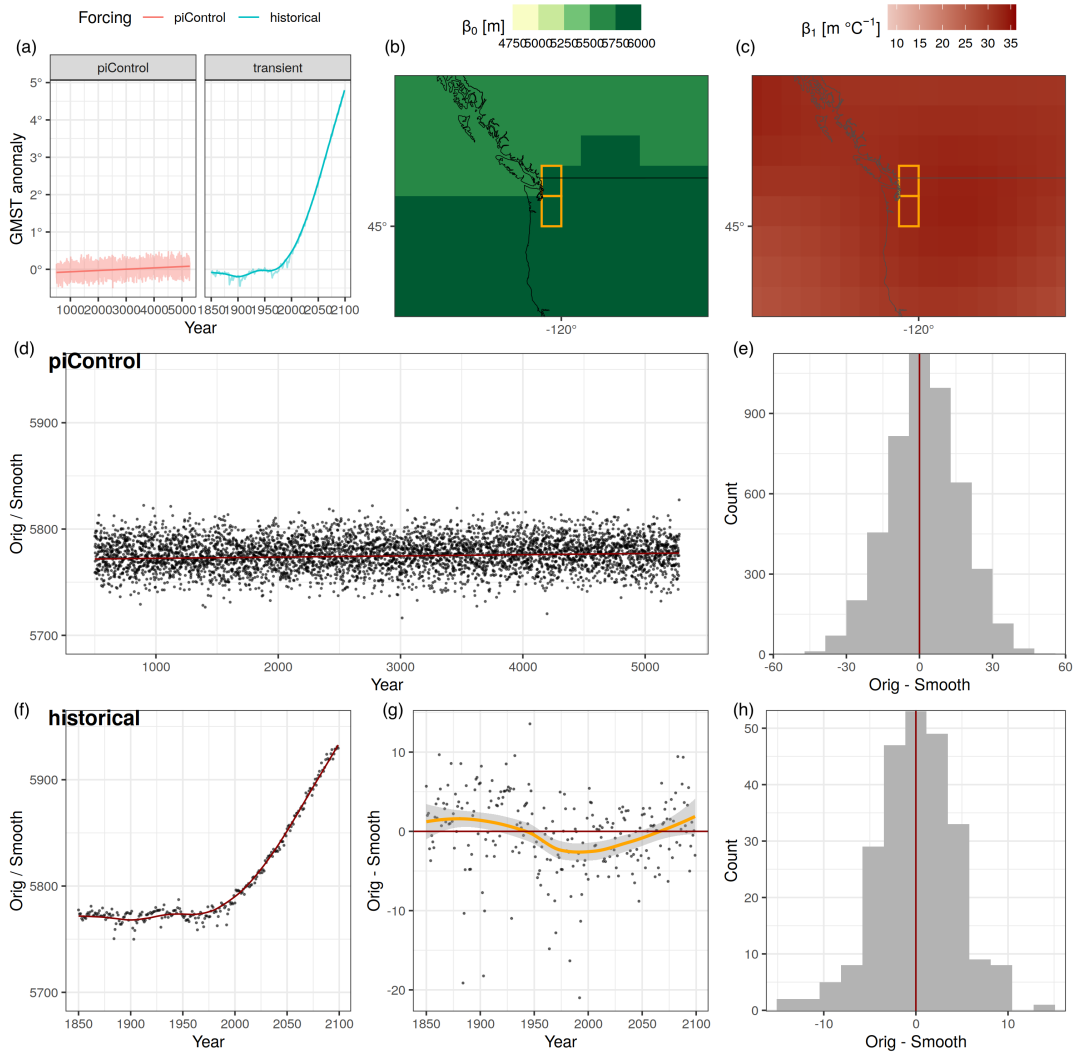

Figure S4. Evaluation plot for least-squares model of CESM12 JJA mean geopotential height regressed against smoothed GMST at the PNW location. (a) The smoothed GMST temperature used as a predictor, (b) and (c) the estimated regression coefficients (as in supplementary Fig. S3), (d) scatter plot of pre-industrial JJA mean geopotential height at WEU grid points and estimated regression line, (e) histogram of residuals, (f) as (d) but for the transient ensemble mean JJA geopotential height, (g) residuals with smoothed trend line and (h) as (e)

.

Table S1. The areal definitions and corresponding actual extreme heatwave events (mid-date of the seven-day period) in the ERA5 reanalysis dataset analysed in this study.

Location	Abbrev.	Lat. extension	Lon. extension	Extreme event
Western Europe	WEU	0° – 5° E	45° – 47.5° N	08 August 2003
Western Russia	WRU	35° – 45° E	52.5° – 57.5° N	05 August 2010
Pacific North-West	PNW	120° – 122.5° W	45° – 50° N	28 June 2021

11. The predictor value $x_{\mathrm{SM}}(t,s,M)$ is the weighted average of the previous and current monthly detrended and scaled, and for the area of interest spatially averaged soil moisture value,

$$x_{\mathrm{SM}}(t,s,M) = \delta x_{\mathrm{SM}}(t_{\mathrm{mon}},s,M) + (1-\delta)x_{\mathrm{SM}}(t_{\mathrm{mon}}-1,s,M) \tag{S7}$$

where $\delta = t_{\mathrm{day}}/30$, i.e. weighting the current month depending on how far into the month the extreme occurred.

12. The vector of predictor values $\boldsymbol{x}_{\mathrm{Z}}(t,s,M)$ is the detrended and scaled geopotential height field at the 512 grid points in a region of $\pm 40°$ lon and $\pm 20°$ lat from the centre of location s, at the date of the BM event t.

13. In the last step before model training, predictor values $\boldsymbol{x}_{\mathrm{Z}}(t,s,M)$ at grid points i,j are standardised across temporal multi-model means, which should make the inference procedure more stable:

$$\tilde{x}_{\mathrm{Z};i,j}(t,s,M) = \frac{x_{\mathrm{Z};i,j}(t,s,M) - \bar{x}_{\mathrm{Z};i,j}(s)}{\mathrm{sd}\big(x_{\mathrm{Z};i,j}(s)\big)} \tag{S8}$$

The result of the pre-processing is a set of four variables, constituting the input of the statistical model, summarised in Table S2. The statistical analysis is always location and model specific. The respective indices s and M will therefore be omitted for the sake of better readability.

Table S2. Pre-processed variables Summary of variables used as predictant and predictors for the statistical model (indices for location s and model M are omitted).

Var.	Dim.	Unit	Physical process variable
$y_{\mathrm{Tx7d}}(t)$	$\in \mathbb{R}_+$	[°C]	Five-year maximum Tx7d
$x_{\mathrm{GMST}}(t)$	$\in \mathbb{R}_+$	[°C]	Global mean surface temperature
$x_{\mathrm{SM}}(t)$	$\in \mathbb{R}$	[sd]	Detr./scaled soil moisture
$\tilde{\boldsymbol{x}}_{\mathrm{Z}}(t)$	$\in \mathbb{R}^{512}$	[sd]	Detr./scaled and stand. Z500 field

S1.3 Pre-processing evaluation

In this section, the pre-processing results are further evaluated to ensure that the input data is comparable across climate model and reanalysis datasets. For supplementary Fig. S5 only years after 1980 were considered, which are later used for the model fitting (for ERA5, all 14 five-year BM events since 1950 are considered to increase the sample size).

- **Seasonality:** Supplementary Fig. S5a) show the temporal distribution of BM throughout the year. For these mid-latitude northern hemispheric locations, the main season for heatwaves is from mid-June to early September, where the peak shifts by a few days to later dates in most of the climate model datasets. However, since these trends are quite marginal and the distributions largely agree, we assume no confounding effect of differences in heatwave seasonality.

- **Mean and variability:** In order to compare coefficients across statistical models derived from different climate model datasets, the predictors should be comparable in distribution. Furthermore, we hope to remove the thermodynamically forced GMST signal from the geopotential height and soil moisture predictors. Thus their distributions should also be independent of the forcing scenario. Densities of detrended and scaled soil moisture x_{SM} and geopotential height $x_{\mathrm{Z}}(s)$ (average over the area of interest s, analogous to x_{SM}) are shown in supplementary Fig. S5b) and c).
  The agreement in the distribution of $x_{\mathrm{Z}}(s)$ across climate model and reanalysis datasets, but also across the three locations, is certainly reasonable, indicating that most five-year BM heatwave events occur under a one- to three-sigma geopotential height anomaly relative to summer averages. Only in the case of CESM2, where there is a small trend towards higher anomalies detectable in stronger forcing scenarios, do the distributions seem rather stationary for the other model/location combinations. This indicates that forced trends in summer mean geopotential height have successfully been removed.
  Variations in density locations and shapes are larger in the case of x_{SM}, which is expected due to the challenges related to the modelling of land-surface processes. However, both the detrending and scaling were at least partially successful, considering the huge differences in absolute values as shown in Fig. 1a).

- **Spatial correspondence of Z500 fields:** The question remains whether also the spatial structure of the Z500

field is similar across different climate model and reanalysis datasets, which is again a prerequisite to compare and apply the statistical models derived thereof. There is good agreement in the average Z500 predictor fields (Fig. 2a), especially when considering that the ERA5 dataset consists of only 14 samples (14 five year block maxima events), which also applies to the variance of the pre-processed geopotential height fields (supplementary Fig. S13). Not only do the first and second moments of the $\tilde{\boldsymbol{x}}_Z$ predictor vector agree well across datasets, but Fig. 2b) also shows good agreement in the six also leading principle component (PC) patterns (Fig. 2b), whose cumulative explained variability amounts to 83 % (ERA5) and 84 % (CESM12). Supplementary Fig. S16 further indicates that the pattern correlation remains strong also for higher order PC loading patterns, and the PC are also stable over the full range of GMST changes (supplementary Fig. S17), confirming that detrending with respect to GMST is successful and that the circulation structure governing heat extremes does not change over time.

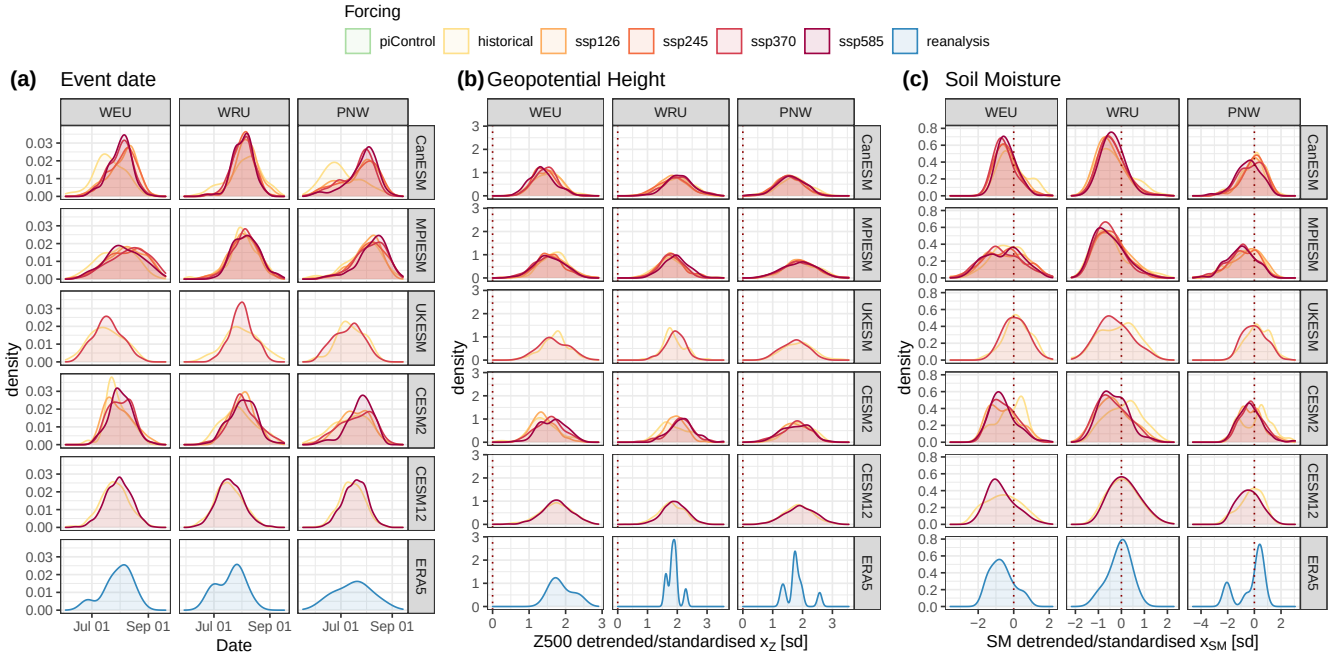

Figure S5. Densities of (a) the extreme event date t_{doy} (day of year), (b) detrended and scaled geopotential height $x_{\text{Z}}(s)$ (over the area of interest), and (c) detrended and scaled soil moisture x_{SM} for locations WEU, WRU and PNW (columns) and different datasets (rows).

S2 The statistical model

S2.1 The block size

The selection of an optimal block size is a classical bias–variance trade-off problem encountered in various fields of statistical modelling. In case a too small block size is chosen, non-extreme data is considered, potentially biasing statistical estimates. However, for too large block sizes, fewer data remains to be analysed, such that uncertainty or variance in the estimates will increase. The effect of block size definitions on extreme value analyses of climatological extreme events are discussed by Huang et al. (2016) and Ben Alaya et al. (2020, 2021). Assessing various block size definitions and testing the max-stability of the shape parameter ξ indicates that a minimum block size of five years is necessary (supplementary Fig. S6). A larger than annual block size also seems reasonable from a climatological perspective, as the potential of experiencing maximum temperatures is mostly confined to a seasonal window of roughly $2-3$ summer months (supplementary Fig. S5a), and strong autocorrelation in temperature further reduces the number of potential heatwave events. In order to minimise the potential influence of warming within a five-year period, temperature maxima y_{Tx7d} were instead sampled from annual blocks of five climate model ensemble members, which is equivalent under the assumption that the distribution of temperature maxima y_{Tx7d} across climate model ensemble members is identical in a given year.

S2.2 Regularisation of the GEV

For the *full* model, the Z500 effect $\mu_Z^*(t)$ in Eq. (2) is a dot product of the coefficient vector $\boldsymbol{\mu}_Z$ and the event specific Z500 anomaly field $\tilde{\boldsymbol{x}}_Z(t)$. In the case of the PNW location, given the spatial extend of the Z500 field considered ($\pm 40°$ in longitude and $\pm 20°$ in latitude), $p = 528$ coefficients $\widehat{\boldsymbol{\mu}}_Z$ have to be estimated (33 grid points in longitude, and 16 in latitude). As only five-year block maximum data from 1980–2089 is considered for the parameter estimation, the number of estimated parameters is not substantially lower relative to the number of input data points (e.g. in the case of the CESM12 large ensemble dataset, $n = 1782$ temperature block maxima were considered for the fit). This would lead to substantial overfitting and thus high variance in the estimates of the coefficient vector $\boldsymbol{\mu}_Z$. Furthermore, the Z500 information is strongly correlated across neighbouring grid points, also substantially increasing variance in the estimated coefficients, thus leading to highly unstable or non-robust estimation results.

