Supplement of Earth Syst. Dynam., 14, 81–100, 2023
https://doi.org/10.5194/esd-14-81-2023-supplement

Earth System Dynamics
Open Access
EGU

Supplement of

Robust global detection of forced changes in mean and extreme precipitation despite observational disagreement on the magnitude of change

Iris Elisabeth de Vries et al.

Correspondence to: Iris Elisabeth de Vries (iris.devries@env.ethz.ch)

S1 Methodological details

S1.1 Data

The CMIP6 models and members used for ridge regression (RR) are listed in table S1. Historical and SSP245 scenario runs of these models are used, and piControl for the selection as indicated in the last column.

SI Table S1: CMIP6 models and members used for RR model training and model forced response estimation. Models for which 450 years of unforced piControl data was available are indicated.

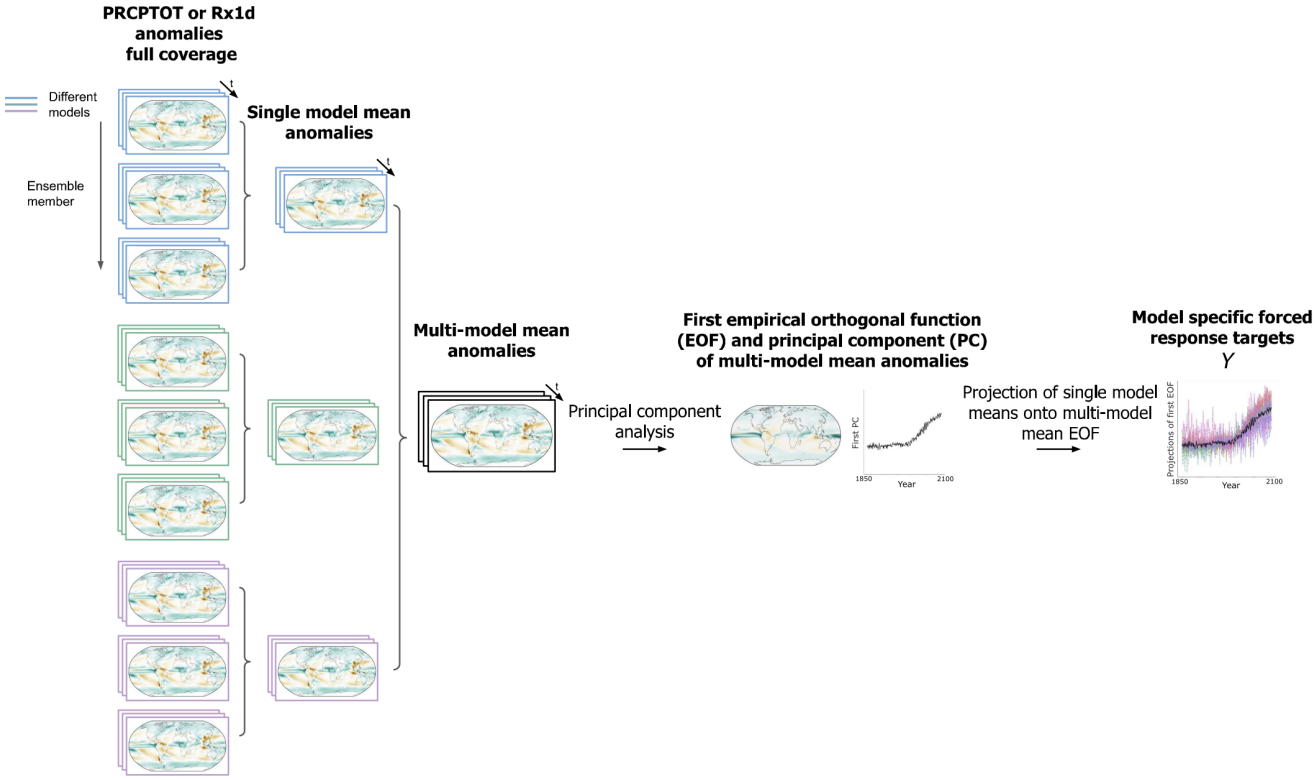
Model	**Member**	**piControl y/n**
ACCESS-CM2	r1i1p1f1, r2i1p1f1, r3i1p1f1	y
ACCESS-ESM1-5	r10i1p1f1, r15i1p1f1, r1i1p1f1	y
CanESM5	r10i1p1f1, r10i1p2f1, r11i1p1f1	y
EC-Earth3	r10i1p1f1, r12i1p1f1, r14i1p1f1	n
EC-Earth3-Veg	r1i1p1f1, r2i1p1f1, r3i1p1f1	n
FGOALS-g3	r1i1p1f1, r3i1p1f1, r4i1p1f1	n
IPSL-CM6A-LR	r10i1p1f1, r11i1p1f1, r14i1p1f1	y
KACE-1-0-G	r1i1p1f1, r2i1p1f1, r3i1p1f1	y
MIROC-ES2L	r10i1p1f2, r11i1p1f2, r12i1p1f2	y
MIROC6	r1i1p1f1, r2i1p1f1, r3i1p1f1	y
MPI-ESM1-2-LR	r10i1p1f1, r1i1p1f1, r2i1p1f1	y
NorESM2-LM	r1i1p1f1, r2i1p1f1, r3i1p1f1	y
UKESM1-0-LL	r10i1p1f2, r11i1p1f2, r12i1p1f2	y

S1.2 Ridge regression forced response targets

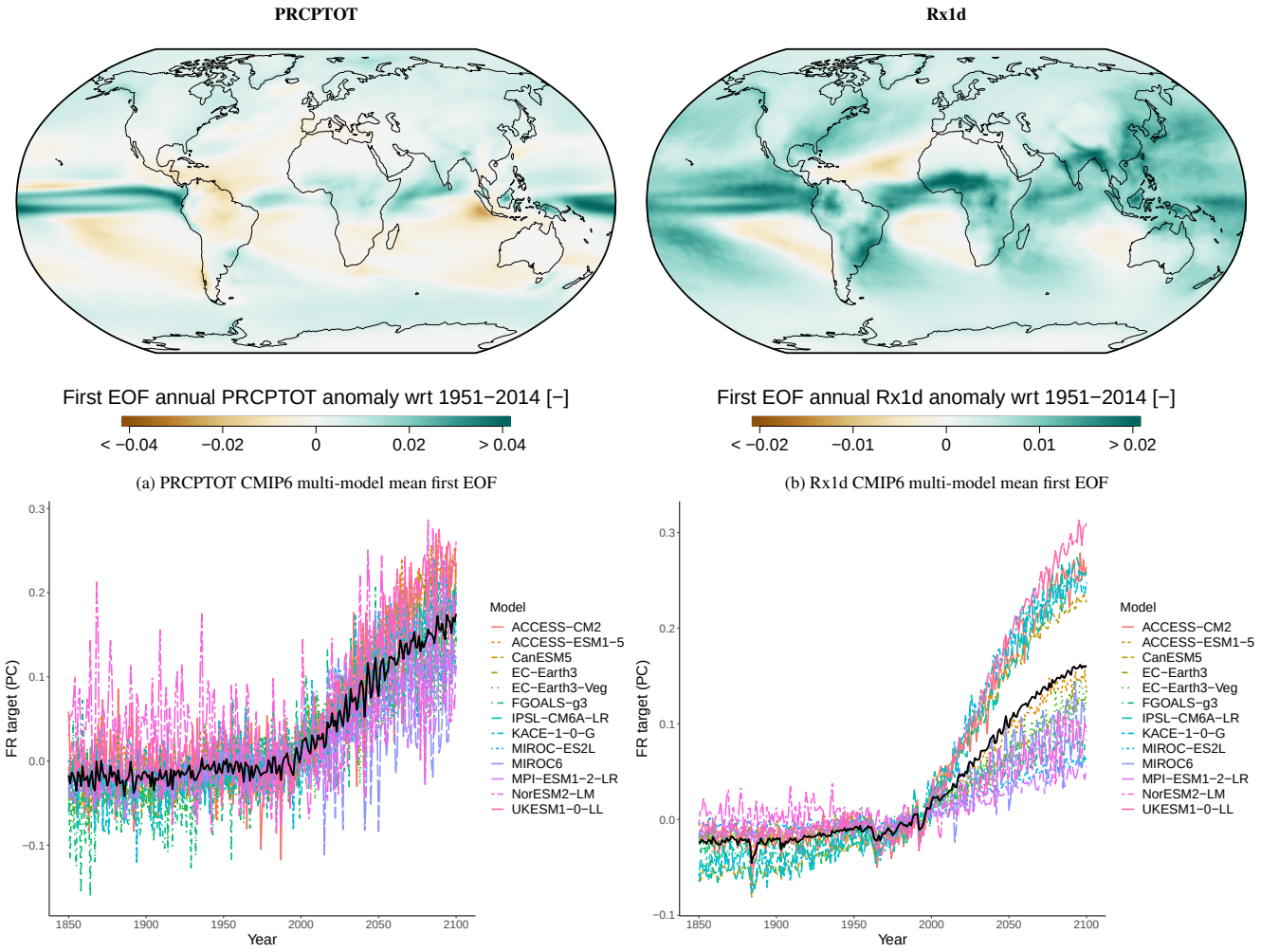
Figure S1 shows a visual representation of the steps taken to construct the targets for the ridge regression models, as described in the methods section of the main text.

Figure S2 shows the first empirical orthogonal function (EOF) of the multi-model mean of PRCPTOT (left) and Rx1d (right) over the full historical and SSP245 future period (1850-2100). The corresponding principal components (PCs) are shown in the bottom panel, where the black line represents the multi-model mean principal component, and the coloured lines the projection of the shown EOF onto individual model ensemble means. These coloured lines make up the effective forced response (FR) targets in the RR training procedure. The PRCPTOT targets are particularly noisy, which is found to be induced by tropical variability primarily, as zonal-region EOFs that exclude the tropics show less variable behaviour. This is due to the high contribution of the tropics to total annual precipitation, and the large variations in the tropics due to e.g. ENSO and variations in the location of the ITCZ.

The correlations between the EOF-based targets and the global means are shown per model in figure S3. Although the correlations are not perfect due to higher spread of the EOF-based targets, they still show large values. In combination with the pattern information enclosed in the EOF, this suggests the EOF-based targets are a suitable choice.



SI Figure S1: Flowchart of the data processing steps required to construct the target variables for the ridge regression model



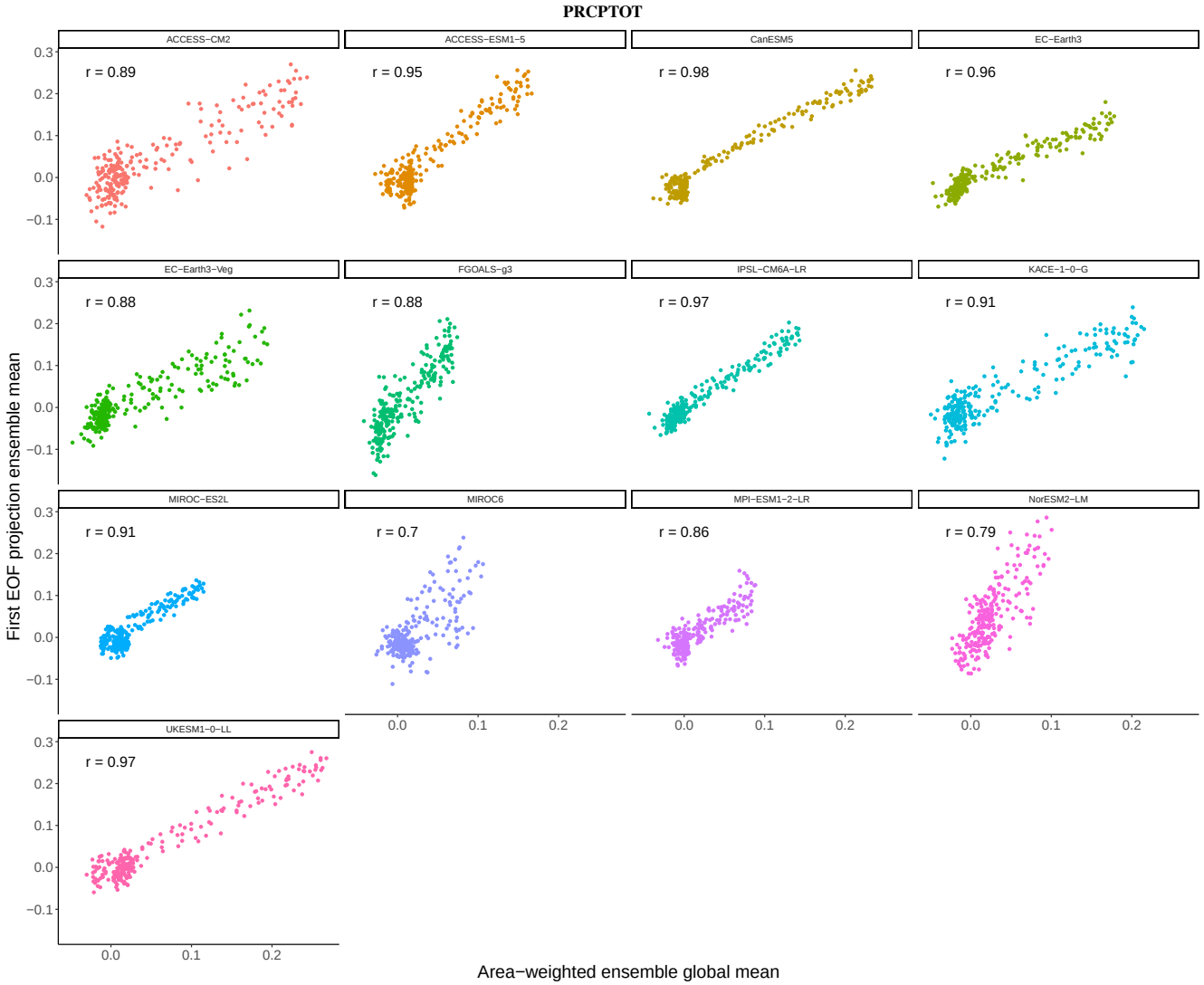
(a) PRCPTOT CMIP6 multi-model mean first EOF