Overfitting and high variance caused by a large number of highly correlated predictors can be addressed by regularising the estimated coefficients (Hastie et al., 2009). In a linear regression context (ordinary least squares regression), parameter estimates $\widehat{\boldsymbol{\beta}}$ are generally obtained by minimising the sum of squared residuals RSS. For regularised regression, an additional penalty term is considered in the objective function, such as the sum of absolute coefficient estimates (lasso) or the sum of squared coefficient estimates (ridge), or a mixture of the two (elastic net). The following equation states the objective function for ridge regression, with an L_2 ridge penalty term

$$\widehat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \Big\{ \underbrace{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2}_{\text{RSS}} + \underbrace{\lambda \boldsymbol{\beta}^\intercal \boldsymbol{\beta}}_{\text{Ridge penalty}} \Big\}, \tag{S9}$$

where $\boldsymbol{X}$ is a matrix of predictor variables, and $\boldsymbol{y}$ is the predictant vector. The minimisation of the RSS and the ridge penalty term, which is also affecting the coefficient vector $\widehat{\boldsymbol{\beta}}$ (referred to as shrinkage), reduces variance but comes at the cost of a bias (in case all classical assumptions of the linear model are fulfilled). The level of shrinkage, and thus the regularisation strength, is determined by the hyper-parameter $\lambda \geq 0$, which can be optimised by cross-validation. In the case of correlated predictor variables, the effect of shrinkage is largest for principle components (or directions in the column space of $\boldsymbol{X}$) with small singular values, i.e. which explain only little variance. The method is closely related to principle component regression, where the smallest eigenvalue components are discarded. Resulting estimates are also equivalent to Bayesian posterior mean estimates of the respective coefficients, where a Gaussian prior with mean zero is selected for the coefficients $\boldsymbol{\beta}$, and the respective variance determines the shrinkage (all derivations can be found in Ch. 3.4/3.5 of Hastie et al., 2009). Sippel et al. (2019) apply regularised regression on surface pressure to retrieve a forced climate signal across variables and spatial scales, also discussing the methodological aspects and impacts of different regression models.

The objective function in the context of the non-stationary GEV model is the likelihood, which is maximised with respect to the available data $y_{\text{Tx7d}}(t)$. However, for numerical reasons, the negative log-likelihood $\ell(\mu, \sigma, \xi)$ is usually minimised, where the general form for a non-stationary GEV is provided by Coles (2001, p. 108). With the GEV model formulation in Eq. (1), the negative log-likelihood is in our case

$$\begin{aligned} -\ell_{\text{reg}}\big(\mu(t), \sigma(t), \xi\big) = \\ \sum_{t=1}^{T} \Big\{ \log\sigma(t) + \big(1 + \frac{1}{\xi}\big) \log\Big[1 + \xi\big(\widetilde{y}_{\text{Tx7d}}\big)\Big] + \\ \Big[1 + \xi\big(\widetilde{y}_{\text{Tx7d}}\big)\Big]^{-1/\xi} \Big\} + \underbrace{\lambda \boldsymbol{\mu}_Z^\intercal \mathbf{Q} \boldsymbol{\mu}_Z}_{\text{Penalty term}}, \end{aligned} \tag{S10}$$

where $\widetilde{y}_{\text{Tx7d}} = \frac{y_{\text{Tx7d}}(t) - \mu(t)}{\sigma(t)}$ with a non-stationary location parameter $\mu(t)$, as in Eq. (2) and scale parameter $\sigma(t)$, as in Eq. (3), and an invariant shape parameter ξ, as in Eq. (4), evaluated at temperature maxima values y_{Tx7d}. The regularisation term is added to the negative log-likelihood, taking

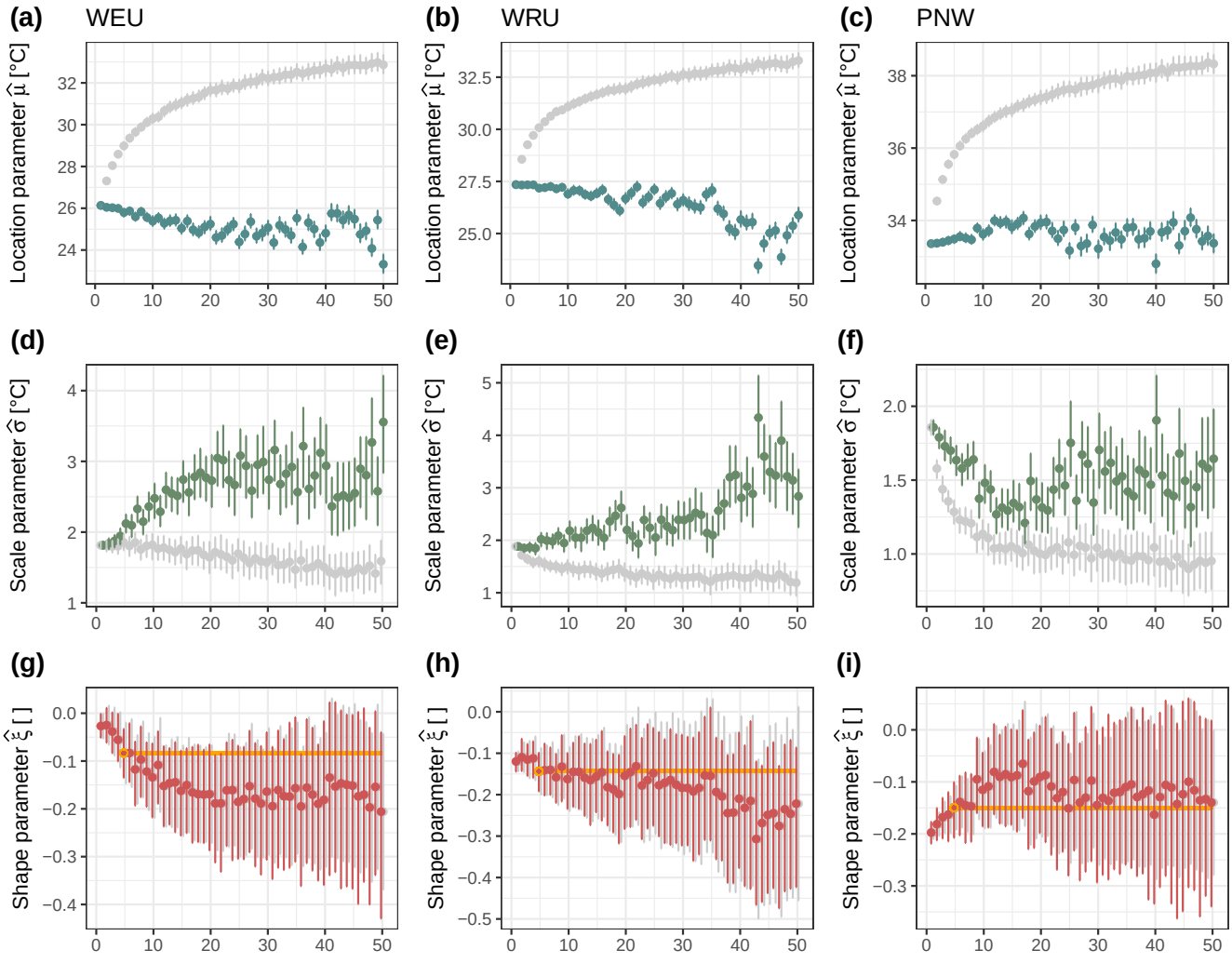

Figure S6. Stationary GEV location $\widehat{\mu}$ (first row, a-c), scale $\widehat{\sigma}$ (second row, d-f) and shape $\widehat{\xi}$ (third column, g-i) parameter estimates based on increasing block size m for grid points in Western Europe (left column a/d/g), in Western Russia (middle column b/e/h), and in the Pacific North-West area (right column c/f/i) in a centennial pre-industrial control run (4780 years). In the case of the location and scale parameter, grey estimates are obtained for the specific block size m, whereas coloured estimates are adjusted for the target block size $m = 1$ year (c.f. supplementary Section S2.5), with the respective bootstrap 99% CI. In the case of the shape parameter, the error bars mark 99% bootstrap CI in colour and 99% normal approximation Wald-type CI in grey. The orange line marks the five-year estimate.

a similar form as in Eq. (S9), with a hyper-parameter λ determining the shrinkage and penalising a large L_2 norm of the Z500 coefficient vector $\widehat{\boldsymbol{\mu}}_{\mathrm{Z}}$. Further prior knowledge on the connectivity of Z500 predictors is introduced via a normalised Laplacian matrix $\mathbf{Q}$ (a symmetric and positive definite matrix), encoding the adjacency of Z500 grid points (Karas et al., 2019). This measure further promotes a spatially smooth field of estimated coefficients $\widehat{\boldsymbol{\mu}}_{\mathrm{Z}}$. If $\mathbf{Q}$ was the identity matrix, the formulation would be equivalent to the L_2 penalty in Eq. (S9).

Before fitting a GEV distribution by maximising the negative log-likelihood as stated in Eq. (S10), the hyper-parameter λ has to be set, and suitable starting values for the optimisation have to be determined. Determining a suitable value of λ is decisive to obtain a statistical model that neither over- nor under-fits the data. The optimisation of λ was conducted on pre-industrial or historical data of five-year temperature maxima before 1900, assuming stationary global climate conditions ($x_{\mathrm{GMST}} = 0$). A ridge regression model with x_{SM} and $\tilde{\boldsymbol{x}}_{\mathrm{Z}}$ as predictors and $\boldsymbol{\beta}_{\mathrm{Z}}$ and β_{SM} as corresponding coefficients is fit to all pre-industrial temperature extreme events (analogous to Eq. (S9), where only the Z500 coefficients $\boldsymbol{\beta}_{\mathrm{Z}}$ are subject to regularisation),

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}}_{\mathrm{Z}} \\ \widehat{\beta}_{\mathrm{SM}} \end{bmatrix} := \underset{\boldsymbol{\beta}_{\mathrm{Z}}, \beta_{\mathrm{SM}}}{\operatorname{argmin}} \Big\{ \| \boldsymbol{y}_{\mathrm{Tx7d}} - \boldsymbol{x}_{\mathrm{SM}} \beta_{\mathrm{SM}} - \boldsymbol{X}_{\mathrm{Z}} \boldsymbol{\beta}_{\mathrm{Z}} \|_2^2 + \widetilde{\lambda} \boldsymbol{\beta}_{\mathrm{Z}}{}^{\intercal} \mathbf{Q} \boldsymbol{\beta}_{\mathrm{Z}} \Big\} \quad \text{(S11)}$$

for a range of ridge regularisation hyper-parameter values $\widetilde{\lambda}$ from 10^{-3} to 10^{8}. Supplementary Fig. S7a) shows estimated Z500 coefficient maps $\widehat{\boldsymbol{\beta}}_{\mathrm{Z}}$ for increasingly strong shrinkage parameters $\widetilde{\lambda}$. The structure becomes increasingly smoother until the pattern becomes that of the first principle component (c.f. Fig 2b), in accordance with theory (Hastie et al., 2009). To determine the optimal hyper-parameter $\widetilde{\lambda}_{\mathrm{opt}}$, the respective ridge models are evaluated on extreme temperature events of the remaining climate model datasets (across-dataset cross-validation), using R^2 as a measure for predictive skill. Maximum skill is reached for values of $\widetilde{\lambda}$ around 10^3 to 10^4, where corresponding coefficient fields are smooth but still show distinct spatial structures, especially a "blob" of positive coefficient estimates in the proximity of the area of interest, consistent across all climate model datasets (supplementary Fig. S7b). It should be noted that with the across-dataset cross-validation, stronger regularisation (higher $\widetilde{\lambda}_{\mathrm{opt}}$ values) and thus more conservative predictive capabilities of the Z500 coefficients $\widehat{\boldsymbol{\beta}}_{\mathrm{Z,opt}}$ should be expected.

The estimates of $\widehat{\boldsymbol{\beta}}_{\mathrm{Z,opt}}$ and $\widehat{\beta}_{\mathrm{SM,opt}}$ serve in the GEV likelihood optimisation procedure in two ways; first, they are used as initial values of the respective parameters (enabling a "warm start" of the maximum likelihood parameter optimisation), which proves to be vital in order to get robust Z500 coefficient estimates $\widehat{\boldsymbol{\mu}}_{\mathrm{Z}}$. The idea of "learning" the circulation effect in pre-industrial simulations and evaluating it in transient conditions was also already applied by Deser et al. (2016). Furthermore, the $\widetilde{\lambda}_{\mathrm{opt}}$ found for the ridge models is not directly applicable in the regularised likelihood in Eq. (S10), as the remaining objective function is fundamentally different compared to Eq. (S11). Therefore, a grid search over a range of λ values form 10^0 to 10^6 is conducted to determine the λ_{opt} value where the L_2 norm of the estimated location parameter coefficients $\widehat{\boldsymbol{\mu}}_{\mathrm{Z}}$ is most similar to that of the optimal ridge regression coefficients $\widehat{\boldsymbol{\beta}}_{\mathrm{Z,opt}}$,

$$\lambda_{\mathrm{opt}} := \underset{\lambda}{\operatorname{argmin}} \, | \, \widehat{\boldsymbol{\beta}}_{\mathrm{Z,opt}}{}^{\intercal} \mathbf{Q} \widehat{\boldsymbol{\beta}}_{\mathrm{Z,opt}} - \widehat{\boldsymbol{\mu}}_{\mathrm{Z}}(\lambda)^{\intercal} \mathbf{Q} \widehat{\boldsymbol{\mu}}_{\mathrm{Z}}(\lambda) \, | . \quad \text{(S12)}$$

Regularised maximum likelihood GEV parameter estimates are then obtained by minimising the negative log-likelihood in Eq. (S10) with the optimised shrinkage hyper-parameter λ_{opt}. The effect of the hyper-parameter selection is also considered when determining the parameter CI. For the parametric bootstrap, 600 $\widetilde{\lambda}$ values are sampled from the set of all $\widetilde{\lambda}$ values where the corresponding cross-validation R^2 score is larger than 95 % of the optimal R^2 score, probability-weighted by the respective R^2 score (a higher R^2 score means higher probability of the respective $\widetilde{\lambda}$ value), thus integrating the uncertainty induced by the selection of this hyper-parameter. For all 600 λ values – determined as in Eq. (S12) – a random sample is drawn from the GEV distribution with the respective regularised maximum likelihood parameters estimated for the corresponding λ value, and another GEV distribution is fit to the respective random sample. Given the set of 600 parametric bootstrap parameter estimates, percentile parameter CI can be estimated (Gilleland, 2020a). Given the fact that the GEV distribution is generally bounded ($\widehat{\xi} < 0$), parametric bootstrap CI should be able to provide reasonable estimates of the uncertainty (Gilleland, 2020b). For the regularisation of the GEV likelihood, existing code in the `extRemes` package (Gilleland and Katz, 2016) was adjusted, and the minimisation of the negative log-likelihood was conducted with the simulated annealing algorithm (Bélisle, 1992) provided in the respective R function `optim()`.

Regularisation in the context of extreme value modelling has already been applied to constrain parameters (generalised/penalised maximum likelihood) in hydrometeorological context (El Adlouni et al., 2007; Cannon, 2010; Bücher et al., 2021), or to constrain the flexibility of non-linear parameter functions (Pauli and Coles, 2001), providing the basis for generalised additive modelling of GEV parameters (Chavez-Demoulin and Davison, 2005). The latter was used to test the improvement of a highly flexible non-linear model relative to the linear additive structure of Eq. (1), see supplementary Fig. S20. As an alternative to the likelihood, the CRPS score (c.f. supplementary Section S2.3) could have been used as an objective function-function (Friederichs and Thorarinsdottir, 2012).

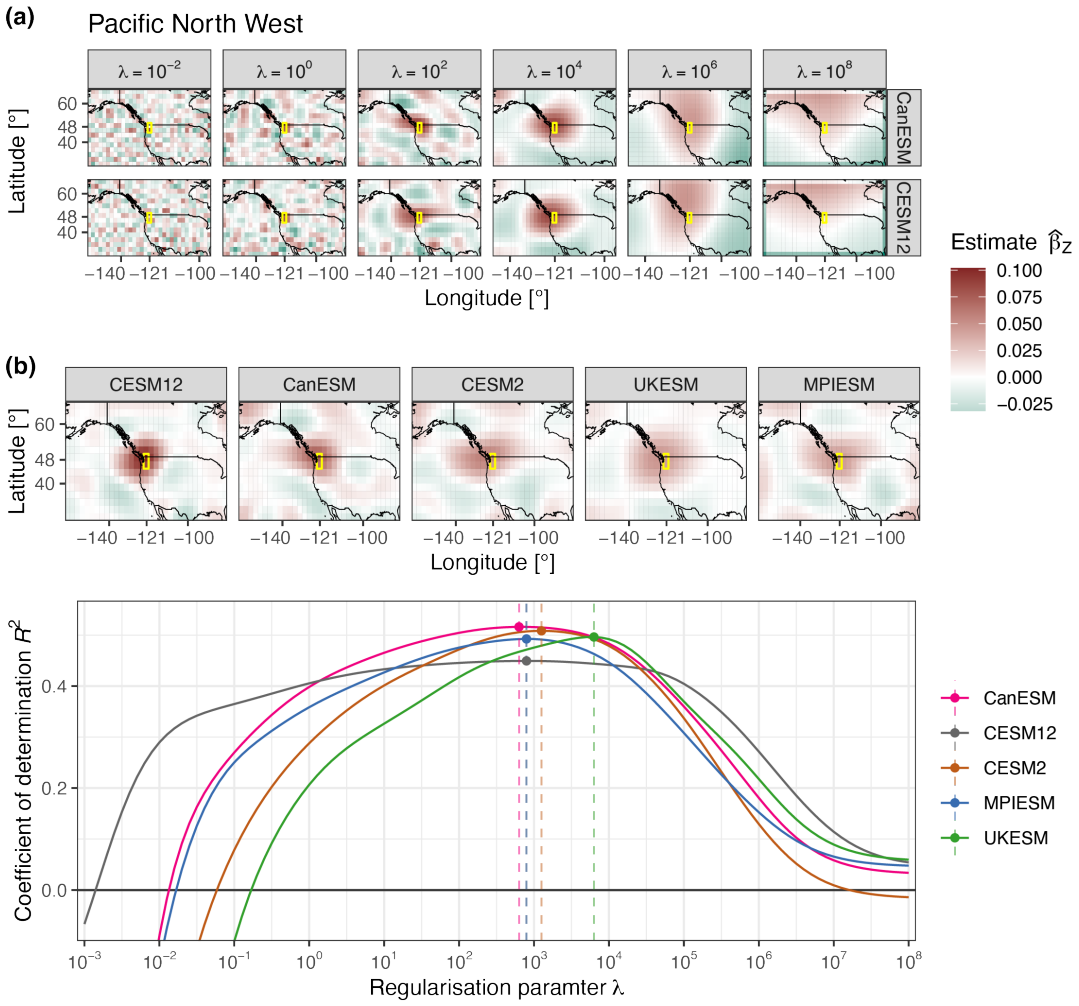

Figure S7. (a) Z500 ridge coefficient estimates $\widehat{\boldsymbol{\beta}}_{\mathrm{Z}}$ for the CanESM and CESM12 datasets (rows) and various shrinkage hyper-parameters $\widetilde{\lambda}$ (columns). (b) Optimal estimates $\widehat{\boldsymbol{\beta}}_{\mathrm{Z,opt}}$ for all climate model datasets, and corresponding R^2 scores for shrinkage values $\widetilde{\lambda}$, where the optimal values $\widetilde{\lambda}_{\mathrm{opt}}$ are indicated with the vertical dashed lines.

S2.3 Evaluation scores

The coefficient of determination R^2 denotes the proportion of variance in the predictand variable (y_{Tx7d}) across all time steps t which is explained by the predictive model – the mean of the *full* GEV model $\bar{F}_{\mathrm{Y},full}(t)$ as in Eq. (1) with the estimated, non-stationary location parameter $\widehat{\mu}(t)$ as in Eq. (2) – relative to the variance explained by a baseline model, in our case, the means of the non-stationary *GMST only* GEV model $\bar{F}_{\mathrm{Y},GMST\ only}$ with $\widehat{\mu}(t)$ as in Eq. (6):

$$R^2 = 1 - \frac{\sum_{t=1}^{T}\left(y_{\mathrm{Tx7d}}(t) - \bar{F}_{\mathrm{Y},full}(t)\right)^2}{\sum_{t=1}^{T}\left(y_{\mathrm{Tx7d}}(t) - \bar{F}_{\mathrm{Y},GMST\ only}(t)\right)^2} \tag{S13}$$

The continuous ranked probability score CRPS (Wilks, 2011) is a strictly proper scoring rule (Gneiting et al., 2007) taking the full probabilistic nature of the fitted GEV distribution into account, determining the sharpness of the estimated probability distribution $F_Y(t)$.

$$\mathrm{CRPS} = \frac{1}{T}\sum_{t=1}^{T}\int_{-\infty}^{\infty}\left(F_Y(y,t) - \mathbb{1}(y_{\mathrm{Tx7d}}(t) \geq y)\right)^2 dy \tag{S14}$$

Additionally, the log- or ignorance score – a proper and local score (Bröcker and Smith, 2007) – was also calculated, but is not shown. It should be highlighted that none of these measures accounts for the larger number of predictors (as the adjusted-R^2 measure would) in the *full* model, compared to the two sub-models. For such an adjustment, the effect of regularisation would need to be quantified, which would require a careful assessment of the effective degrees-of-freedom (Kaufman and Rosset, 2014; Janson et al., 2015) in the context of ill-conditioned models (Karlsson et al., 2019). This was not attempted, but it is assumed that by optimising the regularisation parameter $\widetilde{\lambda}$ in terms of predic-

tive skill across different climate model data, the effective degrees-of-freedom difference is implicitly accounted for, thus a "fair" comparison between the (regularised) *full* and (non-regularised) nested sub-models is conducted.

S2.4 GEV parameter adjustment

The following section summarises the necessary steps to adjust for offsets between climate model and ERA5 reanalysis data, in order to apply GEV models estimated on the former to heatwave events of the latter (given that two CMIP6 models are included, which are subject to strong warming, CanESM and UKESM, see Tokarska et al., 2020). It was decided to only account for constant offsets between GMST input data x_{GMST} and temperature extreme data y_{Tx7d} across datasets. Thus only the intercept parameter $\widehat{\mu}_0$ has to be adjusted, corresponding to a simple shift of the estimated non-stationary GEV distribution.

Specifically, GMST values x_{GMST} of both climate model and reanalysis datasets were shifted by Δx_{GMST} such that the average GMST value for the period 1981 – 2010 equals 0.63 °C, defined as the global average surface temperature in the respective reference period (IPCC, 2018). The estimated intercept $\widehat{\mu}_{\text{GMST}}$ of the location parameter is adjusted by the respective GMST effect $-\widehat{\mu}_{\text{GMST}} \cdot \Delta x_{\text{GMST}}$ (the negative sign is necessary to compensate for the shift in a predictor variable x_{GMST}). Additionally, the respective GMST offset Δx_{GMST} is also subtracted from Tx7d values y_{Tx7d}.

Furthermore, the offset in Tx7d at the corresponding locations is quantified by the difference between median one-year Tx7d from 1950 to 2021 in the climate model dataset and ERA5, Δy_{Tx7d}. The reason for considering one-year block maxima is to reduce the sampling uncertainty of the median ERA5 Tx7d value, assuming that the differences in one-year block maxima are comparable to differences in five-year block maxima. The total effect of these adjustment steps leads to an adjusted intercept estimate $\widehat{\mu}_{0,\text{adj}}$, whereas the remaining parameter estimates remain unchanged:

$$\widehat{\mu}_{0,\text{adj}} = \widehat{\mu}_0 \underbrace{-\widehat{\mu}_{\text{GMST}} \cdot \Delta x_{\text{GMST}}}_{\text{Correction for GMST offset}} \underbrace{+\Delta y_{\text{Tx7d}}}_{\text{Correction for Tx7d offset}} \tag{S15}$$

S2.5 Multi-year block maxima conversion of GEV parameters and exceedance probabilities

The event likelihood can be quantified in terms of conditional annual exceedance probability p_{ex} (Cooley et al., 2019). Most studies applying non-stationary extreme value theory instead use the return period, i.e. the inverse of the AEP, but given the conditioning on event-specific process variables, these could not be interpreted as expected waiting times, thus return periods might be misinterpreted (even though return periods could be adjusted for the non-stationarity effect, c.f. Cooley, 2013). Five-year exceedance probabilities (based on GEV distributions derived from five-year block maxima) were converted to obtain annual exceedance probabilities. This conversion of exceedance probabilities retrieved from a GEV fit based on multi-year block size m maxima (in our case, the block size is $m = 5$ years) to annual exceedance probabilities is possible via an adjustment of GEV parameters given the max-stability property of the GEV, or directly adjusting the exceedance probability based on the relationship of the respective one-year and multi-year complementary cumulative distribution functions.

In case parameters are estimated for maxima of different block sizes m, the respective GEV parameters can be adjusted for the effect of the larger block size $m = m'$, since due to the max-stability property

$$\left(F_{Y,m=1}(y;\mu_1,\sigma_1,\xi_1)\right)^{m'} = F_{Y,m=m'}(y;\mu_{m'},\sigma_{m'},\xi_{m'}) \tag{S16}$$

the equivalent $m = 1$ year parameters are (derivation not shown):

$$\begin{cases} \xi_1 &= \xi_{m'} \\ \sigma_1 &= \sigma_{m'} \cdot m'^{-\xi_1} \\ \mu_1 &= \mu_{m'} - \sigma_1/\xi_1 \cdot (m'^{\xi_1} - 1). \end{cases} \tag{S17}$$

Given the adjusted parameters, all further tail measures (quantiles, exceedance probabilities, etc.) can be retrieved from the respective GEV distribution.

However, multi-year exceedance probabilities $p_{\text{ex},m=m'} = \mathbb{P}\left(Y \geq y_{m=m'}\right)$ retrieved from multi-year block maxima $y_{m=m'}$ can also be adjusted directly to *annual* exceedance probabilities $p_{\text{ex},m=1} = \mathbb{P}\left(Y \geq y_{m=1}\right)$. Realising that a multi-year exceedance probability $p_{\text{ex},m=m'}$ can be understood as the probability of exceeding the respective threshold at least once in the individual m' years, we can express the respective likelihood as one minus a binomial distribution with $k = 0$ successes in m' trials each with probability $p_{\text{ex},1}$:

$$p_{\text{ex},m'} = 1 - \underbrace{(1-p_{\text{ex}})^{m'}}_{\text{Binom}(k=0;\ \pi=p_{\text{ex},1},m=m')} \tag{S18}$$

Reordering the terms provides the formula for expressing multi-year exceedance probabilities as AEP: $p_{\text{ex},1} = 1-(1-p_{\text{ex},m'})^{1/m'}$.

S3 Results

S3.1 Evaluation of GEV model fit to the data

Two aspects of the non-stationary GEV model need to be verified to assess fitness for purpose, i.e. the ability to estimate the relative contributions of physical process variables in the context of heat extremes; first, the regression aspect, i.e. whether the location parameter as a linear combination of input variables is capable of explaining a significant fraction of extreme temperature variability. Second, the conditional GEV probability distribution should be representative of the remaining variability not explained by the non-stationary location parameter.

There are several (heuristic) metrics which may be analysed in this context, like scatter plots of "predicted" (GEV mean) and "observed" (temperature maxima) data, with a focus on the regression aspect in the location parameter and quantile-quantile plots or probability integral transform plots PIT, which further assess the reliability or the overall fit of the GEV distribution, inspired by applications in weather forecast evaluation. These evaluations were also conducted on the input data used for fitting the GEV distributions (1980 – 2089), an independent testing period (2090 – 2100), and on events of both sets with a strong Z500 effect (among the largest 20 %, c.f. Fig 4a). Supplementary Fig. S8 shows the respective figures for the non-stationary GEV model fit at location PNW based on the CESM12 dataset.

Comparing the "observed" Tx7d values $y_{\text{Tx7d}}(t)$ to the mean of the GEV distribution ($\bar{F}_Y(t)$) estimated for the respective predictor values of the event (supplementary Fig. S8c) confirms that the underlying trend induced by global climate change is well represented in the statistical model. For this specific location (PNW) and underlying climate model dataset (CESM12), there is some deviation for extreme temperature values above 32 °C, where the statistical model seems to underestimate the actual event severity. It should be noted that these data points most likely occurred end of century, where the data (beyond the year 2090) has not been used for training.

However, in order to assess the fit of the GEV probability distribution to the data, further metrics are needed. The classical quantile-quantile plot for non-stationary GEV distributions is obtained by standardising both the GEV quantiles (standardised ranks) and "observed" temperature maxima (Gumbel standardised $\widetilde{y}$) data to a Gumbel distribution, $\widetilde{y}_{\text{Tx7d}}(t) = \frac{1}{\hat{\xi}} \log\left\{1 + \hat{\xi}\left(\frac{y_{\text{Tx7d}}(t) - \widehat{\mu}(t)}{\widehat{\sigma}(t)}\right)\right\}$. It should be noted that the quantile plot is not invariant to the choice of Gumbel as the reference distribution. However, due to the status within extreme value theory, it is arguably the most natural choice Coles (2001, p. 110). Supplementary Fig. S8a) shows a reasonably good fit for extremes in the body of the (non-stationary) GEV, but there are some occurrences of heat extremes where the predicted upper tail is too short to capture the event intensity. These events are not those deviating from the predicted GEV mean (supplementary Fig. S8c), as, for example, the strong outlier (with $\widetilde{y}_{\text{Tx7d}} = 12.1$) already occurs in model year 2025 (supplementary Fig. S8b) but is located extremely far in the tail (with an estimated exceedance probability of only $5.6 \cdot 10^{-6}$), even though both the Z500 and soil moisture conditions contributed significantly to the predicted location parameter ($\widehat{\mu}^*_{\text{Z}}(t) = +1.6\,°\text{C}$ and $\widehat{\mu}^*_{\text{SM}}(t) = +1.4\,°\text{C}$).