(b) Rx1d CMIP6 multi-model mean first EOF

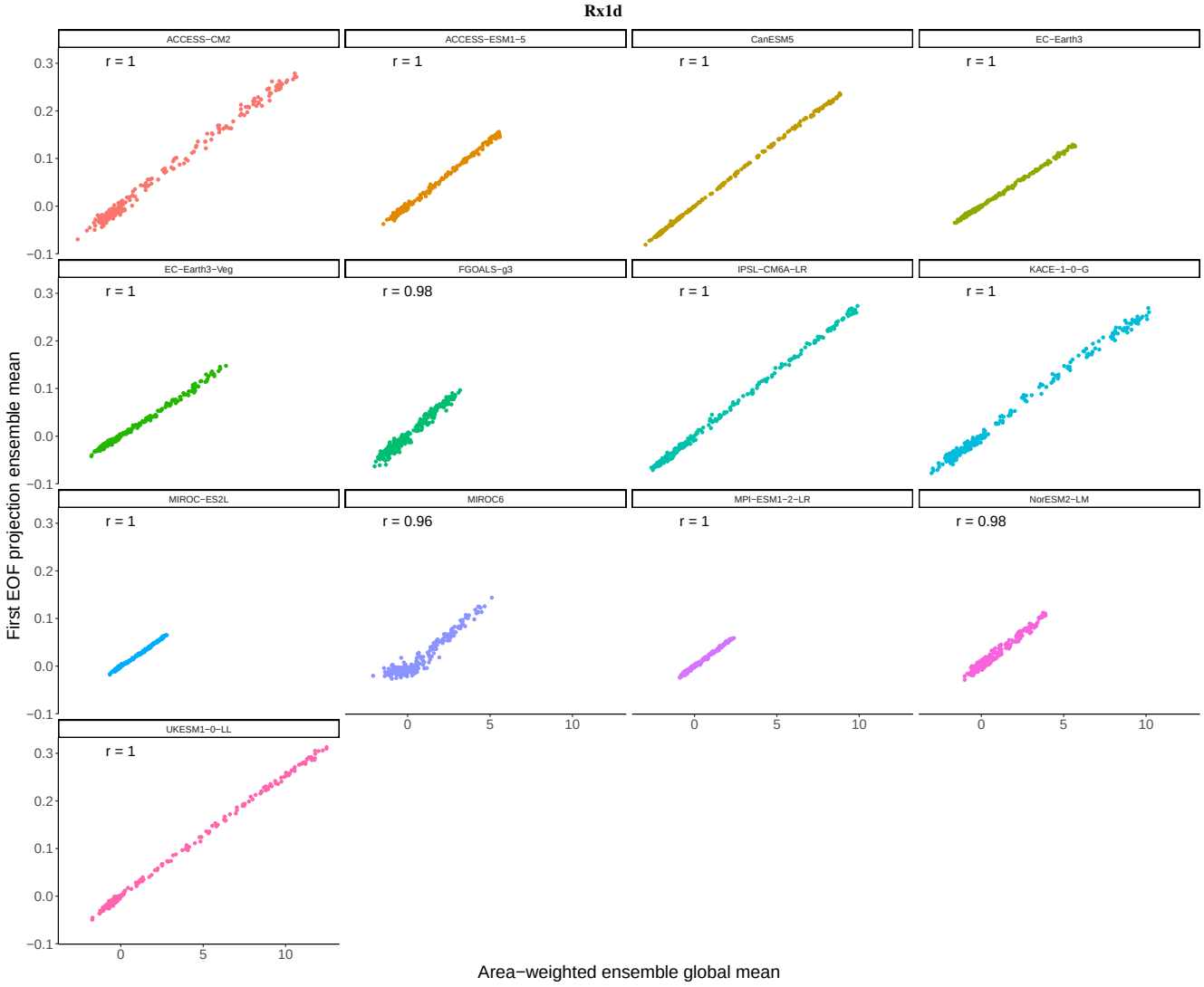
(c) PRCPTOT multi-model mean first principal component and projections onto ensemble means

(d) Rx1d multi-model mean first principal component and projections onto ensemble means

SI Figure S2: First EOF patterns for CMIP6 multi-model mean PRCPTOT (a) and Rx1d (b) over the 1850-2100 period with historical orcing up to 2014 and SSP245 thereafter (Eyring et al., 2016). Corresponding multi-model mean (black) and model ensemble mean (coloured) principal component timeseries (PCs) are shown in c and d. Model ensemble mean principal components are the projection of the multi-model mean EOF (shown) onto individual model ensemble means. These serve as targets in the RR training procedure.



(a) PRCPTOT correlation EOF-based target with global mean



(b) Rx1d correlation EOF-based target with global mean

SI Figure S3: Correlations of model-specific EOF-based targets and area-weights global means, both based on model ensemble means of the models indicated in the subplot headers. Numbers in the upper left corners indicate Pearson correlation coefficients.

S1.3 Ridge regression details

Lambda selection

As mentioned in the main text, the regularisation parameter λ is equal to λ_{sel} in the default case we show. This λ selection is based on the consideration of three possible λs in the optimisation process. These three options depend on the cross-validated error (CVE), or on the post-crossvalidation mean squared error w.r.t. the multi-model mean forced response best estimate (FRBE), referred to as MME, and defined as in equation 1:

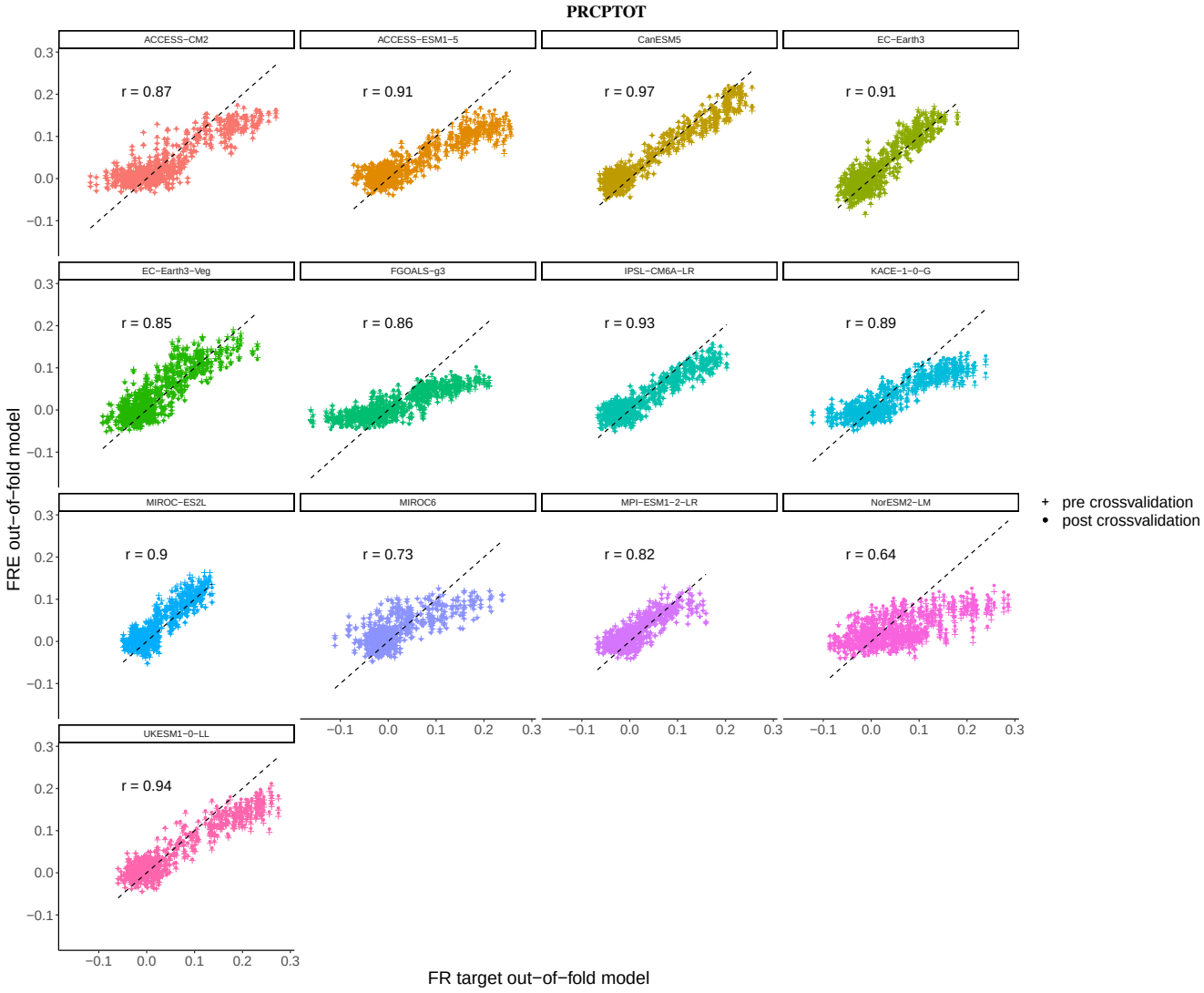
$$\sum_{i=1}^{N} \frac{(\hat{y} - \text{FRBE})^2}{N} \tag{1}$$

The CVE represents the mean squared error of out-of-fold predictions – i.e. the mean squared error of the forced response predicted for a model that was not included in RR training. The λ that results in the smallest CVE is λ_{min}. A common method for λ-selection, however, is to choose the largest λ (more regularisation) with a CVE within one standard error (SE) from the minimum CVE (Friedman et al., 2010; Simon et al., 2011). This λ option is referred to as λ_{1se}. The MME is defined as the mean squared error of all model predictions made with the final RR model, i.e. after cross validation, w.r.t. the multi-model mean first PC: the forced response best estimate. This error thus represents the ability of the RR model to predict one common target – the mean of the training targets – from data from different climate models. It demands relatively high generalisability and thus high regularisation, which is expected to be beneficial when applying the model to observations. The λ that leads to the smallest MME is referred to as λ_{MM}.

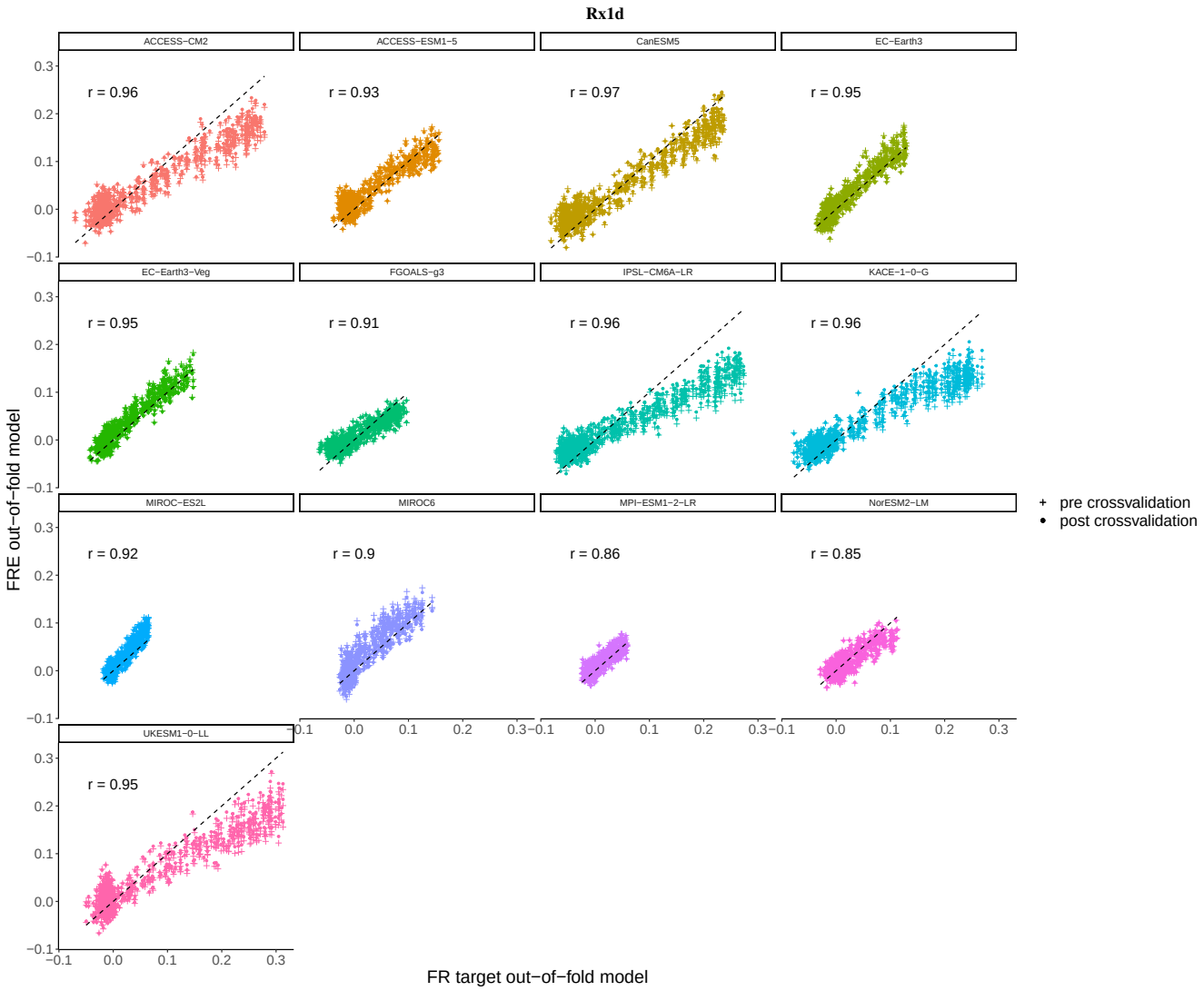
We reason that the most regularised RR model with good performance is a good choice for the detection model, as mentioned in the main text. As both λ_{1se} and λ_{MM} lead to generalisable models and perform well, we select the highest of these two (this differs per case) as our default λ_{sel}.

Cross-validation and application

As discussed in the main text, RR models are applied to the same model data as they have also been trained on using cross-validation. To validate that this application does not significantly influence the model forced response estimates (forced response estimates), and therefore does not jeopardise the relevance of the model forced response estimates, we show model specific correlation plots in figure S4 that include both the pre-crossvalidated forced response estimates (predicted model *not* in training: out-of-fold prediction) and the post-crossvalidated ones (predicted model in training: in-fold prediction). Besides the comparison of in-fold versus out-of-fold prediction, the correlation plots also show the performance of the RR model for ensemble mean forced response prediction in individual models in general. Clearly, the effect of the models being seen in the training is negligible, judging from the similarity of pre- and post-crossvalidation results. The numbers in the upper right corner indicate the Pearson correlation coefficient of the post-crossvalidated predictions with their model specific targets. Note that the horizontal spread of the point clouds is quite large due to the high variability of the EOF-based targets (figures S2c and S2d. Nonetheless, correlations are high, indicating good performance and generality of the RR models for model forced response prediction, although a few individual models have particularly high target spread and/or trends and therefore lower correlations.

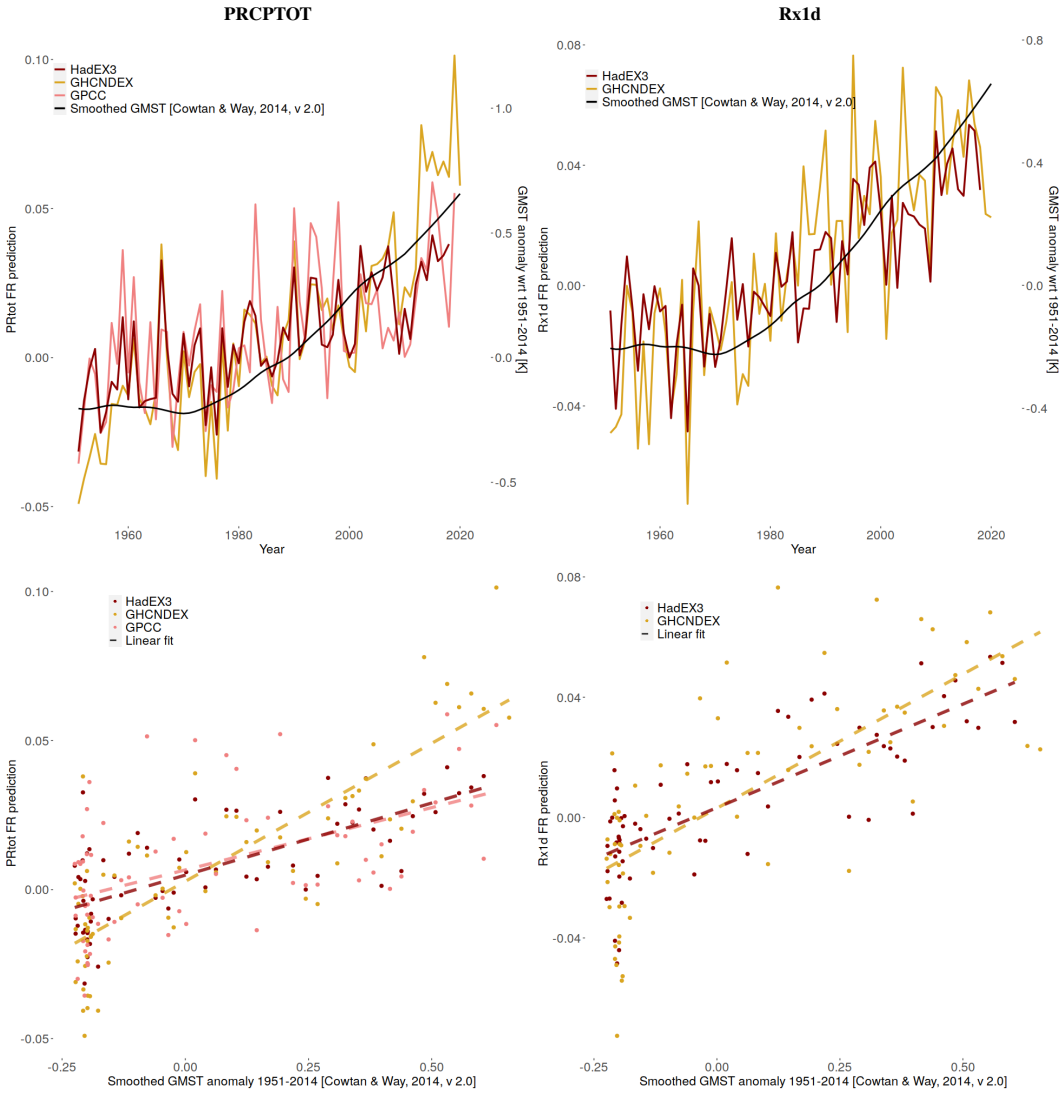


(a) PRCPTOT correlation of EOF-based target with prediction



(b) Rx1d correlation of EOF-based target with prediction

SI Figure S4: Correlations of model-specific EOF-based targets and the forced response estimates obtained from applying the RR model to individual model realisations. The forced response estimates are shown for RR models applied in-fold: i.e. RR models which have been trained and validated on all models (post-crossvalidation), and also for RR-models which have been trained on all-but-one model and are applied out-of-fold, to the model not seen in training (pre-crossvalidation). Numbers in the upper left corners indicate Pearson correlation coefficients.



(a) PRCPTOT correspondence of smoothed GMST and forced response estimates

(b) Rx1d correspondence of smoothed GMST and forced response estimates

SI Figure S5: Top panel: correspondence in shape between observed forced response estimates (coloured lines) and smoothed observed GMST (black line) from Cowtan and Way (2014) in PRCPTOT (a) and Rx1d (b) as a function of year. Bottom panel: PRCPTOT and Rx1d forced response estimates as a function of smoothed GMST, including linear fits (dashed).

Signal-to-noise ratio determination

The relationship between the 21-year LOWESS-filtered global mean surface temperature(GMST) and the forced response estimates for the default PRCPTOT and Rx1d cases are given in figure S5. Here, the top panel shows qualitatively how the GMST and the PRCPTOT and Rx1d forced response estimates are proportional to one another, particularly for Rx1d (right). The bottom plot shows the linear fit of PRCPTOT and Rx1d onto smoothed GMST, used for time of emergence assessment. Note that the GMST curve does not exactly correspond to any of the forced response estimate fits: the forced response estimate fit onto GMST differs between the different datasets. The GMST values shown here are scaled by adjusting the right y-axis manually for visual purposes only, to compare the general long term trends.

S2 Additions to section 3

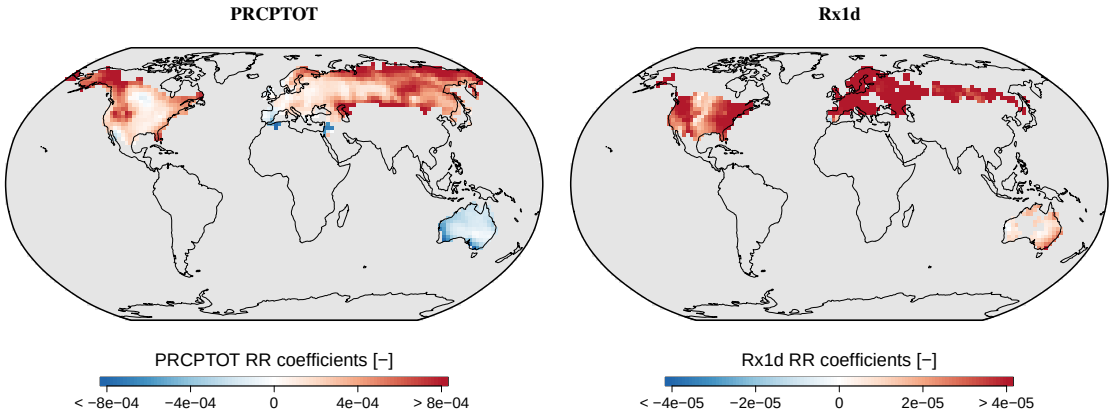
S2.1 Observational dataset and residual consistency

Default case

Figure S6 shows the RR fingerprints for the two observational datasets not shown in the main text: GHCNDEX and GPCC. When compared with figure 2 in the main text, the similarities in coefficient signs are evident. The coverage map of GPCC can be seen to be more scattered, which might interfere to some degree with the extraction of larger-scale patterns using regularised regression.

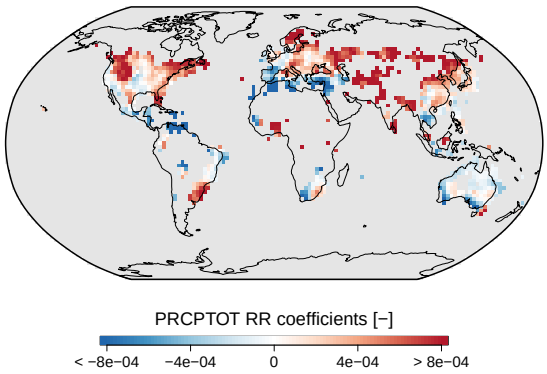
Figure S7 shows the standard deviations of the residuals of the linear trend fits to the forced response estimates over the full period 1951-2014. The standard deviation of the residuals for the observational datasets are shown as vertical lines. For the model forced response estimates, slightly smoothed probability density plots of the residuals standard deviation for all individual realisations are shown, for each coverage mask and for both the forced and the piControl conditions. For both PRCPTOT and Rx1d, all observational datasets' residuals standard deviations lie within the model-derived distributions on their corresponding coverage masks, which validates the consistency of the method used in its application to models and observations. In addition, we also see that the coverage mask influences the spread in a way that corresponds to observations – e.g. GHCNDEX observed forced response estimate residuals are higher than for HadEX3, and model forced response estimate predictions on the GHCNDEX coverage mask also show larger residuals than model forced response estimates on the HadEX3 coverage mask.

Generally, the residuals of model forced response estimates and observed forced response estimates agree better for Rx1d than for PRCPTOT, which is in line with the higher uncertainty in PRCPTOT detection seen throughout this study. For PRCPTOT we also see generally lower forced response estimate residuals for piControl compared to forced model forced response estimates, whereas Rx1d forced response estimate residuals of piControl and forced forced response estimates are more consistent. This potentially results from an already measurable increase in variability in PRCPTOT in the forced simulations.



(a) GHCNDEX-mask PRCPTOT annual RR fingerprint

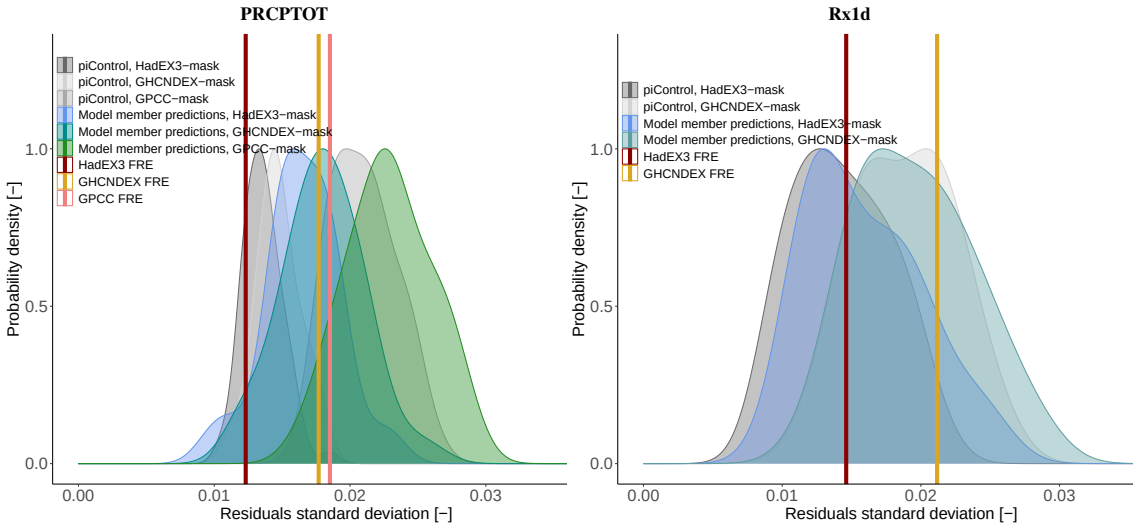
(b) GHCNDEX-mask Rx1d annual RR fingerprint



(c) GPCC-mask PRCPTOT annual RR fingerprint

(d)

SI Figure S6: RR fingerprints for PRCPTOT (a, c), and Rx1d (b) as in main figure 2, for resolution and coverage masks of GHCNDEX and GPCC



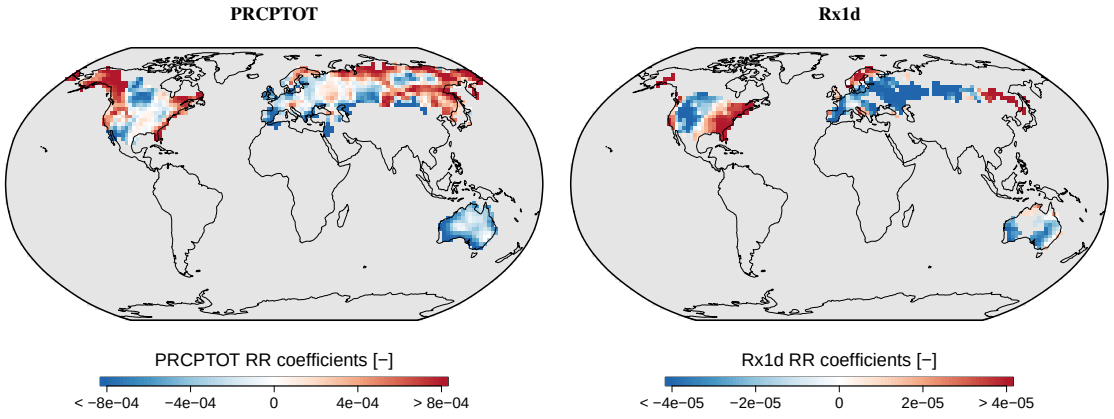
(a) PRCPTOT standard deviation of residuals

(b) Rx1d standard deviation of residuals

SI Figure S7: Slightly smoothed distribution of standard deviations of the residuals of the linear trend fits to the forced response estimates over the full period 1951-2014 for forced response estimates determined from piControl and forced model simulations on all observational masks (shaded density plots, corresponding forced response estimates are those shown in main figure 2). Standard deviation of the residuals of the linear fit to observed forced response estimates are shown by coloured vertical lines. PRCPTOT (a) and Rx1d (b).