PIT plots can be understood as histograms of frequencies that the observed temperature maximum values y_{Tx7d} take in the respective estimated GEV probability distributions, $\text{PIT} = \widehat{F}_Y\big(y_{\text{Tx7d}}(t)\big)$. Assuming the temperature being stochastic draws of the respective probability distributions, the frequency of the PIT should roughly follow a uniform distribution (Gneiting and Raftery, 2007). 95 % probability bounds (in which the PIT frequency bars should lie under the null hypothesis that the fitted GEV distribution is the true probability distribution of the data) were estimated using parametric bootstrapping. The PIT plot in supplementary Fig. S8d) also confirms that the fitted, non-stationary GEV distribution covers the underlying data well in the sense that the frequency of occurrences in the respective quantiles of the distribution are well aligned with the estimated probability. In case the estimated tails of the distribution were too short (i.e. more data points than expected would occur in the tails), the PIT bars would have a convex shape, or if the distribution was too wide, it would be concave.

The analysis was also conducted for independent testing data (2090 – 2100) and data with strong Z500 forcing. From the figures (supplementary Fig. S21) we conclude that the non-stationary GEV model also performs reasonably under unseen conditions (end-of-century climate conditions), and also under extreme conditions with respect to strong circulation forcing. Thus both the regression and the probabilistic aspects of the model can be trusted.

S3.2 Assessment of additional covariates

The formulation of the statistical model does not explicitly account for the effect of seasonality on heatwave intensity, as Tx7d values y_{Tx7d} are not adjusted (e.g. normalised against a seasonal cycle) and no seasonality predictor (e.g. a sine/cosine harmonic regression predictor) is included. However, the monthly detrended and standardised soil moisture predictor x_{SM} is found to decrease at most locations as summer progresses (due to continuous drying over the summer months), thus implicitly providing information on seasonality. Put differently, a seasonality predictor would be heavily correlated with the SM predictor with little added information. Supplementary Fig. S9 shows that for Tx7d events occurring later in the summer season (roughly after mid-August at location PNW, corresponding to approximately 10 % of all events), the model tends to overestimate the heatwave intensity. But for the peak season (July to mid-August), where the

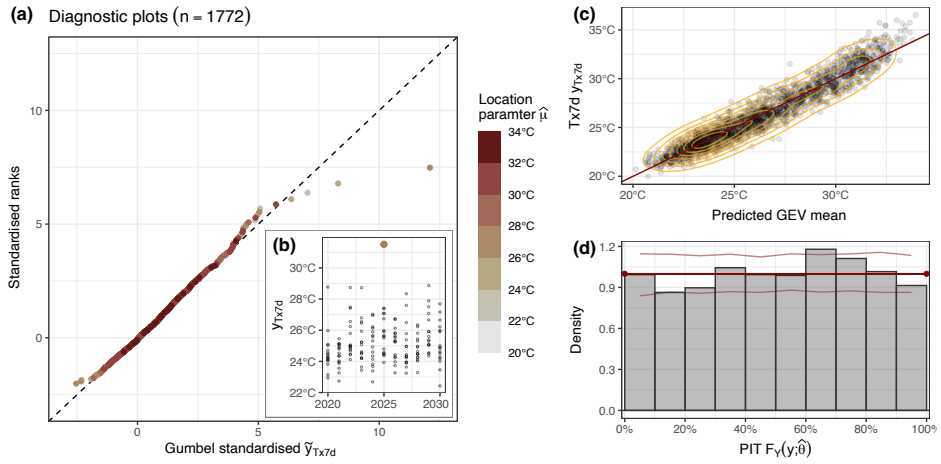

Figure S8. Model diagnostic plots for GEV model fits to CESM12 climate model dataset and location PNW: quantile plot (a) with a subplot of outlier event (b), a scatterplot of estimated GEV mean on the abscissa and actual Tx7d value on the ordinate (with kernel density in orange, c), and probability integral plot (with bootstrapping based 95 % probability bounds for the PIT under the null-hypothesis, d).

frequency of five-year block maxima is highest, the residuals hardly show any dependency on seasonality.

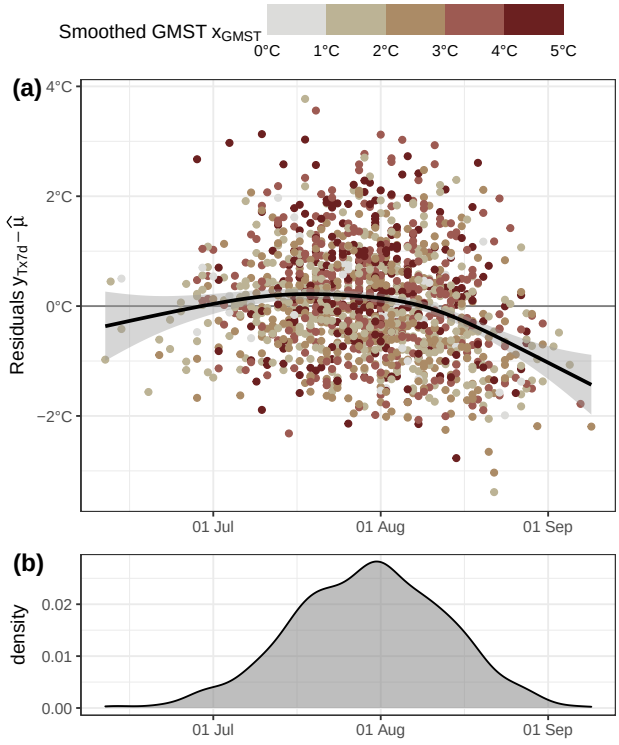

Figure S9. (a) Tx7d residuals – predictand values $y_{Tx7d}(t)$ minus the respective estimated location parameter $\hat{\mu}(t)$ – over the coarse of the summer season. The colouring refers to the respective smoothed GMST value x_{GMST}, and the black line with grey shading shows a smoothed trend line. (b) The density of occurrence of five-year Tx7d block maxima events throughout the year.

The GMST covariate x_{GMST} is intended to represent the combined effect of thermodynamic global warming on local Tx7d intensity (c.f. Section S1.2), for this reason, the annual ensemble mean temperature values are further smoothed in order to remove the effect of any low-frequency variability in x_{GMST}. However, it remains to evaluate whether low-frequency climate variability is still detectable in Tx7d residuals. Ensemble member specific GMST time series are correlated with Tx7d residuals (subtracting the estimated location parameter $\hat{\mu}(t)$ from the predictand value y_{Tx7d}), where the respective GMST series is smoothed continuously to represent modes of lower frequency internal GMST variability (increasing the smoothing hyper-parameters α, analogous to a larger window in a running mean filter). Supplementary Fig. S10a) shows raw GMST anomalies against the ensemble mean trend (black dots) and corresponding smoothed time series as coloured lines (weak smoothing in pink, strong smoothing in blue). To assess whether the correlation between smoothed GMST and Tx7d residuals is significant for a specific frequency (α hyper-parameter), the correlation of GMST is also conducted with resampled (shuffled) Tx7d residuals. Supplementary Fig. S10b) shows that the estimated correlation (red line) does not leave the grey band of bootstrapped correlation, thus, is not considered significant (the departure for very low smoothing values is most probably an artefact, as it would indicate a negative association of local Tx7d with global, non-smoothed GMST). Thus we conclude that there is no low-frequency thermodynamically forced signal in Tx7d residuals which would have to be integrated into the statistical model formulation.

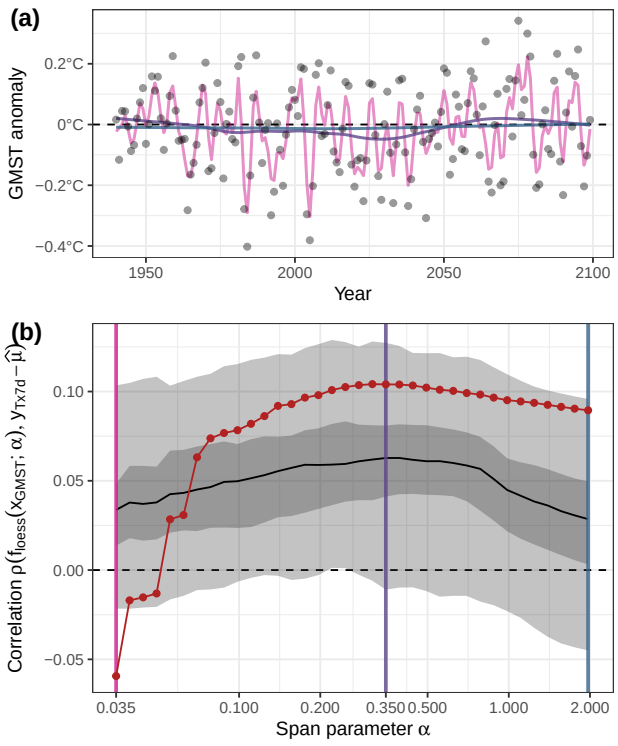

Figure S10. (a) GMST anomalies in ensemble member 23 of the CESM12 climate model dataset (smoothed ensemble mean GMST x_{GMST} subtracted from annual GMST in ensemble 23). The coloured lines indicate loess-smoothed fits to the anomalies with differing span parameter α. (b) The dark red line shows the correlation of Tx7d residuals (predictand values y_{Tx7d} minus the respective estimated location parameter $\widehat{\mu}(t)$) and loess-smoothed GMST anomalies with varying span parameter α, ranging from 0.35 (almost no smoothing) to 2.0 (very strong smoothing). The grey bands indicate 50 % and 95 % bootstrap CI of correlation where Tx7d values are matched to random GMST values.

S3.3 Aggregated evaluation of GEV model skill

The coefficient of determination R^2 is a measure for goodness-of-fit of the non-stationary regression component of the GEV model, evaluating how well the point estimates (the mean of the estimated GEV distribution, not the location parameter) correspond to the actual event intensity y_{Tx7d}. The *local Z500* and the *full* GEV models, informed by not just global warming but also event-specific physical process variables, are able to explain a larger fraction of the variability in heatwave intensity, i.e. showing a higher R^2 score. Over the testing period 2090 – 2100 (middle panel of Fig 4b), the *local Z500* model explains 31 % of the remaining variability not accounted for by the pure GMST effect, and the *full* model explains 43 %. Not surprisingly, the scores are higher when conditioning on events dominated by a strong Z500 effect (right panel of Fig 4b), but the ratio of *full* to *local Z500* model scores is comparable to those of the unconditional testing set (middle panel).

The CRPS score provides a measure for the fit of the overall GEV probability distribution, as it also rewards concentration of the probability density (sharpness) around the event intensity (Wilks, 2011). CRPS scores, which have a negative orientation (i.e. smaller values indicate higher skill), are shown in Fig 4c). The skill of the *full* model is again the largest across all evaluation datasets, even though the difference in skill with respect to the *local Z500* model is again smaller over the testing period (middle panel) than over the estimation period (left panel).

The *full* model has substantially more parameters and thus higher flexibility, being provided with the full geopotential height field as predictor vector, an aspect that would be heavily penalised if measures like the adjusted R^2 or the Akaike (AIC) and Bayesian (BIC) information criterion were applied. However, these scores assume the number of predictors to translate directly to the degrees of freedom, which is not the case for a regularised model, which structurally reduces the degrees of freedom via shrinkage. The regularisation hyper-parameter λ, responsible for the degree of parameter shrinkage, was retrieved with an across-dataset cross-validation approach (c.f. supplementary Sect. 2.2). This hyper-parameter selection assures a significant shrinkage of the coefficients $\widehat{\boldsymbol{\mu}}_Z$ and guards against overfitting of the *full* GEV model. We, therefore, argue that the resulting higher skill of the estimated GEV model is representative of only the desired information gain obtained from including the Z500 field, without any artificial skill due to overfitting.

S3.4 Climate model and location specific evaluation of GEV model skill

As outlined in Sect. 4.2, the scatter plot in Fig. S11a) confirms that the *full* model (considering the full Z500 field as predictor vector) but also models with only localised Z500 information (*local Z500* model) and only GMST (*GMST only* model) as predictor variables are capable of representing the large scale climate change driven trends in extreme temperature data. However, the *full* model shows larger skill both in terms so R^2 and CRPS throughout all location/model dataset combinations.

Higher skill of the *full* model is also observed when only considering data after 2090, which was not used for training (supplementary Fig. S11b). Please note that differences in skill between CESM12/UKESM (first and fourth row of supplementary Fig. S11b) and the remaining datasets are primarily driven by the fact that these models only have one future forcing scenario, and thus the *GMST only* model has no skill (as GMST is almost constant within the testing period 2090 – 2100). The skill of the *full* and *local Z500* model for these two climate model datasets (CESM12 and UKESM) can be interpreted as indicators of skill added purely from non-GMST predictors. Also, the fact that the skill of the *local Z500* model is higher than the skill of the *full* model for the UKESM testing dataset could indicate that the *full* model

overfitted on the training data. It should be noted that the UKESM testing dataset consists of only 20 data points (five-year block size maxima in 10 years of 10 SSP3-7.0 ensemble members). Thus sample uncertainty in these skill scores is probably large, as it is only the case for this specific location/model dataset combination that the *local Z500* model shows a higher skill score than the *full* model.

In supplementary Fig. S11c) skill scores are again displayed but with a focus on cases with strong Z500 forcing. The skill is generally comparable to the skill for the overall training data, besides the clearly weaker skill of the *GMST only* model (caused by the specific selection of data). It can be concluded that the model skill does not deteriorate for stronger geopotential height forcing situations.

S3.5 The Western Russian CESM12 (within model) heatwave

For a discussion of the method's capabilities, the statistical model is evaluated for a heatwave event of the CESM12 large ensemble dataset, thereby avoiding the uncertainty related to the use of reanalysis data. A climate model heatwave event was selected for the area affected by the Western Russian heatwave of 2010 (Barriopedro et al., 2011; Trenberth and Fasullo, 2012; Watanabe et al., 2013), with comparable relative magnitude and within the historical simulation period. The event occurred in ensemble member 4 on July 15 in the model year 1999, reaching an intensity of $y_{\mathrm{Tx7d}} = 28.1\,^{\circ}\mathrm{C}$ and thereby exceeding previous temperature maxima over the historical simulation period by large margins (supplementary Fig. S12a). A south-west to north-east elongated anomaly dominates the geopotential height field (supplementary Fig. S12c), with a clear imprint in surface temperature anomalies (supplementary Fig. S12d). The soil moisture anomaly associated with this event is at an all-time low with respect to the historical simulation (supplementary Fig. S12b).

Applying both the *GMST only* and the *full* GEV model (the parameter estimates of the latter are shown in supplementary Fig. S16) to the predictor values of this specific CESM12 heatwave event provides the estimated effect sizes of GMST ($\widehat{\mu}^{*}_{\mathrm{GMST}}$), Z500 ($\widehat{\mu}^{*}_{\mathrm{Z}}$), and SM ($\widehat{\mu}^{*}_{\mathrm{SM}}$), which are shown in supplementary Fig. S12e/f) with the corresponding GEV probability density curves. The estimated location parameter $\widehat{\mu}^{*}$ is the sum of these event-specific effects and the intercept estimate $\widehat{\mu}_0$, see Eq. (2). The intercept represents the pre-industrial average intensity of a five-year maximum temperature (vertical solid line). In this example, all process variables and the respective effect sizes contribute positively to the location parameter of the *full* model, thus shifting the distribution to higher temperature values. The finding that the relative Z500 effect is negligible for some climate model datasets should not be interpreted as if Z500 did not contribute to the event, but only that the Z500 effect was close to the average for a five-year temperature maximum heatwave event. There remains a large fraction of event intensity ($+2.5\,^{\circ}\mathrm{C}$) not explained by the predictor process variables, visualised as the grey horizontal double-arrow between the estimated location parameter $\widehat{\mu}^{*}$ and the actual event intensity y_{Tx7d}. As there remains an unexplained intensity gap, the event intensity still lies in the tail of the *full* GEV distribution, as the probability distribution is also narrower than the *GMST only* distribution, where the SM and Z500 effects are not accounted for (supplementary Fig. S12e).