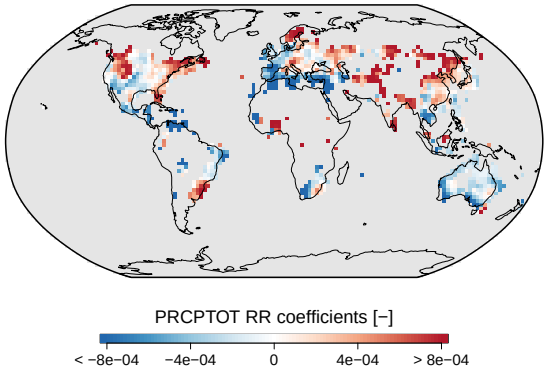
Detrended case

Also for the mean removed case, figure S8, the similarities show. For GHCNDEX Rx1d, however, it can also be seen that the missing coverage in South-East Asia, South America and South Africa, as compared to HadEX3, is detrimental for the RR model's ability to estimate the forced response (compare to main figure 4d). The residuals for the detrended case are consistent across models and observations too, as figure S9 shows, apart from GHCNDEX PRCPTOT. The fact that GHCNDEX PRCPTOT forced response estimates show considerably higher variance than model PRCPTOT forced response estimates, implies that the very high forced response estimate trend seen in this case is unreliable.



(a) GHCNDEX-mask PRCPTOT annual RR fingerprint

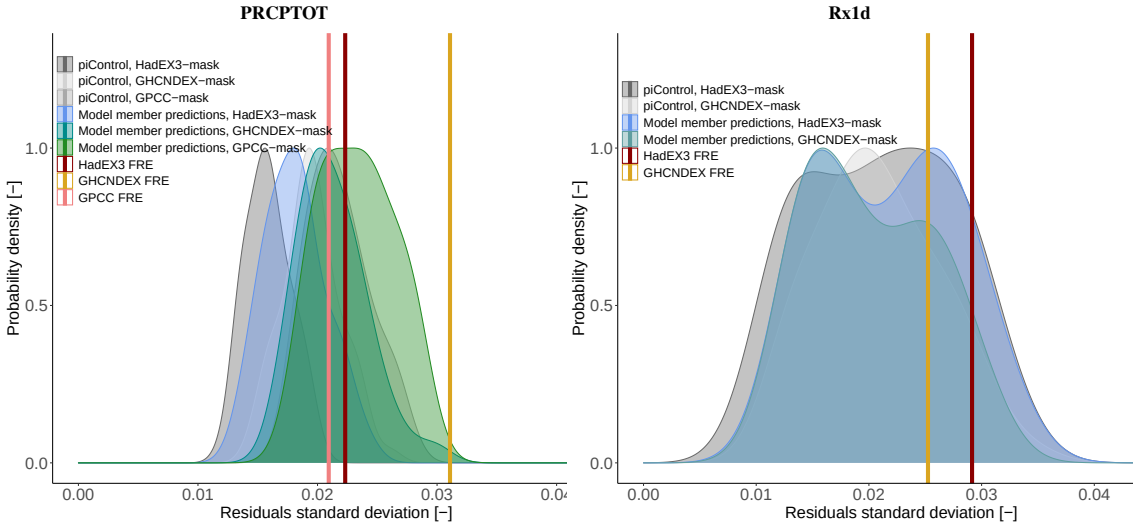
(b) GHCNDEX-mask Rx1d annual RR fingerprint



(c) GPCC-mask PRCPTOT annual RR fingerprint

(d)

SI Figure S8: RR fingerprints for PRCPTOT (a, c), and Rx1d (b) as in main figure 3, for resolution and coverage masks of GHCNDEX and GPCC, trained on model data from which the global mean was subtracted (detrended).



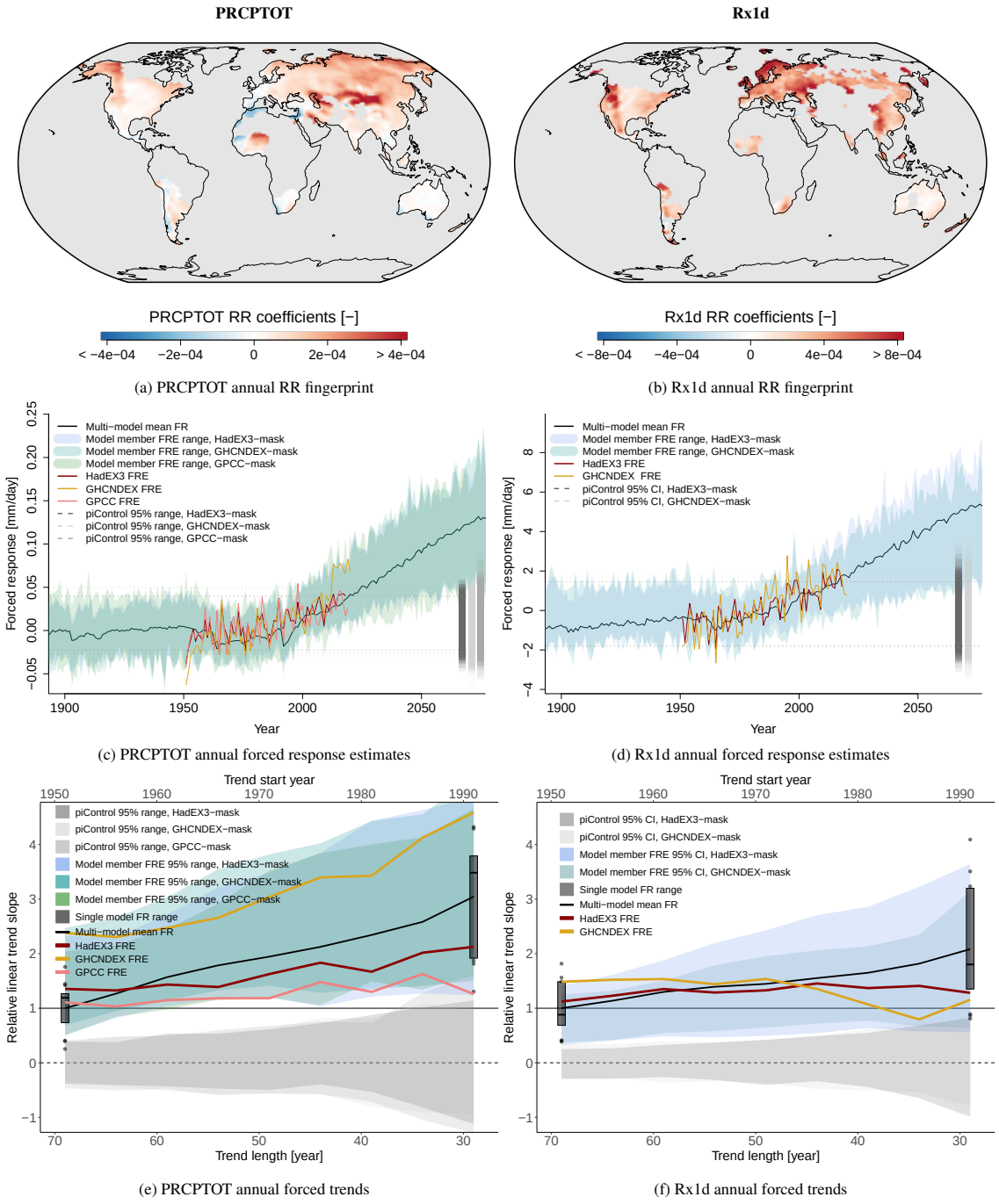
(a) PRCPTOT standard deviation of residuals, detrended case

(b) Rx1d standard deviation of residuals, detrended case

SI Figure S9: As figure S7 but for RR models trained on data from which the global mean is subtracted (detrended). Corresponding forced response estimates are those shown in main figure 4.

S2.2 Alternative forced response target: global mean

In figure S10, the results of the RR procedure with area-weighted annual global means (model ensemble means) as targets is shown. Comparing figure S10 to its counterpart, main figure 2, shows that the choice of target metric only has negligible impact on the results of this study. In the figure below, the target metric (black lines) are smoother than in the default case, especially for PRCPTOT, however, the fingerprints and trends are virtually identical. This also leads to nearly identical times of emergence (not shown). A reason to choose the EOF-based target rather than the global mean based one shown here, is the effect of non-GHG forcings. In the global mean, these effects of large volcanic eruptions or aerosols might have a larger and long-term effect on the trend of the forced response best estimate than in an PC dominated by GHG forcing. A more direct way to isolate the GHG-forced response would be by using single-forcing ensembles.



(a) PRCPTOT annual RR fingerprint

(b) Rx1d annual RR fingerprint

(c) PRCPTOT annual forced response estimates

(d) Rx1d annual forced response estimates

(e) PRCPTOT annual forced trends

(f) Rx1d annual forced trends

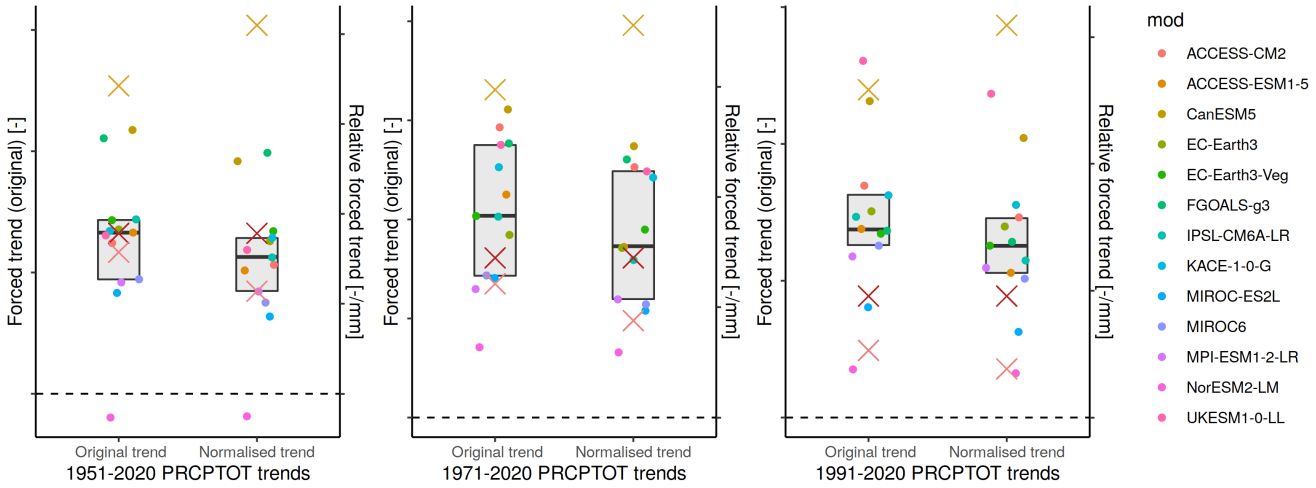
SI Figure S10: As main figure 2 but for RR models trained with area-weighted global mean PRCPTOT and Rx1d as forced response target

S2.3 Absolute versus normalised precipitation metrics

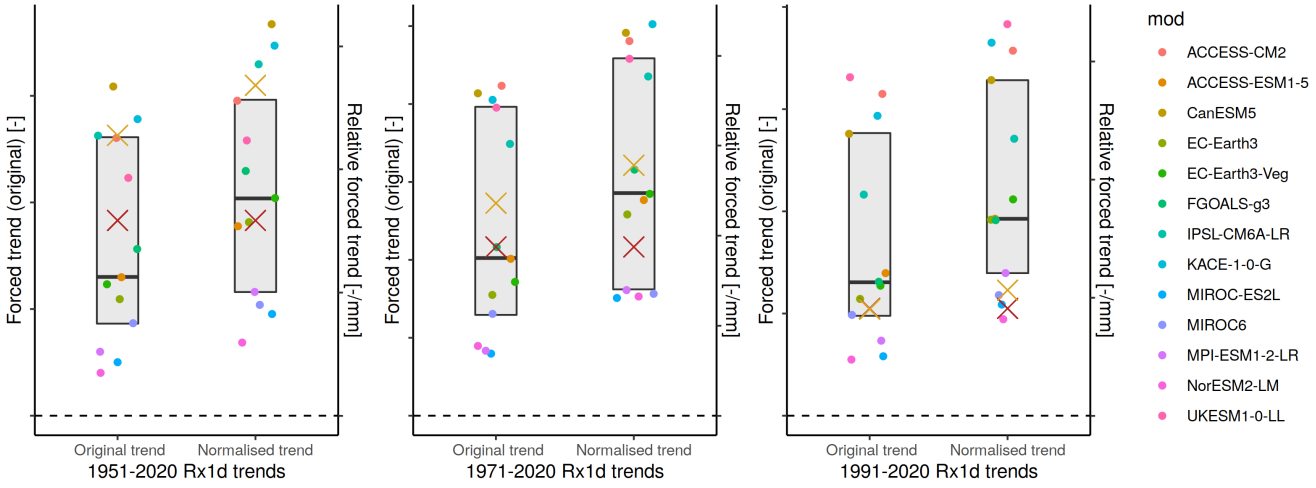
In the main text, we focus on absolute units of precipitation (although not in mm, since the EOF-based targets do not support this). Since the picture may change if other commonly used precipitation metrics are used, we here show a sensitivity study verifying that our conclusions hold regardless of the precipitation metric assessed.

Firstly, we normalise PRPTOT and Rx1d forced response estimates by their respective (masked) global mean values, which gives relative change and removes the sensitivity of trends to differing climatological precipitation levels across models and observations. In a second normalisation step, we normalise the relative change over a certain period by the temperature change over the corresponding period. This removes the dependence of results on differing climate sensitivities across models and observations. In the default case, where forced response estimates are no measure of global mean PRCPTOT or Rx1d (but of forced pattern strength), the normalised quantity is not directly comparable to Clausius-Clapeyron scaling, yet, for illustration purposes, we also apply the normalisation to the forced response estimates resulting from the procedure with PRCPTOT and Rx1d global means as a target. The latter gives a change in %/K.

Note we do not assess local relative changes (per-gridpoint normalisation w.r.t. climatology), since this would lead to inflation of very small positive changes in arid regions with near-zero climatological precipitation, which would then disproportionally affect the result.



(a) PRCPTOT



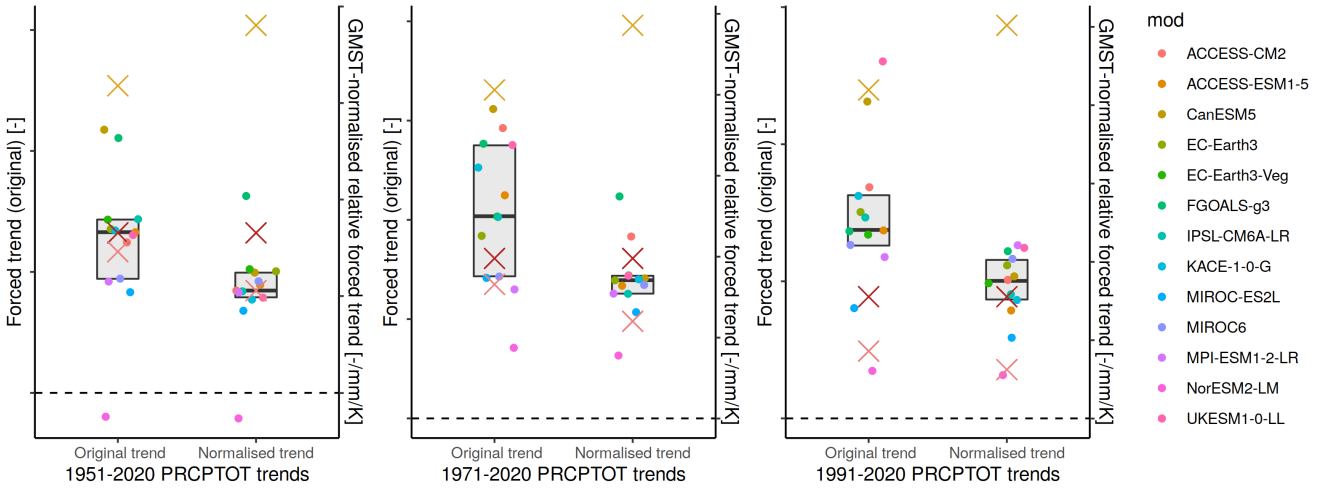
(b) Rx1d

SI Figure S11: Comparison of original PRCPTOT (a) and Rx1d (b) trends (as in manuscript) and trends normalised by the model's/observation's corresponding climatological PRCPTOT/Rx1d levels being the 1951-2014 mean, averaged over the observational masks. Trends of single-model targets (points and corresponding boxplot indicating the interquartile range), and observed forced response estimates (X-marks). Non-physical units, black dashed line indicates 0.

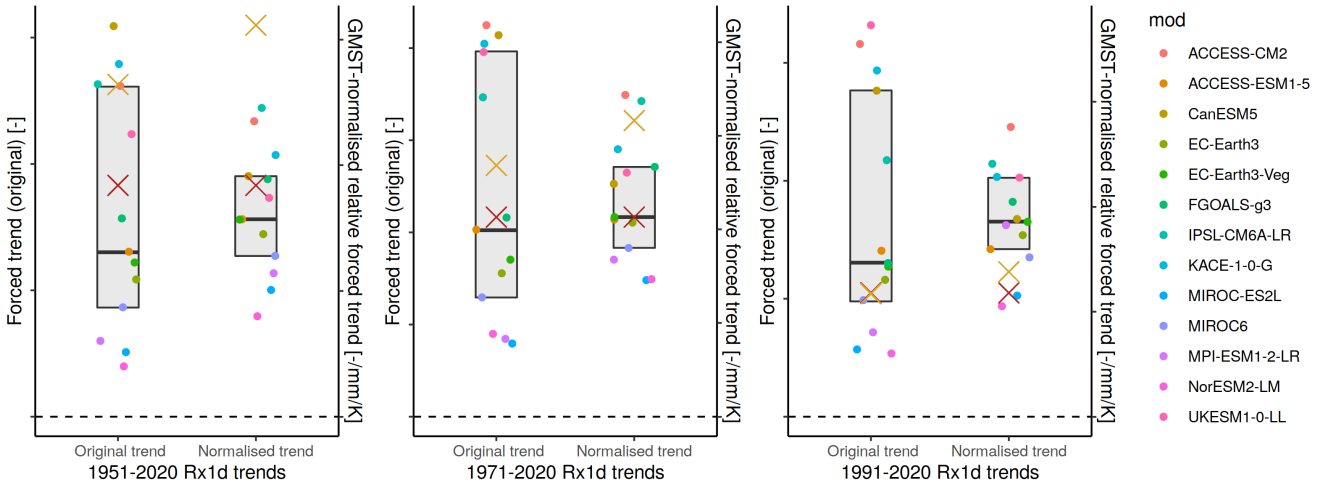
Figure S11 shows how normalising by precipitation climatology changes the results. Note that these plots represent three points (start years 1951, 1971, and 1991, from left to right) in figure 3 in the manuscript. The different start years, as in the

manuscript, allow for assessment of changing relative trends depending on trend period. Comparing the left and right half of each plot reveals the difference between the original trends as in the manuscript (left) and the normalised ones (right). For PRCPTOT (figure S11a), we see that normalising trends w.r.t. climatological mean precipitation shifts the modelled forced trends down relative to observations, consistent with the models exhibiting slightly higher climatological PRCPTOT levels - a known persistent systematic bias (Stephens et al., 2010). Despite slight decreases in model forced trends, it remains the case that the relative magnitude of model forced trends and observed forced trend estimates depends on the period and observational dataset.

For Rx1d (figure S11b), on the contrary, normalising trends w.r.t. climatological mean Rx1d increases forced model trends relative to observed forced trend estimates, suggesting climatological mean levels of Rx1d are lower in models than in observations, which is also a known model bias (Sillmann et al., 2013; Bador et al., 2020). Nonetheless, again, the main conclusions on the relative model vs. observational trend magnitudes do not change. These opposing findings regarding PRCPTOT and Rx1d, align well with the findings of Fischer and Knutti (2016), who suggest PRCPTOT changes are overestimated by models, whereas Rx1d changes are underestimated.



(a) PRCPTOT

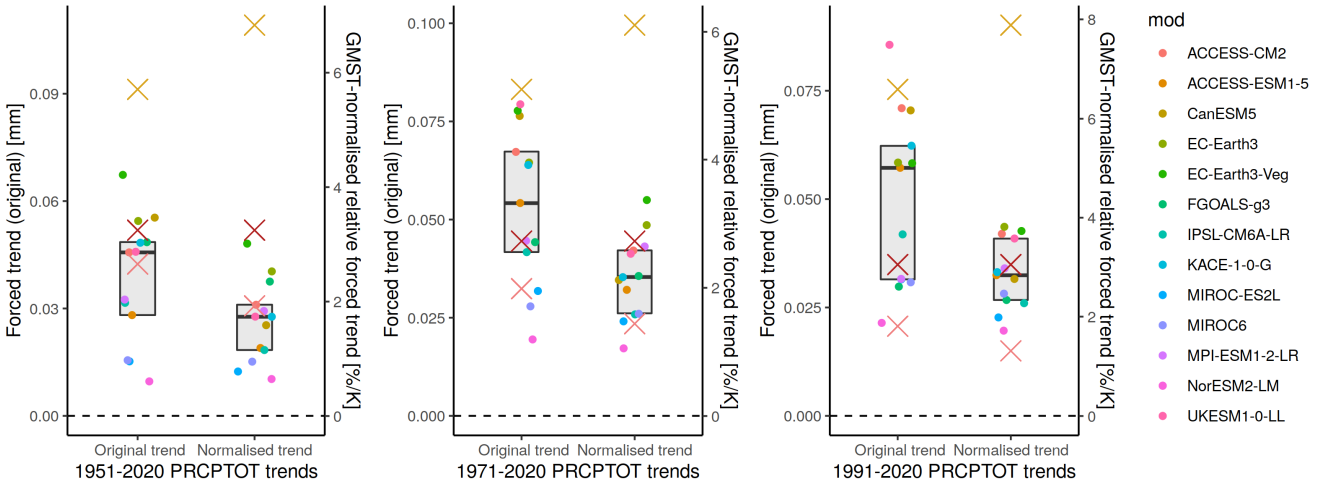


(b) Rx1d

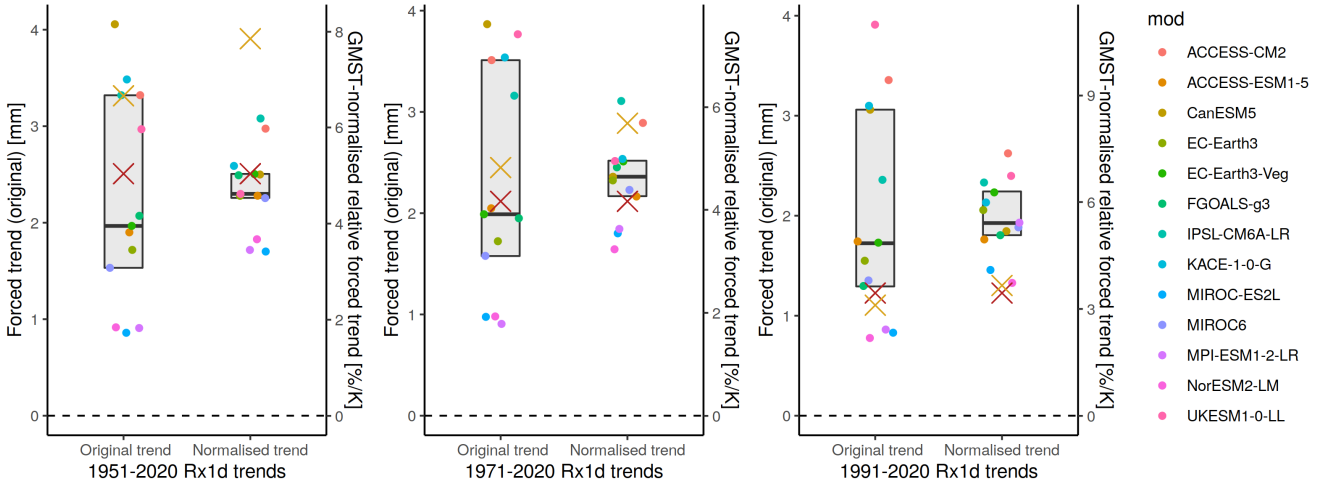
SI Figure S12: Comparison of original PRCPTOT (a) and Rx1d (b) trends (as in manuscript) and trends normalised by the model's/observation's corresponding climatological PRCPTOT/Rx1d levels and temperature change (difference between the 2020 value and the values in 1951, 1971, and 1991 of the 21-year LOWESS-smoothed global mean surface (air) temperature (from Cowtan and Way (2014) for observations).

The comparison between original trends, as in the manuscript, and relative GMST-normalised trends is shown for PRCPTOT in figure S12a and Rx1d in figure S12b. Comparing the left and right column in each panel shows that normalising the forced relative trends from figure S11 w.r.t. their corresponding temperature change reduces model spread, which is to be expected. For PRCPTOT, GMST-normalisation further reduces model trend magnitude relative to observed forced trend estimates, since model warming rate in CMIP6 is higher than in observations. Therefore, for Rx1d, GMST-normalisation

reduces model trends as well, and offsets some of the effect of normalising w.r.t climatological Rx1d levels seen in figure S11b. However, more importantly, figure S11 and S12 show that, compared to the original trends, the relative magnitude of model and observational trends changes somewhat in response to normalising w.r.t climatology and warming rate, but the main picture does not change - relative trend magnitudes still differ between periods and observational datasets. The main conclusion of our study – forced trends are detected, but observations lie on different ends of the model-projected spectrum – holds also for normalised trends.



(a) PRCPTOT



(b) Rx1d

SI Figure S13: As Fig. S12 but target trends (points/boxplot) and forced response estimates are based on the procedure using global mean PRCPTOT and Rx1d as forced response metric (as opposed to the EOF-based metric), leading to physical units.

For trends in %/K (i.e. physical units), obtained by using the forced response estimates based on the ridge model with the global mean target (see Sect S2.2), the main conclusion also does not change qualitatively, as shown in figure S13. Relative model and observational trends remain dependent on observational dataset and trend period. Normalising even suggests larger differences across different observational datasets. This check also shows that global mean based ridge regression also reproduces numbers in the range of the well-known 2-3%/K change in global mean PRCPTOT. For Rx1d, the 5%/K change we find is lower than the 7%/K change prescribed by Clausius-Clapeyron, which has been found for CMIP models of different generations before (Allan and Soden, 2008; Kotz et al., 2022). Note that we are restricted to normalising with respect to a climatological precipitation value that is based on the mean over the grid cells with observational coverage, in order to "treat" model and observational data the same. Therefore, the percentages may be off, since the global mean differs from the mean we use.