It should not come as a surprise that the intensity of a record-breaking heatwave event is not explained by a first-order linear combination of three process variables in a statistical model. For such an event to unfold, there is an intractable amount of complex physical interactions and feedbacks and the spatio-temporal evolution of all the atmospheric and surface processes determining the properties of a heatwave event. These processes are not representable in a linear statistical model, such as potential long-term variability changes in full temperature distribution (Schär et al., 2004; Fischer et al., 2012), diabatic heating along the advection of air masses (Bieli et al., 2015), the effect of sea surface temperature anomalies (Schubert et al., 2014) or the effect of non-local land-surface conditions (Merrifield et al., 2019; Schumacher et al., 2019; Zhou and Yuan, 2022).

Considering the contribution of the different effects relative to the intensity exceeding the pre-industrial average ($y_{\mathrm{Tx7d}} - \widehat{\mu}_0$), almost 60 % is explained by the effects of GMST warming and the respective SM and Z500 anomalies (supplementary Fig. S12h). The estimates of the relative GMST effect in supplementary Fig. S12g/h) agree well for the *GMST only* and the *full* GEV model, indicating that the detrending of the SM and Z500 variables were successful, and thus all the full global warming signal is represented in the GMST covariate. However, estimates of the relative intensity explained differ significantly across GEV models trained on different climate model datasets. These differences reflect both the representation of heatwave events and their relationship with the chosen process variables in the climate models, the uncertainties related to internal variability, and the generally large stochasticity associated with extreme events.

How would the event intensity change under different global warming or soil moisture conditions? The location of the yellow dot in supplementary Fig. S13a), marking the estimated SM effect $\widehat{\mu}^{*}_{\mathrm{SM}}$ discussed above, again highlights the anomalous conditions and resulting pertinent contribution to the intensity of the event. Given the same prevalent event conditions (Z500 and SM anomalies and also unexplained remainder term), and under a pessimistic future warming scenario (RCP 8.5), the estimated event intensity would increase by more than 5 °C in 2060, following the dot-dashed horizontal arrow in supplementary Fig. S13b). Conversely, to reach the same intensity but under standard SM conditions (moving along the diagonal dashed arrow), warming conditions as reached in the model year 2020 would be necessary to compensate for the lack of an SM anomaly. To also compensate

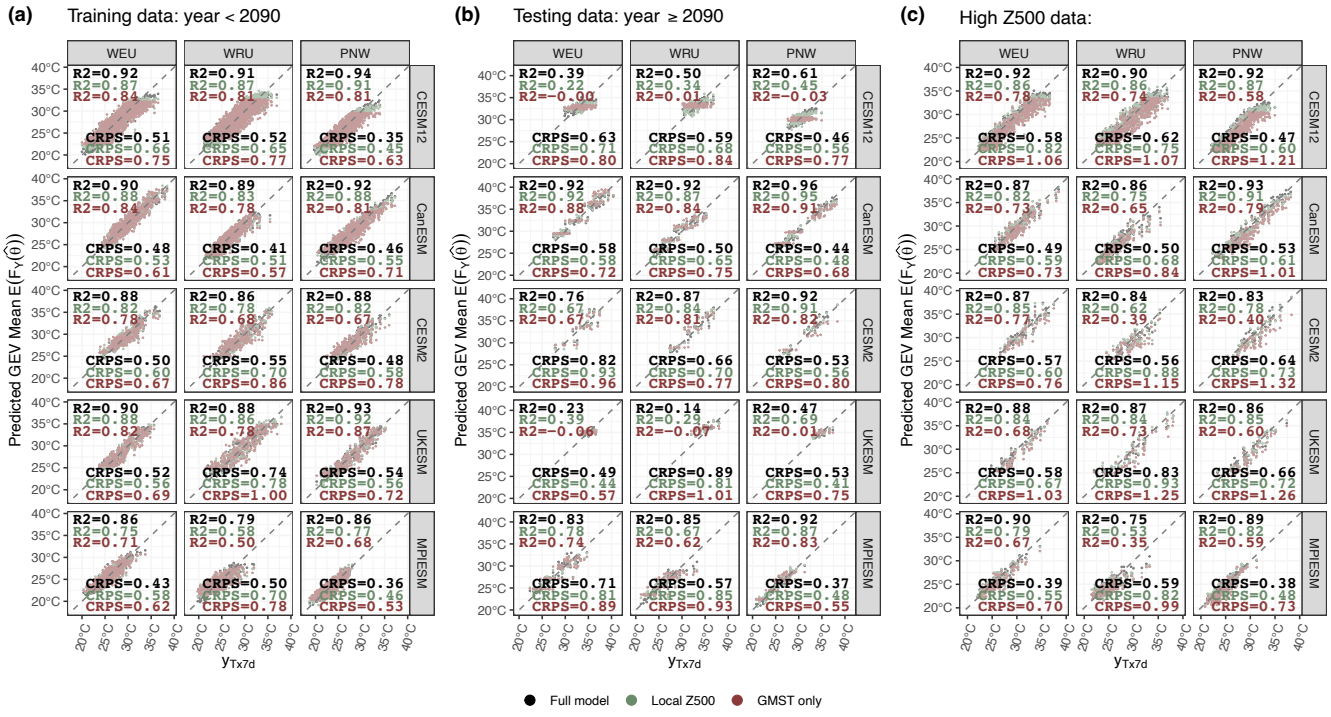

Figure S11. Scatter plots of Tx7d y_{Tx7d} values on the abscissa and predicted GEV mean $\bar{F}_Y(\widehat{\boldsymbol{\theta}})$ on the ordinate at locations WEU, WRU, and PNW (columns) and for different climate model datasets (rows). (a) Values for the training data (1980 – 2089), (b) for testing data (not used for model training, 2090 – 2100), and (c) for strong Z500 forcing cases ($\widehat{\mu}^*_{\text{Z}}(t) \geq q_{0.8}\{\widehat{\boldsymbol{\mu}}^*_{\text{Z}}\}$). Black colours refer to the *full* model in Eq. (2), green to a model with only localised Z500 $x_{\text{Z,loc}}$ information as a predictor (*local Z500* model), and red to a model with only GMST x_{GMST} as a predictor (*GMST only* model). The coefficient of determination R^2 (top-left) and CRPS skill score values (bottom-right) are displayed in the panel corners.

for the Z500 anomaly (with respect to average Z500 conditions during five-year maximum temperature events) and the unexplained remainder term, a warming level of close to 2.8 °C would need to be reached (moving along the dotted white arrow). Put differently, at this warming level, the event intensity of the analysed 1999 event would roughly correspond to the average five-year maximum heatwave intensity.

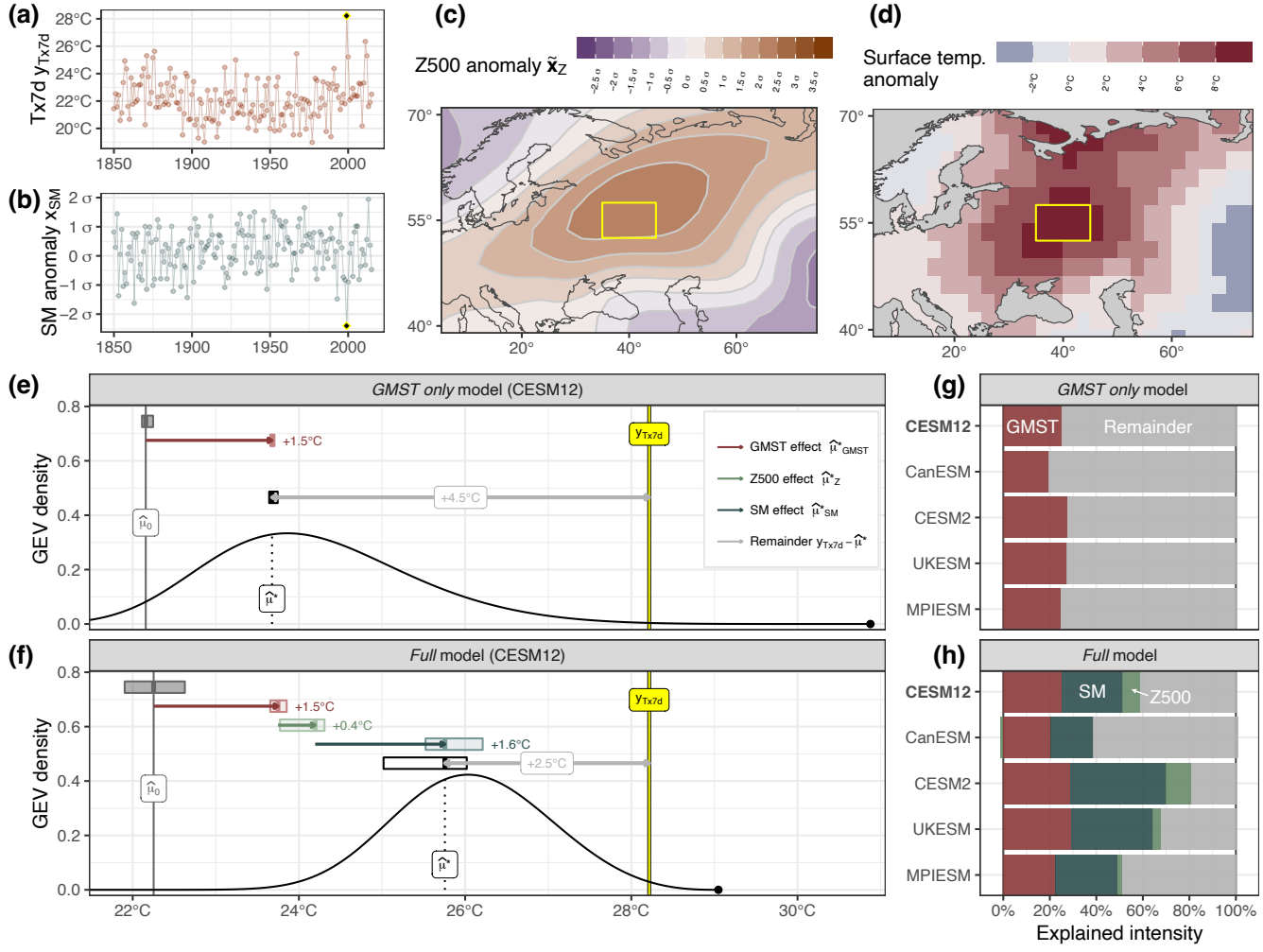

Figure S12. Time series of (a) one-year Tx7d values y_{Tx7d} and (b) corresponding detrended and standardised soil moisture anomalies x_{SM}, the yellow point marks the WRU (within model) heatwave event (CESM12 model year 1999, ensemble member 4). (c) Detrended and standardised Z500 anomaly $\tilde{x}_{\mathrm{Z}}$ and (d) temperature anomaly fields of the respective event.
Conditional GEV densities of (e) the *GMST only* and (f) *full* model with estimated effect sizes (horizontal arrows) and 95 % CI (horizontal bars), event intensity y_{Tx7d} (vertical yellow line), intercept $\hat{\mu}_0$ (vertical grey line) and location parameter $\hat{\mu}^*$ (vertical dotted line). Multi-model assessment of relative effect sizes of (g) the *GMST only* and (h) *full* GEV models.

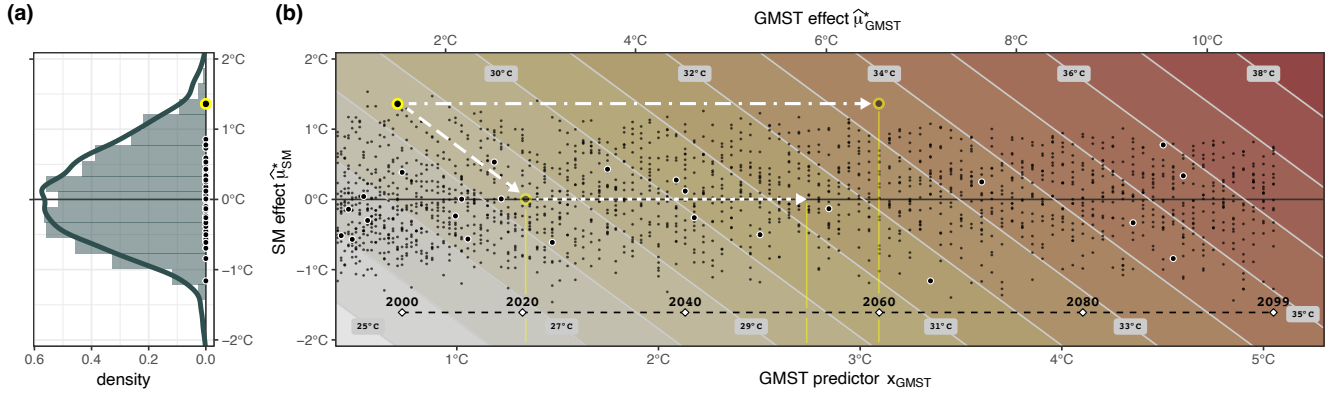

Figure S13. (a) Density of SM effect magnitudes $\widehat{\mu}^*_{\mathrm{SM}}$ on the abscissa in blue (density in CESM12 climate model dataset as histogram, and for all climate model datasets pooled as smoothed density) at the WRU location. Dots indicate corresponding values in CESM12 ensemble member 4 five-year heatwave events. The dot with a yellow stroke corresponds to the 1999 heatwave event.
(b) Event intensity (coloured background) as a function of the GMST predictor x_{GMST} (bottom ordinate) or GMST effect $\widehat{\mu}^*_{\mathrm{GMST}}$ (top ordinate) and SM effect $\widehat{\mu}^*_{\mathrm{SM}}$ (on the abscissa, same as in a), conditional on the 1999 event specific Z500 effect $\widehat{\mu}^*_{\mathrm{Z}} = 0.4\,^{\circ}\mathrm{C}$ and unexplained remainder term of 2.5 °C (c.f. supplementary Fig. S12f). The yellow dot marks the 1999 event under discussion, further dots show events in the CESM12 climate model dataset in terms of the respective SM/GMST effects (white dots are from the same ensemble member, black dots from other ensemble members), whose intensity is not related to the background colouring. The black dashed line at the bottom of the plot indicates the timing when certain GMST levels are reached in the CESM12 large ensemble under RCP 8.5 forcing. White arrows and transparent yellow dots/lines refer to "what if" scenarios discussed in the text.

References

Barriopedro, D., Fischer, E. M., Luterbacher, J., Trigo, R. M., and García-Herrera, R.: The Hot Summer of 2010: Redrawing the Temperature Record Map of Europe, Science, 332, 220–224, https://doi.org/10.1126/science.1201224, 2011.