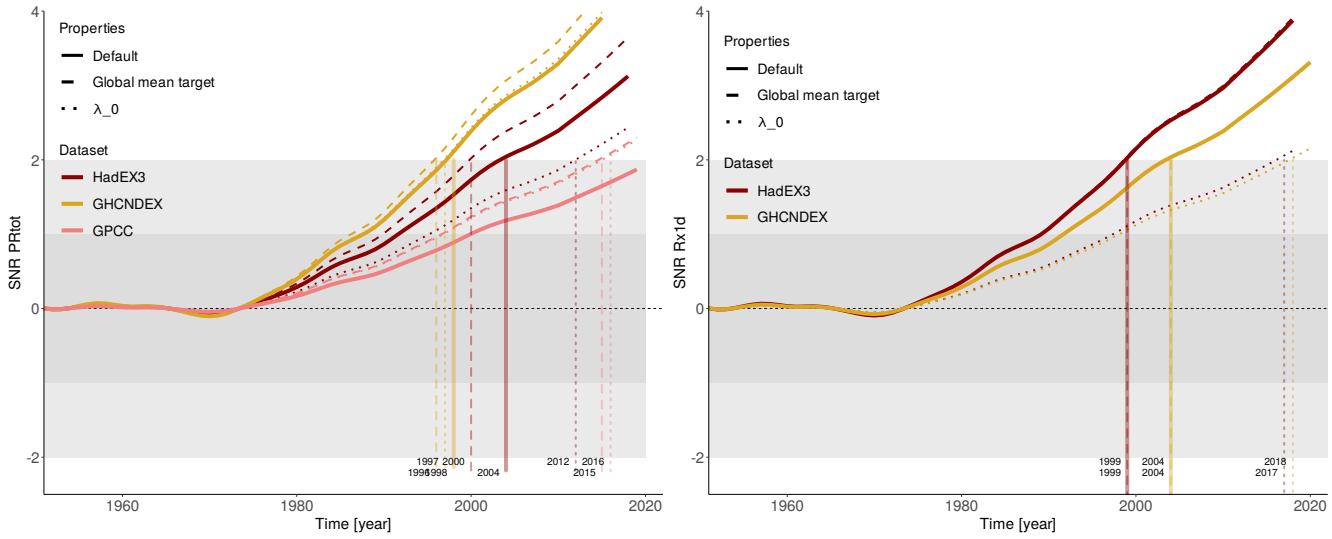
Finally, table S14 shows an overview of D&A studies assessing PRCPTOT and Rx1d, including their judgment on whether models over- or underestimate changes with respect to observations.

SI Figure S14: Previous D&A studies on PRCPTOT and Rx1d, including their main findings on whether modelled forced changes are smaller (-), similar (0) or larger (+) than observed forced changes. (Noake et al., 2012; Wu et al., 2013; Polson et al., 2013; Fischer and Knutti, 2014; Knutson and Zeng, 2018; Min et al., 2011; Zhang et al., 2013; Fischer and Knutti, 2016; Borodina et al., 2017; Paik et al., 2020; Sun et al., 2022)

Paper	Model archive	Obs dataset	Spatial region	Variable	Method	Trend periods	Models vs. observations	Remarks
PRCPTOT								
Noake et al. (2012)	CMIP3	GHCN, CRU, VASClimO	Global land, separated into 5deg latitude bands. Scaling factors determined for spatiotemporal aggregate, not per latitude band.	Seasonal PRCPTOT percentage change per latitude band	Optimal fingerprinting	1952-1990, 1960–1999, 1951–1990, 1975–1999	-	"-" applies to scaling factor best estimate for seasons and observational datasets in which significant change is detected (confidence interval does not include 0), and holds for all trend periods.
Wu et al. (2013)	CMIP5	GHCN	Northern hemisphere land	PRCPTOT percentage change	Optimal fingerprinting	1952-2011	-	
Polson et al. (2013)	CMIP5	GHCN, CRU, VASClimO, GPCC	Global land, separated into 5deg latitude bands. Scaling factors determined for spatiotemporal aggregate, not per latitude band.	Seasonal PRCPTOT percentage change per latitude band	Optimal fingerprinting	1951–2005 (2000 for VASClimO)	-	Applied Noake's method to CMIP5. "-" applies to scaling factor best estimate for seasons and observational datasets in which significant change is detected (confidence interval does not include 0). GPCC never shows a detectable climate signal.
Fischer & Knutti (2014)	CMIP5 + CESM initial condition ensemble	HadEX2, GHCNDEX	Global	Spatial distibution of gridpoint trends in PRCPTOT, expressed in terms of local sigma (based on 1986-2005 interannual variability)	Spatial probability distribution comparison	1960-2010	+	Models estimate more regions with positive trends in PRCPTOT, but not enough negative trends --> too much wetting.
Knutson & Zeng (2018)	CMIP5	GPCC	Global, per gridpoint	Linear trend in PRCPTOT	Linear trend fitting to grid point timeseries	1901-2010, 1951-2010, 1981-2010	-	Models cannot produce the magnitude of positive nor negative trends in obs. Discrepancy gets stronger in later trend periods.
Rx1d								
Min et al. (2011)	CMIP3	HadEX	NH land, separated into (overlapping) regions: mid-latitudes, tropics	Rx1d and Rx5d Probability index: 0-1 quantile per value, based on fit GEV per gridpoint	Optimal fingerprinting	1951-1999	-	"-" applies to scaling factor best estimate for regions where there is detection.
Zhang et al. (2013)	CMIP5	HadEX2 + Russian station data	NH land, separated into (overlapping) regions: Western Eurasia, Eastern Eurasia, North America, mid-latitudes, tropics	Rx1d and Rx5d Probability index: 0-1 quantile per value, based on fit GEV per gridpoint/station and then interpolated	Optimal fingerprinting	1951-2005	0/+	Scaling factor estimates include 1, but best estimates are still below 1.
Fischer & Knutti (2014)	CMIP5 + CESM initial condition ensemble	HadEX2, GHCNDEX	Global	Spatial distibution of gridpoint trends in Rx5d, expressed in terms of local sigma (based on 1986-2005 interannual variability)	Spatial probability distribution comparison	1960-2010	-	Models don't show a large enough land fraction exhibition positive trends, and do not reproduce the magnitude of the largest trends seen in observations.
Fischer & Knutti (2016)	CMIP5 and EURO-CORDEX	E-OBS/Ensembles	Europe	Changing occurrence of historical >90 percentile daily precipitation	Probability distribution comparison	1951-1980 and 1981-2013	-	Models show smaller increase in intensity of >90th percentile daily precipitation amounts.
Borodina et al. (2017)	CMIP5 + CESM initial condition ensemble	GHCNDEX, HadEX2	Global land, selected wet regions only (wettest 40%, agreed across models)	Rx1d percentage change per gridpoint as a function of GMST [%/K], averages over wet regions, as well as land area fraction experiencing positive Rx1d trends	Trend comparison	1951-2005	-	Models show smaller trends than both observational datasets, but HadEX2 shows smaller trends than GHCNDEX.
Paik et al. (2020)	CMIP5	HadEX2	Global land, separated into (overlapping) regions: Western Eurasia, Eastern Eurasia, North America, mid-latitudes, tropics, wet and dry regions.	Rx1d and Rx5d Probability index: 0-1 quantile per value, based on fit GEV per gridcell/station and then interpolated	Optimal fingerprinting	1950-2020	0/+	"0" applies to EU and dry regions, where models and observations agree. For all other regions with detection, models overestimate the change ("+").
Paik et al. (2020)	CMIP5	HadEX2	Global land, separated into (overlapping) regions: Western Eurasia, Eastern Eurasia, North America, mid-latitudes, tropics, wet and dry regions.	Rx1d and Rx5d Probability index: 0-1 quantile per value, based on fit GEV per gridcell/station and then interpolated. Spatially averaged trends, normalised by GMST	Trends in %/K	1950-2020	-	Note: same study as above. In all regions where forced change is detected, models underestimate observations when trends in %/K are assessed. In these same regions, scaling factors suggest that models overestimate change.
Sun et al. (2022)	CMIP6 and CanESM2 LE	HadEX2 stations + Russian and Chinese station data	Global, continental, regional	Rx1d and Rx5d, Non stationary spatiotemporal varying GEV-based optimal fingerprinting, no normalisation: absolute units of precipitation (log)	Non-optimal variant of optimal fingerprinting; scaling factor determination but no internal variability covariance corrections	1950-2014	+	"+" applies to all continents/regions, and also global level, but Northwestern Europe (Scandinavia/UK) where scaling factors are around 1 ("0").

S2.4 Effect of design choices on time of emergence

The contribution of target choice and λ choice to signal-to-noise ratio (SNR) is shown in figure S15, which closely relates to main text figure 5.



(a) PRCPTOT

(b) Rx1d

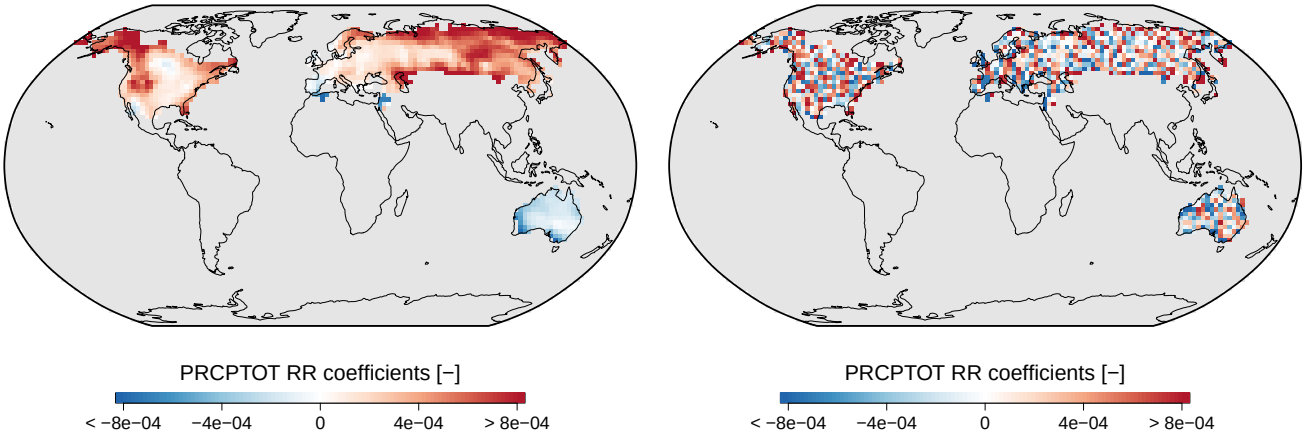
SI Figure S15: SNRs of mean total precipitation (PRCPTOT) (a) and extreme precipitation (Rx1d) (b) forced response estimates in GHCNDEX, HadEX3 and GPCC, including senstivities to target metric and regularisation. Exceedance of an SNR of 2 implies emergence. Signal is defined as forced response estimate regressed onto 21-year LOWESS filtered global mean surface temperature, noise as residuals of this fit.

In these plots, "default" refers to the results of the procedure using EOF-based regression targets and our λ_{sel}, chosen as outlined in SI Sect. S1.3, "Global mean target" shows the results for model ensemble global mean PRCPTOT and Rx1d targets, and λ_{sel}, "λ_0" shows the results for EOF-based targets, but very small λ, equivalent to almost no regularisation.

The SNR does not necessarily increase by using the EOF-based target instead of the global mean target; for PRCPTOT (Fig. S15a) the EOF-based target exhibits lower SNR, whereas for Rx1d it does not make any difference whether we use the global mean based target or the EOF-based target. The choice of using the EOF-based metric for PRCPTOT thus requires some explanation. The global mean based target leads to higher SNR because the trend in global mean precipitation is stronger than the trend in the first principal component of mean precipitation, and models are more in agreement on global mean precipitation change than on the first EOF-pattern. However, since forced changes in mean precipitation behave according to a pattern of wetting and drying regions (Held and Soden, 2006), the global mean trend in precipitation is not a very refined measure of forced precipitation changes. The first EOF captures the forced pattern of change, and its corresponding principal component time series captures the strength of that pattern. The advantage of using the EOF pattern is that the forced response in all regions is somewhat reflected in the corresponding principal component, and not averaged out as in the global mean. In addition, individual models' deviations from the multi-model pattern due to uncertainties in e.g. the forced response in circulation, are reflected in the projections of the EOF on the model ensemble means which serve as our model-specific forced response targets. Since the EOF-based target metric has a weaker trend and more variability for PRCPTOT, the ridge model and the forced response estimates are "pushed" in a more conservative direction. We argue that this is the better approach, given that the goal is not to construct a ridge model that generates the strongest forced response estimate, but one that is most likely to predict the true forced response given the observations that are available. Our default, therefore, is to use the more conservative targets, which implicitly include pattern information and uncertainties. We point out, however, that the main conclusions, which are detection of a forced response but disagreement among observational datasets on the strength of the observed forced response relative to the simulated forced response, are insensitive to the choice of target metric.

Comparing times of emergence of the default case and λ_0 indicates the benefit of using regularised regression. λ_0 is not equivalent to ordinary least squares, in that λ is not set to 0, but it is the smallest λ_0 used in the training procedure, and in all cases at least two orders of magnitude smaller than$\lambda_s el$. A smaller λ increases the variability in the forced response estimate, but, likely, also the trend. Therefore, when it comes to SNR, the effect of λ choice is a trade-off between the increased variability and the increased trend. For Rx1d, we see that λ_0 deteriorates the detectability; overfitting leads to large variability increase without reducing a low trend bias. In PRCPTOT, the effect is messier. For HadEX3, the SNR clearly decreases for λ_0 but for GHCNDEX and GPCC this is not the case. Analysis shows that the strong uptick at the end of the GHCNDEX record (referred to in the manuscript) is somewhat dampened by larger λs. When λ_0 is close to zero, the GHCNDEX forced response

estimate shows this strong increase in the last few years of the record, which amplifies the overall trend, and therefore high SNRs are seen. For λ_0, however, physical consistency of the fingerprints is strongly impaired, exemplified by Fig. S16



(a) Regularisation parameter: λ_{sel}

(b) Regularisation parameter: λ_0

SI Figure S16: Fingerprints for GHCNDEX Annual PRCPTOT for λ_{sel} (a) $lambda_0$ (b)

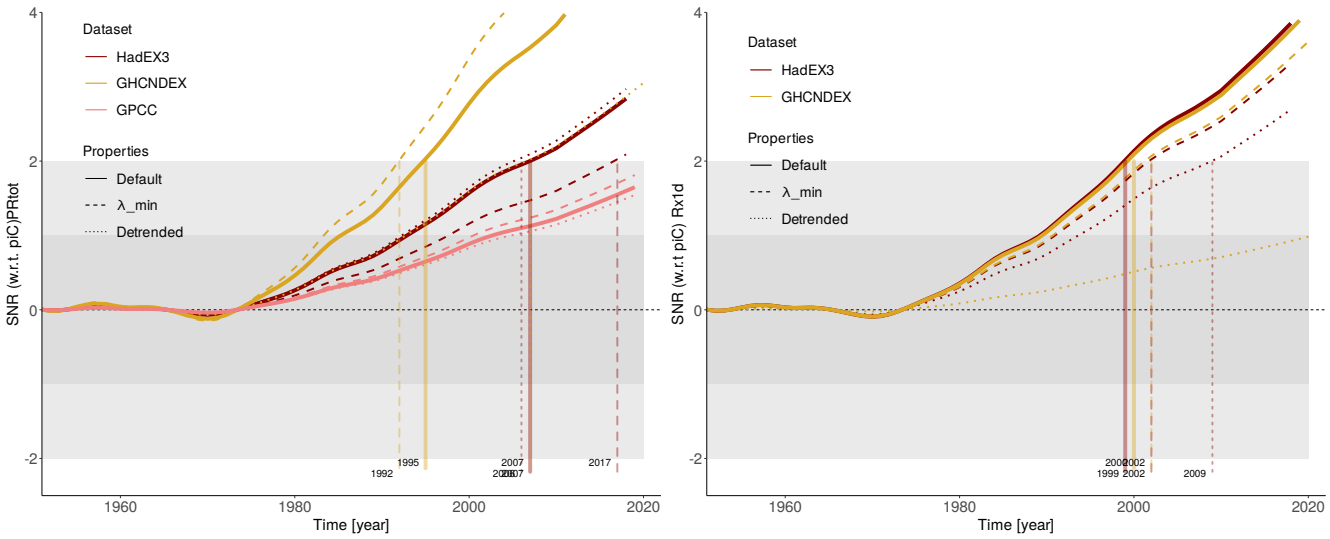
For GPCC, the low coverage leads to generally very high variability in the forced response estimate, as also witnessed by the low SNRs. λ_0 leads to a slightly larger increase in trend relative to the increase in variability, however, the fingerprints no longer reflect any physical consistency. Polson et al. (2013) also found it is difficult to detect forced responses in GPCC.

The above shows that it is important to assess the complete result of fingerprints, forced response estimates, and SNRs to judge the quality of the detection model and the detected response. PRCPTOT is generally a more difficult variable to detect forced trends in, due to the spatial pattern of change and high internal and model variability in the representation of this pattern. This was also found by e.g. Fischer and Knutti (2014). For the most recent, higher-resolution and higher-coverage HadEX3 dataset, however, ridge regression also has clear benefits for the detection of forced trends in PRCPTOT, besides the fingerprint interpretability advantages which we see in all three observational datasets.

In Fig. S17 time of emergence for the same models as in the main text is shown, however, the noise N in this calculation of SNR is now the standard deviation of the forced response estimates obtained when applying the detection fingerprints to piControl model output. I.e., instead of assessing when the observed PRCPTOT and Rx1d emerge from "their own" spread around the mean, we assess when they emerge from the simulated unforced spread. This is, evidently, a less consistent metric since the simulated spread in piControl forced response estimates is not directly comparable to the spread in observed forced response estimates. Furthermore, piControl forced response estimates for different observational datasets only differ in their coverage (observation-based mask), but not in the underlying data. Between the observational datasets, both coverage as well as actual data values differ.

Minor changes are seen, e.g. the earlier emergence of PRCPTOT and the higher SNR of λ_{min} relative to the default shows that piControl spread in PRCPTOT is smaller than observed spread. The increased agreement between HadEX3 and GHCNDEX in Rx1d emergence is due to the fact that the higher variability in GHCNDEX relative to HadEX3 is no longer reflected in the SNR because N is now based on the same model runs for both datasets. This shows again that the coverage difference only explains part of GHCNDEX's lower detectability – other disagreements are due to observational differences between the different datasets.

Note that one can determine SNR in many different ways, including e.g. the ratio of observed trends to (standard deviation of) piControl trends. This indicates when the trend statistically emerges. The measure we use in the main text, on the contrary, indicates when the accumulated change in PRCPTOT or Rx1d is so large that the mean level no longer lies within 2 standard deviations from the starting, unforced climatalogy.



(a) PRCPTOT

(b) Rx1d

SI Figure S17: SNRs of mean total precipitation (PRCPTOT) (a) and extreme precipitation (Rx1d) (b) forced response estimates in GHCNDEX, HadEX3 and GPCC, where noise is spread in forced response estimates from piControl simulations. Signal is defined as forced response estimate regressed onto 21-year LOWESS filtered global mean surface temperature, noise as residuals of this fit.

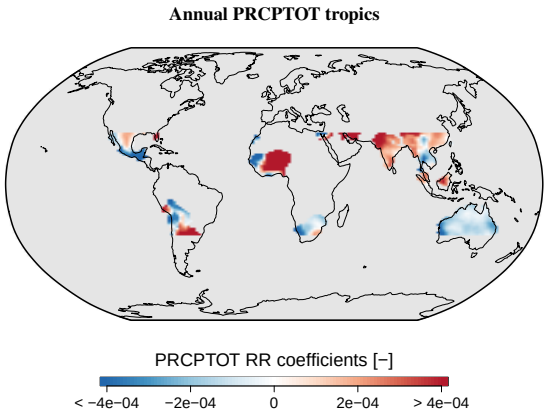
S3 Regional and seasonal analysis

As mentioned in the main text, the Northern Hemisphere (NH) signals make up the largest contribution to primarily to the total forced response estimate. To exemplify this we show fingerprints and forced response estimates for three separate regions, namely the extratopical NH (30N-90N), extratropical Southern Hemisphere (SH) (30S-90S) and the tropics (30S-30N). In the season-free tropical region, we continue to use the annual timescale only. The two extratropical regions, however, have distinct seasons with season-specific climatological patterns, meaning that seasonal timescales provide more specific information than annual timescales. We therefore also assess December-January-February (DJF) and June-July-August (JJA) in the extratropical regions.

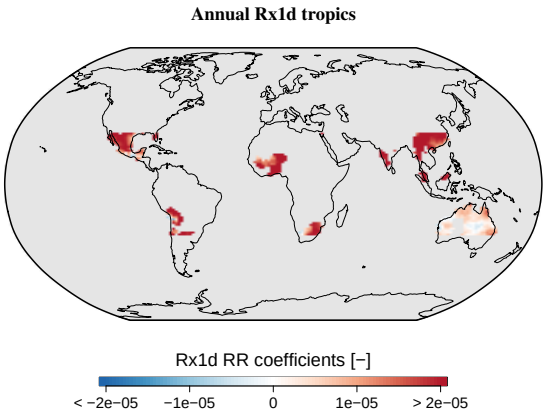
For the figures shown, the forced response targets used for RR model training are once again the projections of the multi-model mean first EOF onto ensemble means, following the procedure described in section 2 in the main text. Separate EOFs and corresponding forced response targets were determined for each region, to capture the region-specific forced response in the target.

Tropics, annual

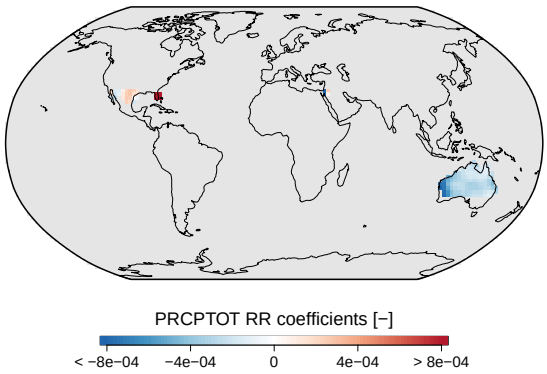
Figure S18 shows the fingerprints for the annual tropics case, all observational coverage masks are shown for comparison, as well as the corresponding forced response estimates. For PRCPTOT, the topical signal is clearly very noisy, and despite the generous definition of the tropics, including almost all of Australia and South Africa, this region alone does not contain robust enough signals to construct an RR model that can extract the forced tropical PRCPTOT signal from observations. This reflects the high internal variability in the tropics, but also the high degree of model disagreement on the pattern of forced change to total precipitation. For Rx1d, the more uniform increase in the tropics does enable signal isolation from observations that is consistent with models for HadEX3 (S18g). The RR model trained on the very limited GHCNDEX data, however, cannot do better than predicting the time average.



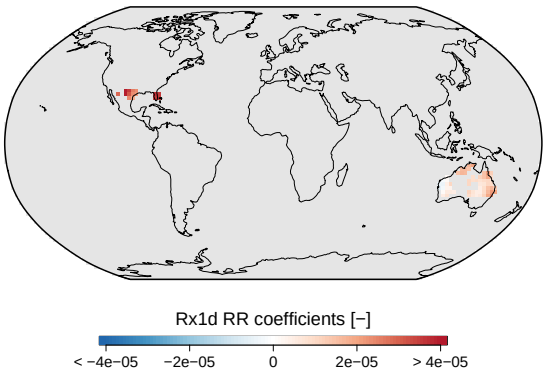
(a) HadEX3-mask PRCPTOT annual RR fingerprint



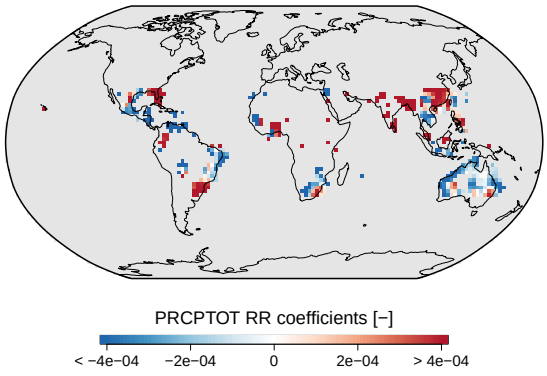
(b) HadEX3-mask Rx1d annual RR fingerprint



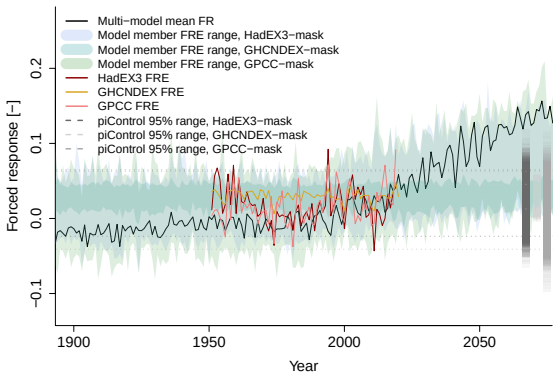
(c) GHCNDEX-mask PRCPTOT annual RR fingerprint



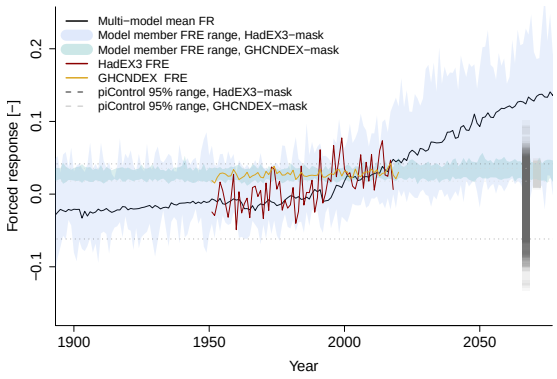
(d) GHCNDEX-mask Rx1d annual RR fingerprint