Bélisle, C. J. P.: Convergence theorems for a class of simulated annealing algorithms on $\mathbb{R}$d, Journal of Applied Probability, 29, 885–895, https://doi.org/10.2307/3214721, 1992.

Ben Alaya, M. A., Zwiers, F., and Zhang, X.: An Evaluation of Block-Maximum-Based Estimation of Very Long Return Period Precipitation Extremes with a Large Ensemble Climate Simulation, Journal of Climate, 33, 6957–6970, https://doi.org/10.1175/JCLI-D-19-0011.1, 2020.

Ben Alaya, M. A., Zwiers, F. W., and Zhang, X.: On estimating long period wind speed return levels from annual maxima, Weather and Climate Extremes, 34, 100 388, https://doi.org/10.1016/j.wace.2021.100388, 2021.

Bieli, M., Pfahl, S., and Wernli, H.: A lagrangian investigation of hot and cold temperature extremes in europe, Quarterly Journal of the Royal Meteorological Society, 141, 98–108, https://doi.org/10.1002/qj.2339, 2015.

Bröcker, J. and Smith, L. A.: Scoring Probabilistic Forecasts: The Importance of Being Proper, Weather and Forecasting, 22, 382–388, https://doi.org/10.1175/WAF966.1, 2007.

Bücher, A., Lilienthal, J., Kinsvater, P., and Fried, R.: Penalized quasi-maximum likelihood estimation for extreme value models with application to flood frequency analysis, Extremes, 24, 325–348, https://doi.org/10.1007/s10687-020-00379-y, 2021.

Cannon, A. J.: A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology, Hydrological Processes, 24, 673–685, https://doi.org/10.1002/hyp.7506, 2010.

Chavez-Demoulin, V. and Davison, A. C.: Generalized additive modelling of sample extremes, Journal of the Royal Statistical Society: Series C (Applied Statistics), 54, 207–222, https://doi.org/10.1111/j.1467-9876.2005.00479.x, 2005.

Christidis, N. and Stott, P. A.: Changes in the geopotential height at 500 hPa under the influence of external climatic forcings, Geophysical Research Letters, 42, 10 798–10 806, https://doi.org/10.1002/2015GL066669, 2015.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I.: STL: A Seasonal-Trend Decomposition Procedure Based on Loess, Journal of Official Statistics, 6, 3–73, http://www.ncbi.nlm.nih.gov/pubmed/2207653, 1990.

Coles, S.: An introduction to statistical modeling of extreme values, Springer series in statistics, London: Springer, London, 3rd print edn., 2001.

Cooley, D.: Return Periods and Return Levels Under Climate Change, in: Extremes in a Changing Climate: Detection, Analysis and Uncertainty, edited by AghaKouchak, A., Easterling, D., Hsu, K., Schubert, S., and Sorooshian, S., chap. 4, pp. 97–114, Springer Netherlands, Dordrecht, Netherlands, https://doi.org/10.1007/978-94-007-4479-0, 2013.

Cooley, D., Hunter, B. D., and Smith, R. L.: Univariate and Multivariate Extremes for the Environmental Sciences, in: Handbook of Environmental and Ecological Statistics, edited by Gelfand, A., Fuentes, M., Hoeting, J. A., and Smith, R. L., chap. 8, pp. 153–180, Chapman and Hall/CRC, Boca Raton : Taylor & Francis, 2018., https://doi.org/10.1201/9781315152509-8, 2019.

Deser, C., Terray, L., and Phillips, A. S.: Forced and internal components of winter air temperature trends over North America during the past 50 years: Mechanisms and implications, Journal of Climate, 29, 2237–2258, https://doi.org/10.1175/JCLI-D-15-0304.1, 2016.

El Adlouni, S., Ouarda, T. B., Zhang, X., Roy, R., and Bobée, B.: Generalized maximum likelihood estimators for the nonstationary generalized extreme value model, Water Resources Research, 43, 1–13, https://doi.org/10.1029/2005WR004545, 2007.

Fischer, E. M., Rajczak, J., and Schär, C.: Changes in European summer temperature variability revisited, Geophysical Research Letters, 39, 1–8, https://doi.org/10.1029/2012GL052730, 2012.

Friederichs, P. and Thorarinsdottir, T. L.: Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction, Environmetrics, 23, 579–594, https://doi.org/10.1002/env.2176, 2012.

Gessner, C., Fischer, E. M., Beyerle, U., and Knutti, R.: Very rare heat extremes: quantifying and understanding using ensemble re-initialization, Journal of Climate, 34, 6619–6634, https://doi.org/10.1175/JCLI-D-20-0916.1, 2021.

Gilleland, E.: Bootstrap Methods for Statistical Inference. Part I: Comparative Forecast Verification for Continuous Variables, Journal of Atmospheric and Oceanic Technology, 37, 2117–2134, https://doi.org/10.1175/JTECH-D-20-0069.1, 2020a.

Gilleland, E.: Bootstrap Methods for Statistical Inference. Part II: Extreme-Value Analysis, Journal of Atmospheric and Oceanic Technology, 37, 2135–2144, https://doi.org/10.1175/JTECH-D-20-0070.1, 2020b.

Gilleland, E. and Katz, R. W.: ExtRemes 2.0: An extreme value analysis package in R, Journal of Statistical Software, 72, https://doi.org/10.18637/jss.v072.i08, 2016.

Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, Journal of the American Statistical Association, 102, 359–378, https://doi.org/10.1198/016214506000001437, 2007.

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness, Journal of the Royal Statistical Society. Series B: Statistical Methodology, 69, 243–268, https://doi.org/10.1111/j.1467-9868.2007.00587.x, 2007.

Greve, P. and Seneviratne, S. I.: Assessment of future changes in water availability and aridity, Geophysical Research Letters, 42, 5493–5499, https://doi.org/10.1002/2015GL064127, 2015.

Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning, Springer Series in Statistics, Springer New York, New York, NY, https://doi.org/10.1007/b94608, 2009.

Huang, W. K., Stein, M. L., McInerney, D. J., Sun, S., and Moyer, E. J.: Estimating changes in temperature extremes from millennial-scale climate simulations using generalized extreme value (GEV) distributions, Advances in Statistical Climatology, Meteorology and Oceanography, 2, 79–103, https://doi.org/10.5194/ascmo-2-79-2016, 2016.

IPCC: Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change,, Tech. rep., World Meteorological Organisation, Geneva, Switzerland, 2018.

Janson, L., Fithian, W., and Hastie, T. J.: Effective degrees of freedom: a flawed metaphor, Biometrika, 102, 479–485, https://doi.org/10.1093/biomet/asv019, 2015.

Jézéquel, A., Yiou, P., and Radanovics, S.: Role of circulation in European heatwaves using flow analogues, Climate Dynamics, 50, 1145–1159, https://doi.org/10.1007/s00382-017-3667-0, 2018.

Jones, P. W.: First- and Second-Order Conservative Remapping Schemes for Grids in Spherical Coordinates, Monthly Weather Review, 127, 2204–2210, https://doi.org/10.1175/1520-0493(1999)127<2204:FASOCR>2.0.CO;2, 1999.

Karas, M., Brzyski, D., Dzemidzic, M., Goñi, J., Kareken, D. A., Randolph, T. W., and Harezlak, J.: Brain Connectivity-Informed Regularization Methods for Regression, Statistics in Biosciences, 11, 47–90, https://doi.org/10.1007/s12561-017-9208-x, 2019.

Karlsson, P. S., Behrenz, L., and Shukur, G.: Performances of Model Selection Criteria When Variables are Ill Conditioned, Computational Economics, 54, 77–98, https://doi.org/10.1007/s10614-017-9682-8, 2019.

Kaufman, S. and Rosset, S.: When does more regularization imply fewer degrees of freedom? Sufficient conditions and counterexamples, Biometrika, 101, 771–784, https://doi.org/10.1093/biomet/asu034, 2014.

Kim, Y.-H., Min, S.-K., Zhang, X., Sillmann, J., and Sandstad, M.: Evaluation of the CMIP6 multi-model ensemble for climate extreme indices, Weather and Climate Extremes, 29, 100 269, https://doi.org/10.1016/j.wace.2020.100269, 2020.

Merrifield, A. L., Simpson, I. R., McKinnon, K. A., Sippel, S., Xie, S. P., and Deser, C.: Local and Nonlocal Land Surface Influence in European Heatwave Initial Condition Ensembles, Geophysical Research Letters, 46, 14 082–14 092, https://doi.org/10.1029/2019GL083945, 2019.

Padrón, R. S., Gudmundsson, L., Decharme, B., Ducharne, A., Lawrence, D. M., Mao, J., Peano, D., Krinner, G., Kim, H., and Seneviratne, S. I.: Observed changes in dry-season water availability attributed to human-induced climate change, Nature Geoscience, 13, 477–481, https://doi.org/10.1038/s41561-020-0594-1, 2020.

Pauli, F. and Coles, S.: Penalized likelihood inference in extreme value analyses, Journal of Applied Statistics, 28, 547–560, https://doi.org/10.1080/02664760120047889, 2001.

Schär, C., Vidale, P. L., Lüthi, D., Frei, C., Häberli, C., Liniger, M. A., and Appenzeller, C.: The role of increasing temperature variability in European summer heatwaves, Nature, 427, 332–336, https://doi.org/10.1038/nature02300, 2004.

Schubert, S. D., Wang, H., Koster, R. D., Suarez, M. J., and Groisman, P. Y.: Northern Eurasian heat waves and droughts, Journal of Climate, 27, 3169–3207, https://doi.org/10.1175/JCLI-D-13-00360.1, 2014.

Schumacher, D. L., Keune, J., van Heerwaarden, C. C., Vilà-Guerau de Arellano, J., Teuling, A. J., and Miralles, D. G.: Amplification of mega-heatwaves through heat torrents fuelled by upwind drought, Nature Geoscience, https://doi.org/10.1038/s41561-019-0431-6, 2019.

Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., Orlowsky, B., and Teuling, A. J.: Investigating soil moisture-climate interactions in a changing climate: A review, Earth-Science Reviews, 99, 125–161, https://doi.org/10.1016/j.earscirev.2010.02.004, 2010.

Sillmann, J., Kharin, V. V., Zhang, X., Zwiers, F. W., and Bronaugh, D.: Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate, Journal of Geophysical Research: Atmospheres, 118, 1716–1733, https://doi.org/10.1002/jgrd.50203, 2013.

Sippel, S., Meinshausen, N., Merrifield, A., Lehner, F., Pendergrass, A. G., Fischer, E., and Knutti, R.: Uncovering the forced climate response from a single ensemble member using statistical learning, Journal of Climate, pp. 5677–5699, https://doi.org/10.1175/jcli-d-18-0882.1, 2019.

Suarez-Gutierrez, L., Maher, N., and Milisnki, S.: Evaluating the internal variability and forced response in Large Ensembles, US Clivar Variations, 18, 27–35, 2020.

Tebaldi, C., Dorheim, K., Wehner, M., and Leung, R.: Extreme metrics from large ensembles: investigating the effects of ensemble size on their estimates, Earth System Dynamics, 12, 1427–1501, https://doi.org/10.5194/esd-12-1427-2021, 2021.

Terray, L.: A dynamical adjustment perspective on extreme event attribution, Weather and Climate Dynamics, 2, 971–989, https://doi.org/10.5194/wcd-2-971-2021, 2021.

Thorarinsdottir, T. L., Sillmann, J., Haugen, M., Gissibl, N., and Sandstad, M.: Evaluation of CMIP5 and CMIP6 simulations of historical surface air temperature extremes using proper evaluation methods, Environmental Research Letters, 15, 124 041, https://doi.org/10.1088/1748-9326/abc778, 2020.

Tokarska, K. B., Stolpe, M. B., Sippel, S., Fischer, E. M., Smith, C. J., Lehner, F., and Knutti, R.: Past warming trend constrains future warming in CMIP6 models, Science Advances, 6, 1–14, https://doi.org/10.1126/sciadv.aaz9549, 2020.

Trenberth, K. E. and Fasullo, J. T.: Climate extremes and climate change: The Russian heat wave and other climate extremes of 2010, Journal of Geophysical Research Atmospheres, 117, 1–12, https://doi.org/10.1029/2012JD018020, 2012.

Vautard, R., Yiou, P., Otto, F., Stott, P., Christidis, N., Van Oldenborgh, G. J., and Schaller, N.: Attribution of human-induced dynamical and thermodynamical contributions in extreme weather events, Environmental Research Letters, 11, https://doi.org/10.1088/1748-9326/11/11/114009, 2016.

Watanabe, M., Shiogama, H., Imada, Y., Mori, M., Ishii, M., and Kimoto, M.: Event Attribution of the August 2010 Russian Heat Wave, SOLA, 9, 65–68, https://doi.org/10.2151/sola.2013-015, 2013.

Wehner, M., Gleckler, P., and Lee, J.: 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, 100 283, https://doi.org/10.1016/j.wace.2020.100283, 2020.

Wehrli, K., Guillod, B. P., Hauser, M., Leclair, M., and Seneviratne, S. I.: Identifying Key Driving Processes of Major Recent Heat Waves, Journal of Geophysical Research: Atmospheres, 124, 11 746–11 765, https://doi.org/10.1029/2019JD030635, 2019.

Wilks, D. S.: Statistical methods in the atmospheric sciences, Elsevier, Oxford, United Kingdom, third edn., 2011.

Zhou, S. and Yuan, X.: Upwind Droughts Enhance Half of the Heatwaves Over North China, Geophysical Research Letters, 49, https://doi.org/10.1029/2021GL096639, 2022.

Supplementary tables and figures

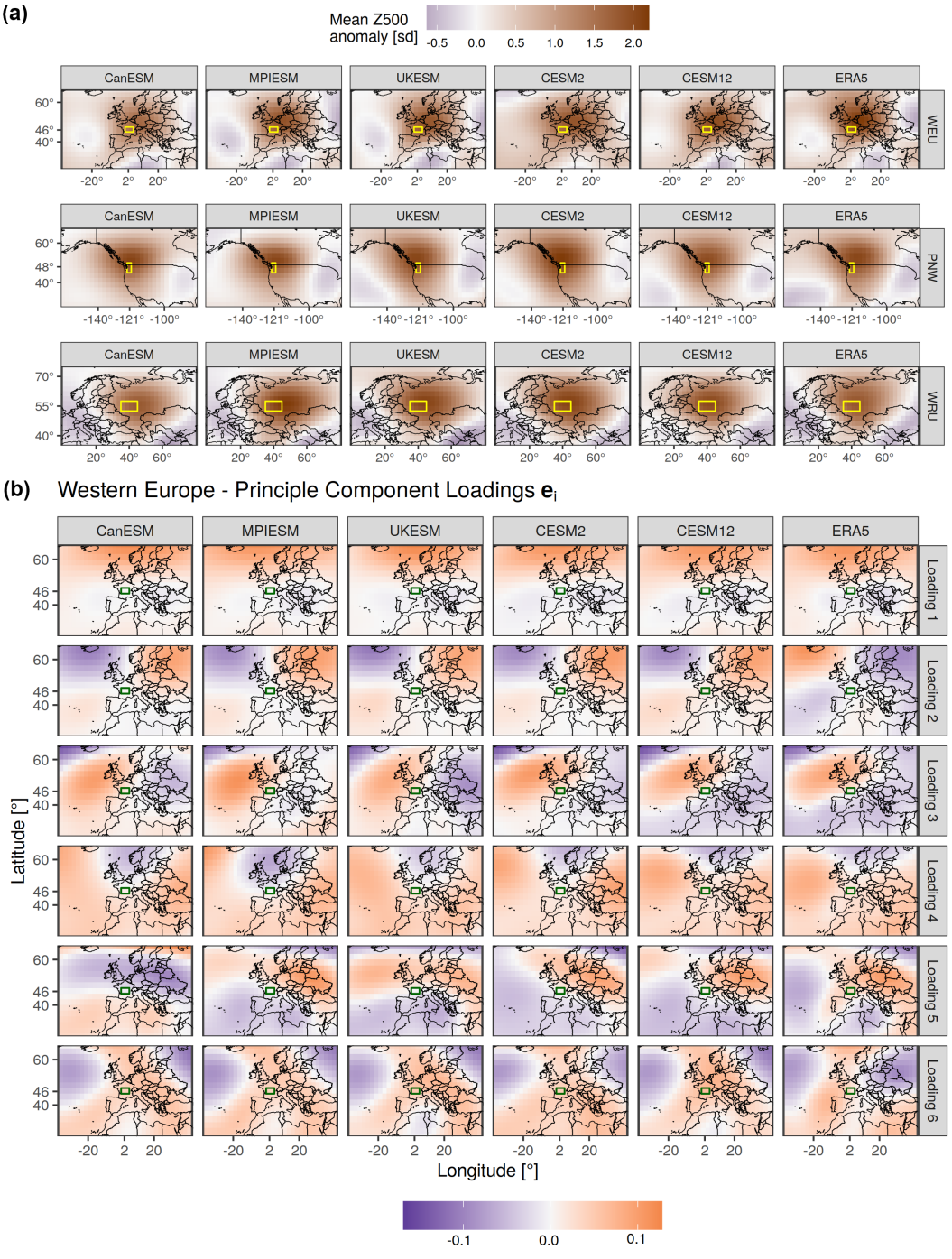

Figure S14. (a) Average and (b) leading six PC loading patterns of geopotential height anomaly fields $\bar{\boldsymbol{x}}_Z$ (before standardisation) during five-year BM heatwave events across climate model/reanalysis datasets.

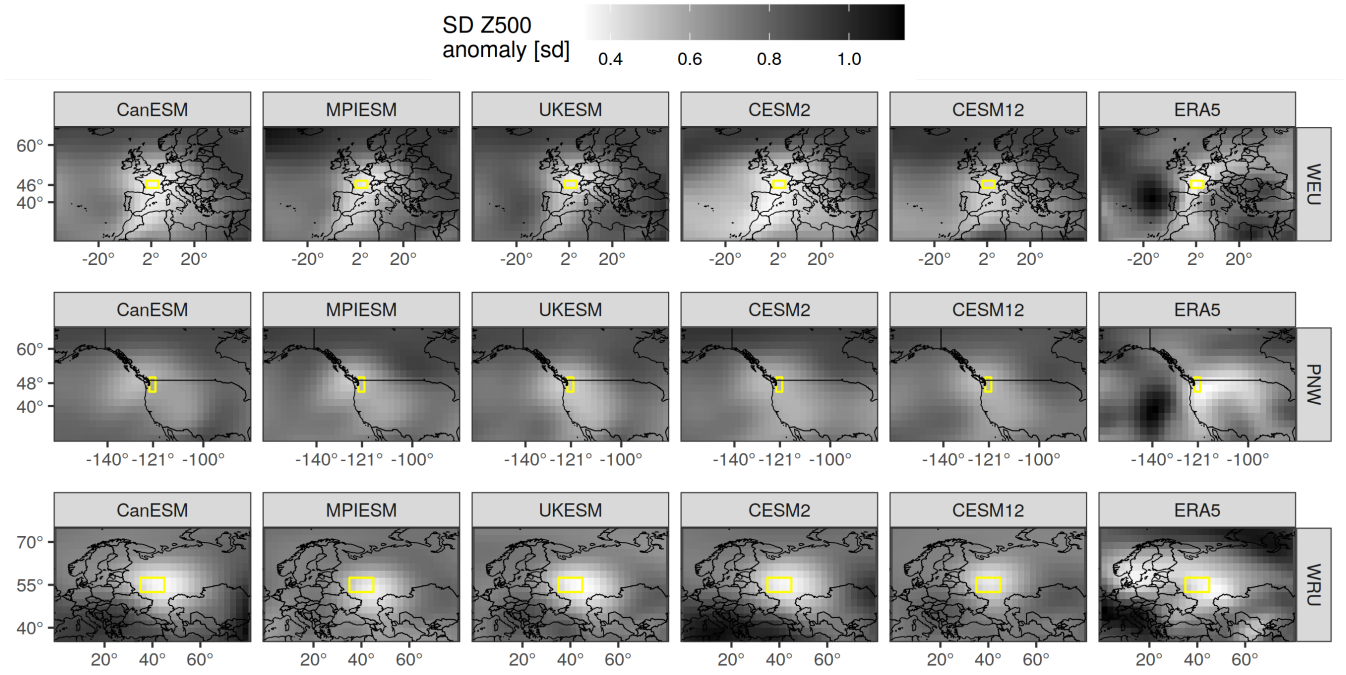

Figure S15. Variance (in standard deviations) of geopotential height anomaly fields $\mathrm{sd}(\boldsymbol{x}_Z)$ during five-year BM heatwave events across different climate model/reanalysis datasets.

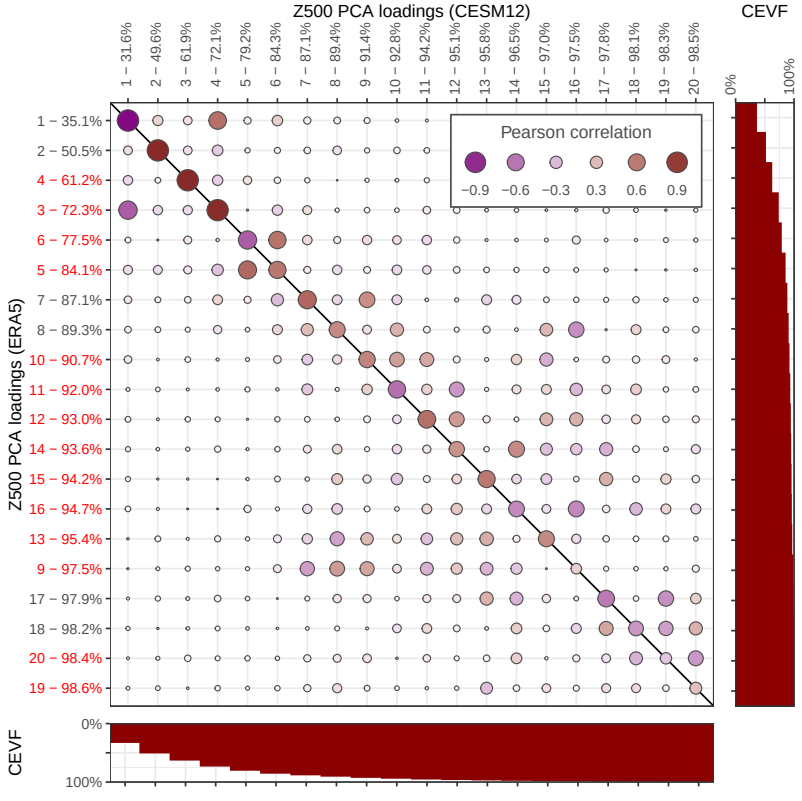

Figure S16. Pattern correlation of the leading 20 CESM12/ERA5 PCs of $\tilde{\boldsymbol{x}}_Z$ during five-year BM heatwave events. The bottom and right sub-figures show the cumulative explained variance fraction (CEVF). Red indices show where the PC order has been adjusted.

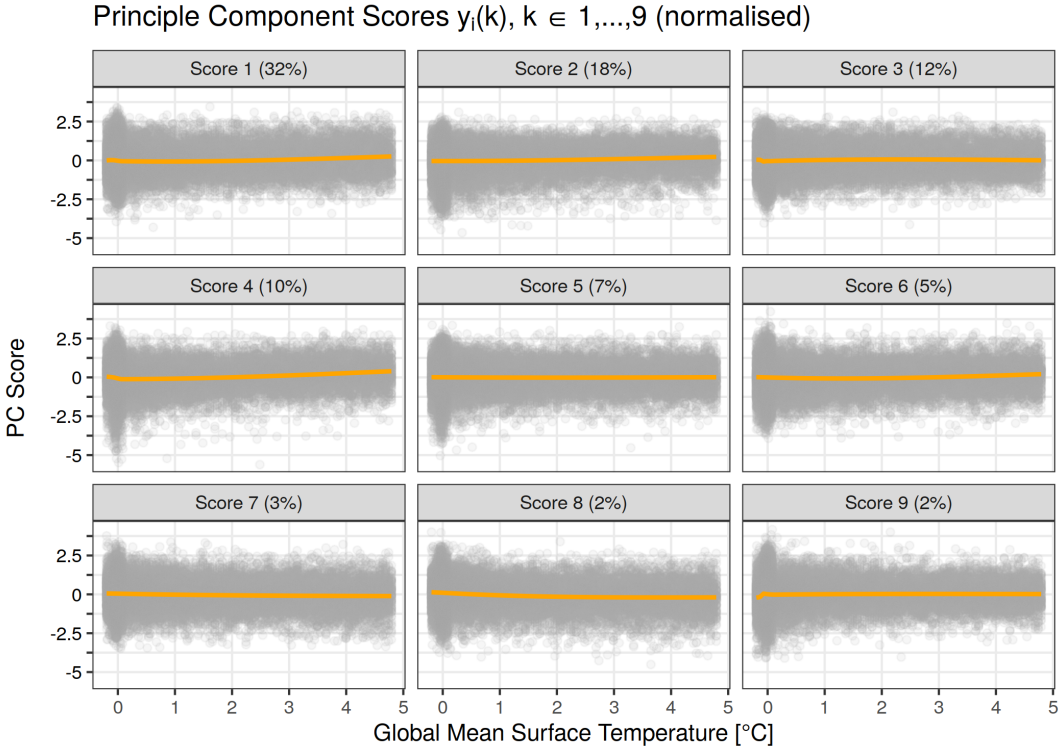

Figure S17. Leading nine CESM12 PCs (normalised) of $\tilde{\boldsymbol{x}}_Z$ during five-year BM heatwave events as function of GMST. In orange a smoothed trend line.

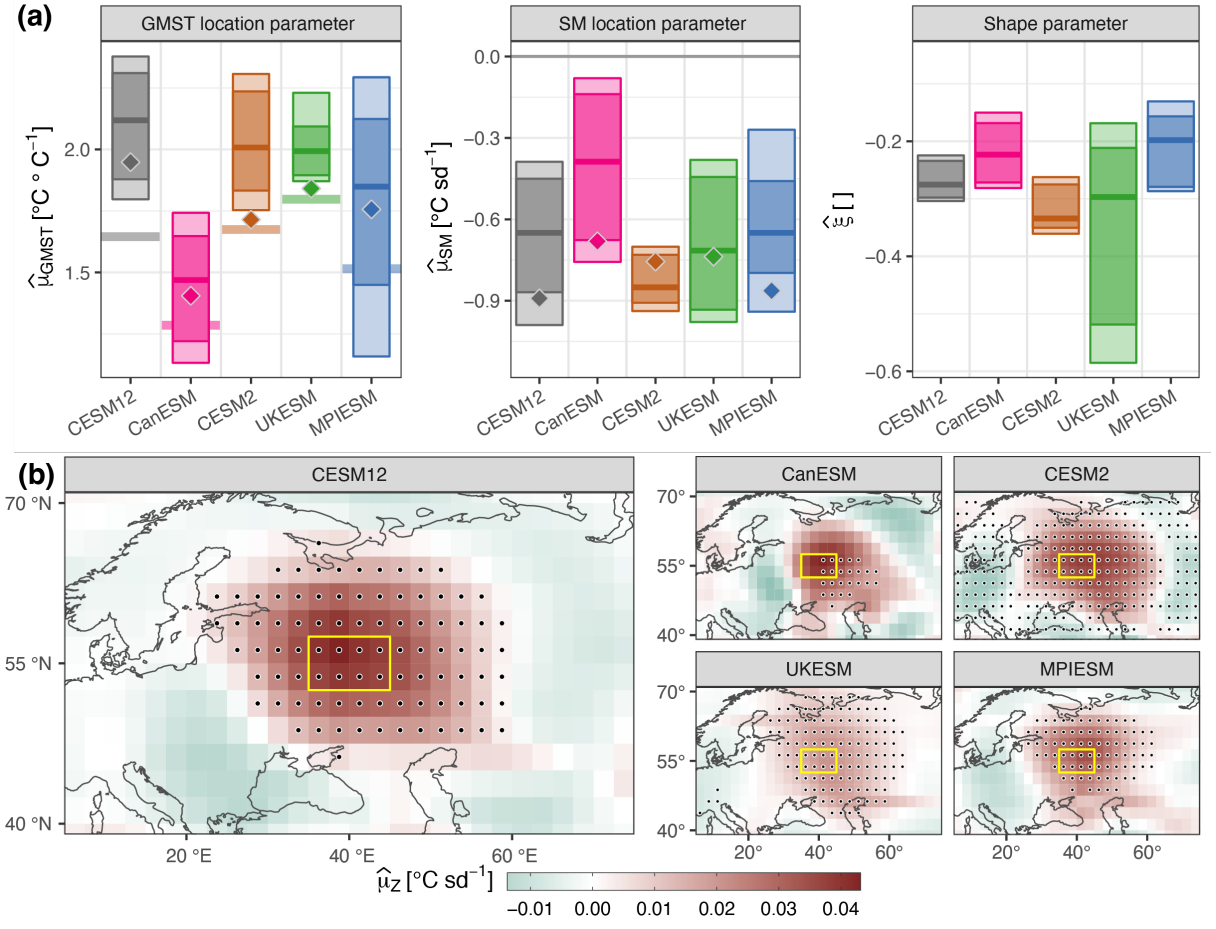

Figure S18. GEV parameter estimates at the WRU location, analogous to Fig. 3.

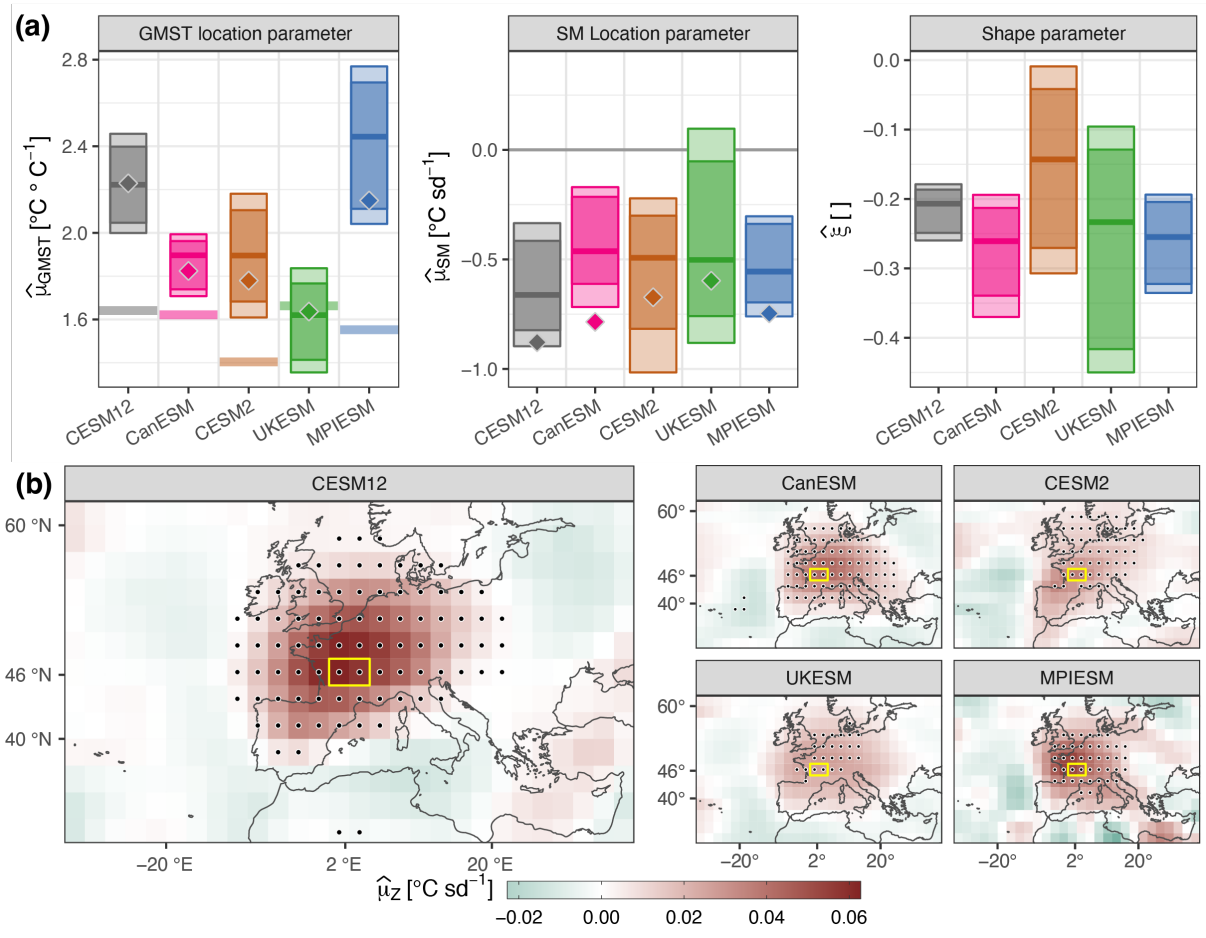

Figure S19. GEV parameter estimates at the WEU location, analogous to Fig. 3.

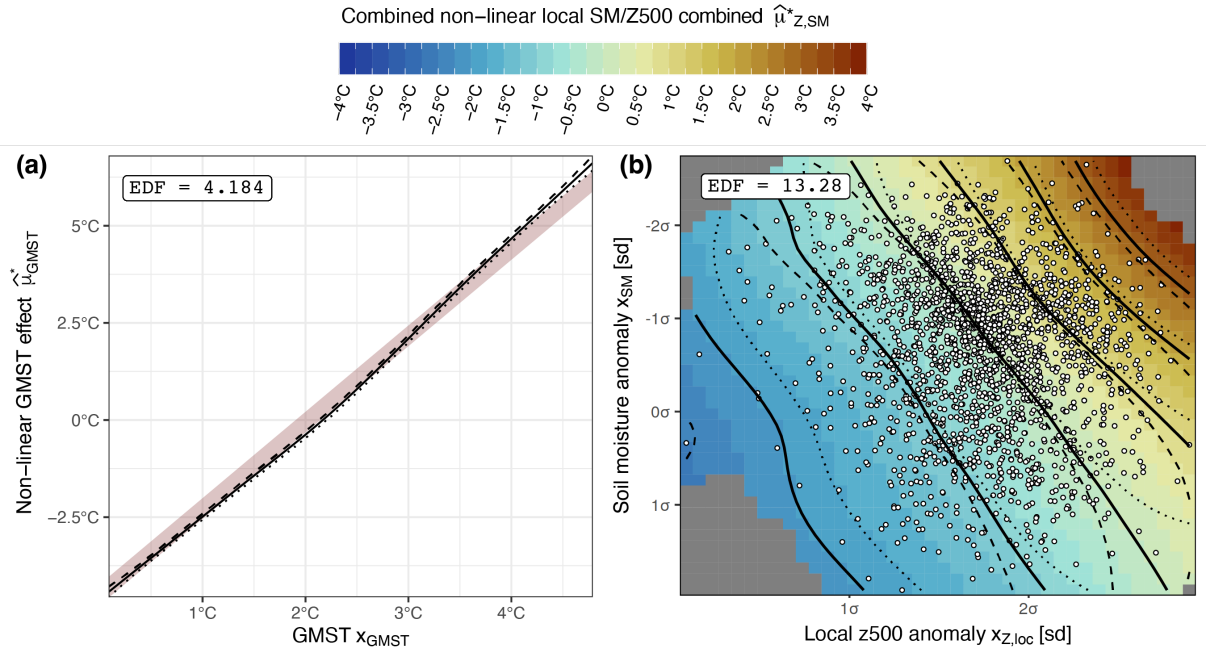

Figure S20. (a) Predicted non-linear GMST effect with 99 % CI (lower dotted line, upper dashed line) based on a smoothing spline basis function. (b) Predicted joint SM and local Z500 effect with 99 % CI (dotted, solid and dashed contour lines), based on a full tensor product smooth. The white dots mark input data points, used for the estimation. Corresponding effective degree-of-freedom measures are provided in the top-left corners.

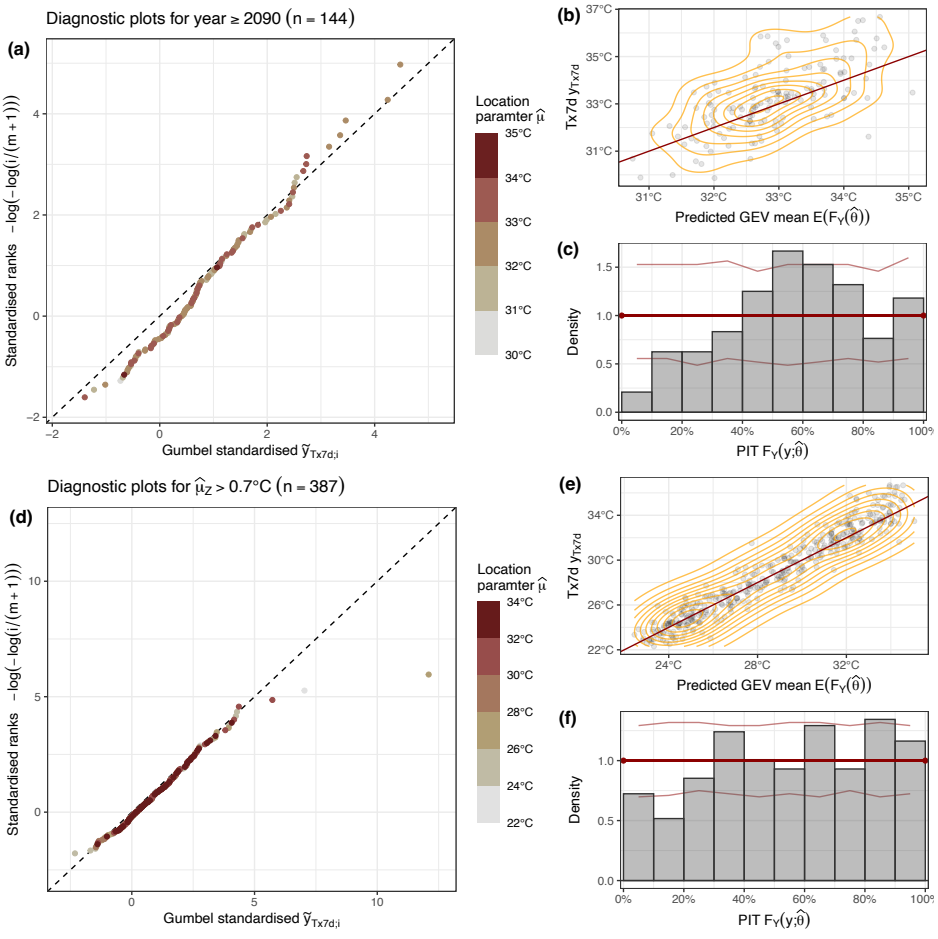

Figure S21. Model diagnostic plots for GEV model fits to CESM12 climate model dataset as in Fig. S8, but for testing data (a-c) and strong Z500 forcing data points (d-f).

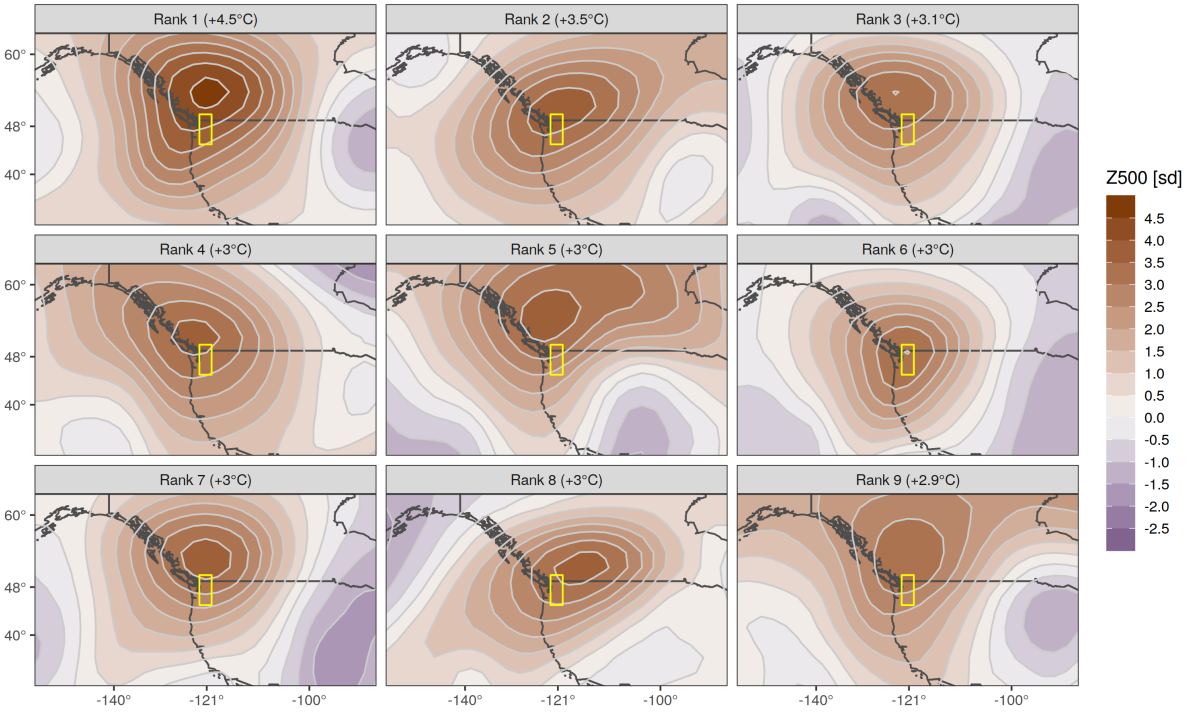

Figure S22. Z500 anomaly fields across all climate model datasets with the highest Z500 effect (in brackets) according to the CESM1.2 based GEV model.

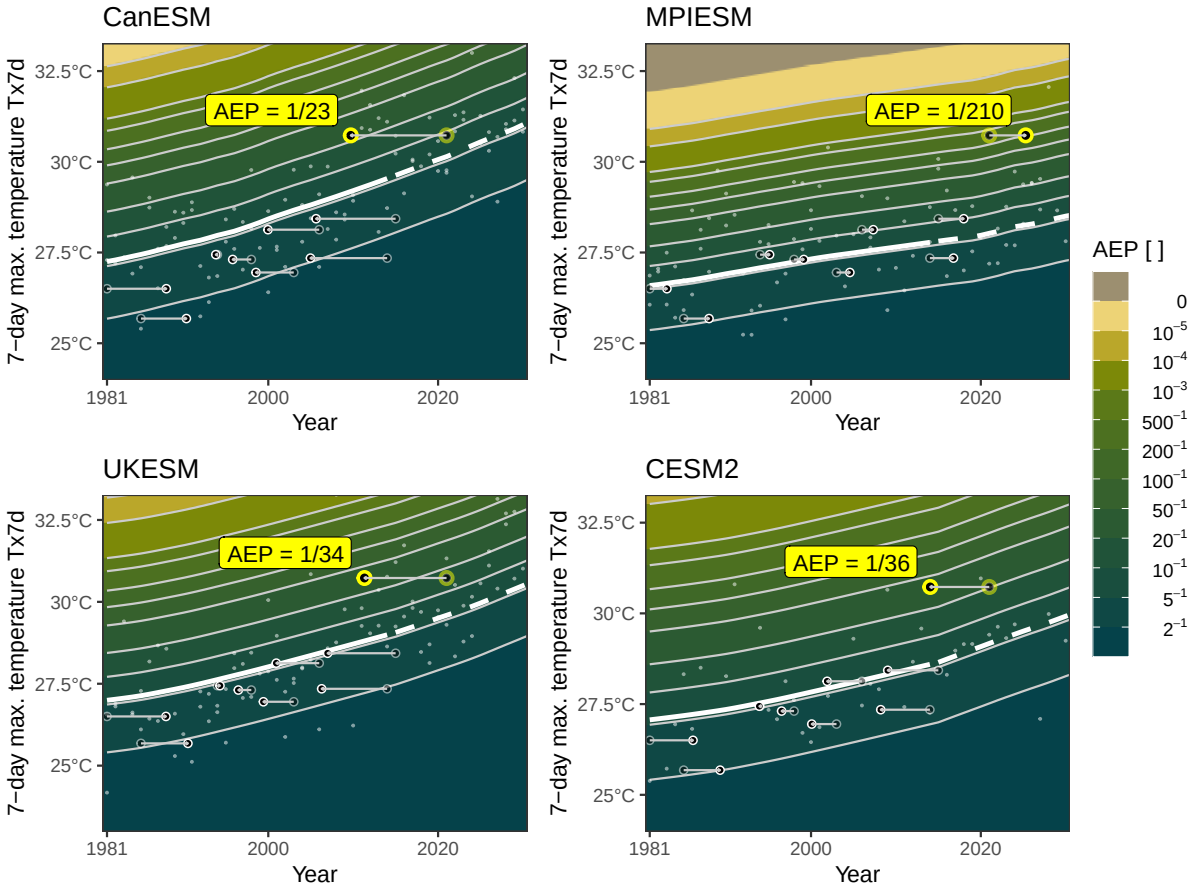

Figure S23. Annual exceedance probability of intensities y_{Tx7d} (abscissa) as function of model (panels) year (ordinate), as in Fig. 7