(e) GPCC-mask PRCPTOT annual RR fingerprint



(f) PRCPTOT annual forced response estimates, tropics



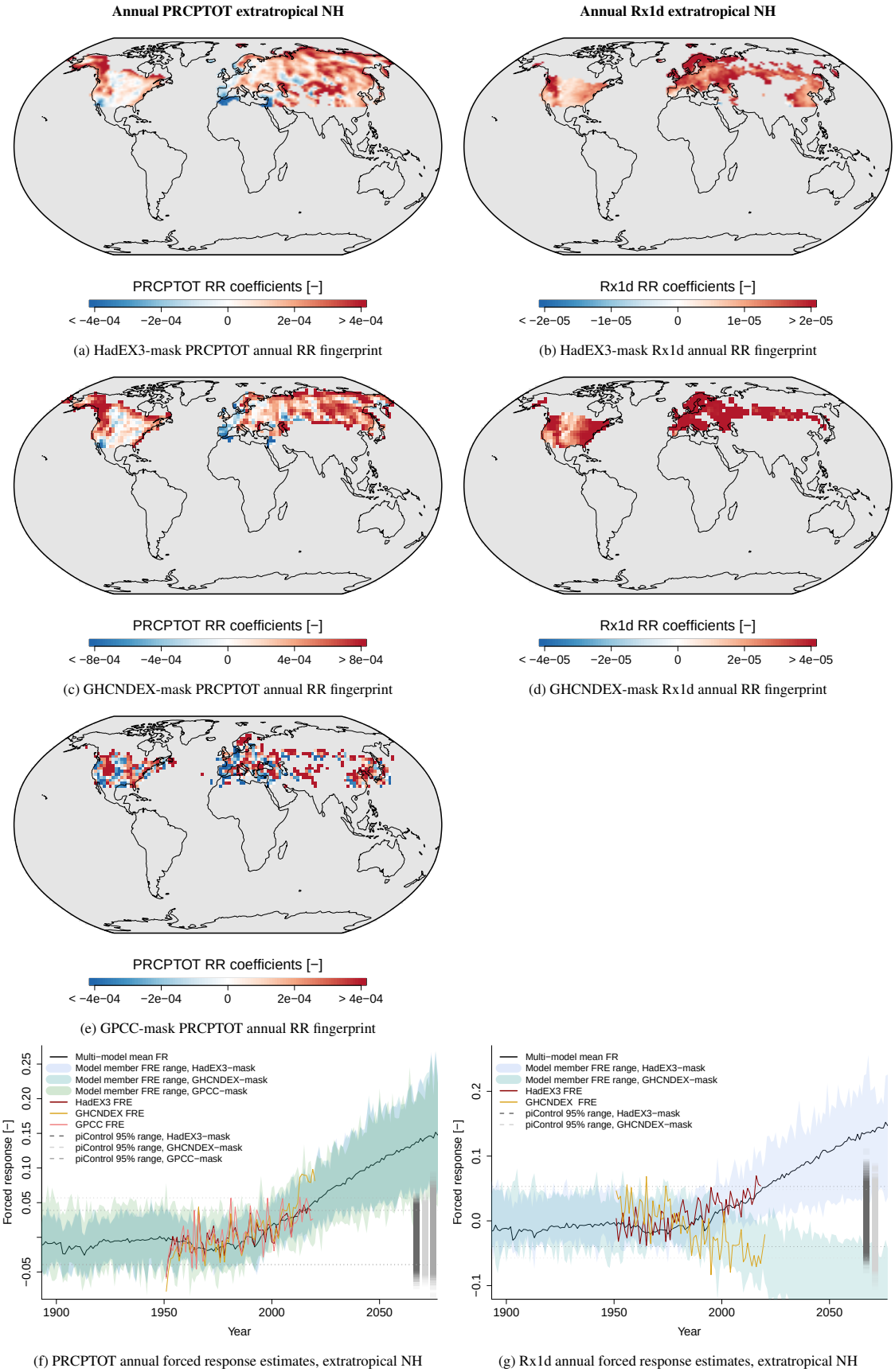
(g) Rx1d annual forced response estimates, tropics

SI Figure S18: RR fingerprints on all observational coverage masks (a, b, c, d, e) and forced response estimates (f, g) as in main figure 2 but for tropical annual PRCPTOT and Rx1d.

Extratropical Northern Hemisphere, annual and seasonal

Figure S19 shows the annual fingerprints and forced response estimates for the extratropical NH. Unsurprisingly, these results are very similar to the global annual detection results in the main text, due to the fact that most of the coverage is in the NH. Figure S20 shows the fingerprints and forced response estimates for NH winter (DJF). As mentioned above, the extratropical forced response target (black line) is much less noisy than the tropical one, reflecting lower internal variability and higher model agreement. The higher granularity of the fingerprints, especially for PRCPTOT, might be a consequence of this smoother target; the smoothness results in relatively smaller forced response estimate errors and less of an error increase when variance of the forced response estimate increases, leading to lower regularisation parameters. Nonetheless, the general large scale patterns can still be distinguished in the form of mostly positive weights in mid to high latitudes, and negative weights in regions with lower projected changes or drying. For Rx1d, the primarily positive response is clearly represented in the fingerprint, as well as small regions of lower extreme precipitation, such as the Mediterranean. From the forced response estimates (lowermost panel) it is evident that the forced signal in both PRCPTOT and Rx1d can be extracted from the observational datasets when only NH winter is addressed.

The fingerprints for extratropical NH summer (JJA), figure S21 are physically interpretable, picking up the Mediterranean drying and Northern European wettening signal (especially for Rx1d), associated with northward stormtrack displacement. PRCPTOT GPCC looks highly overfit, however, due to the low and spatially discontinuous coverage. Despite the interpretability, the forced response estimates from observations (S21e and S21f) do not show strong consistent trends that can be distinguished from the noise. This is found consistently in other studies as well. A potential explanation for this is the nature of summer precipitation being mostly convective: the models used are not convection-permitting, and the spatial RR fingerprints thus also do not represent the regions where changes in convective precipitation are strong. It would be instructive to find out if convection permitting simulations can be used in combination with RR to detect forced changes in summer precipitation in the NH. In addition, the GHG-signal in NH summer precipitation is likely to be obscured by changing precipitation-inhibiting aerosol effects. Particularly summer convective precipitation is negatively affected by aerosols due to their decreasing effect on surface temperature and increasing effect on droplet number concentrations (Undorf et al., 2018; Stjern and Kristjánsson, 2015). Between roughly 1951 and 1975 industrial aerosol emissions in Europe and the US reached their peak and inhibited convective precipitation increases. From 1975 onwards, aerosol concentrations over Europe and the US decreased, in concert with increases in convective precipitation, while they continued to rise in (South-)East Asia leading to more convective precipitation (Stjern and Kristjánsson, 2015). The spatial and temporal changes in aerosol forcing compromise the appropriateness of one fingerprint to detect these forced changes, and call for an approach that separates individual forcings (and regions).



(a) HadEX3-mask PRCPTOT annual RR fingerprint

(b) HadEX3-mask Rx1d annual RR fingerprint

(c) GHCNDEX-mask PRCPTOT annual RR fingerprint

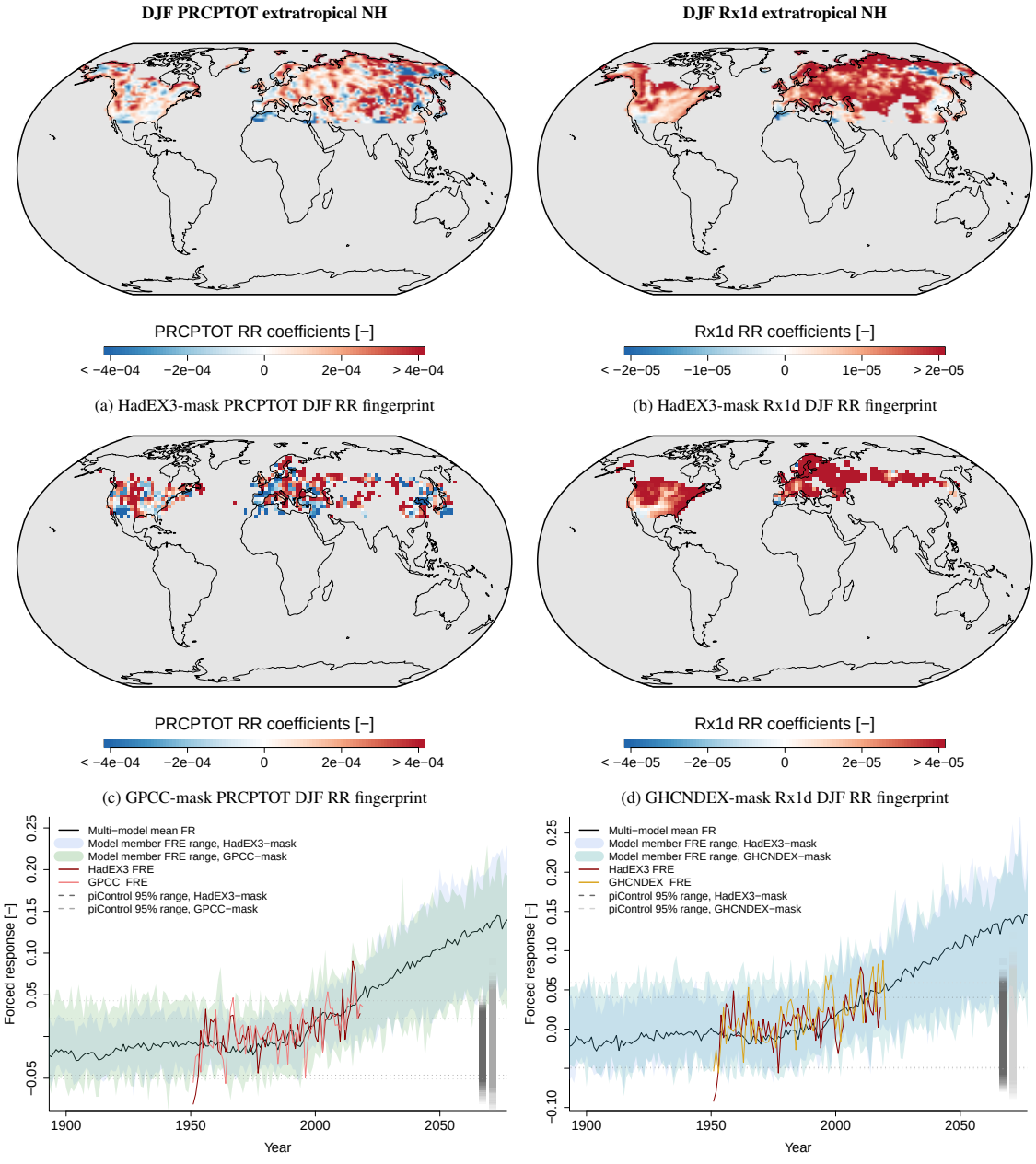
(d) GHCNDEX-mask Rx1d annual RR fingerprint

(e) GPCC-mask PRCPTOT annual RR fingerprint

(f) PRCPTOT annual forced response estimates, extratropical NH

(g) Rx1d annual forced response estimates, extratropical NH

SI Figure S19: RR fingerprints on all observational coverage masks (a, b, c, d, e) and forced response estimates (f,g) as in main figure 3 but for annual extratropical NH PRCPTOT and Rx1d.



(a) HadEX3-mask PRCPTOT DJF RR fingerprint

(b) HadEX3-mask Rx1d DJF RR fingerprint

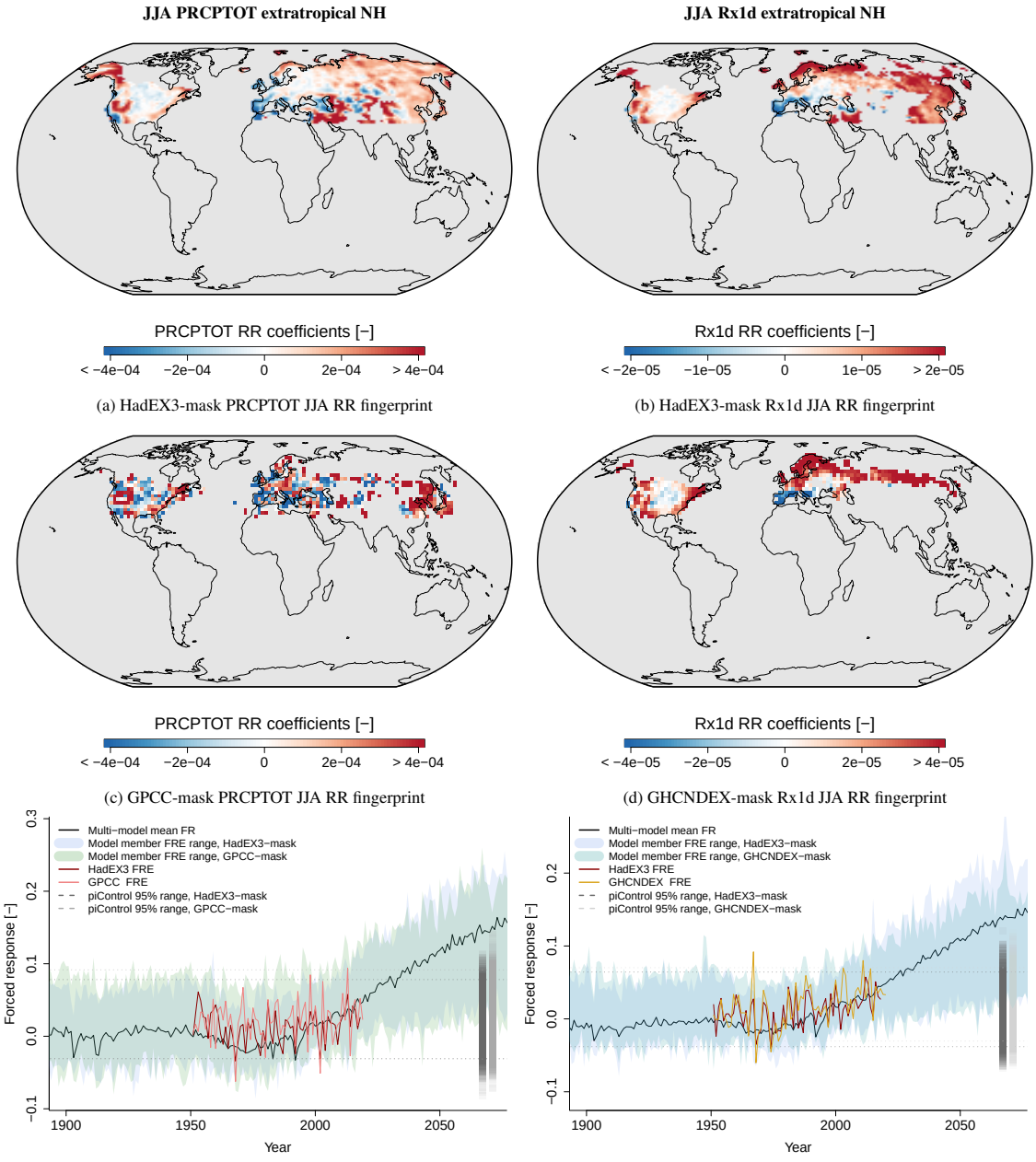
(c) GPCC-mask PRCPTOT DJF RR fingerprint

(d) GHCNDEX-mask Rx1d DJF RR fingerprint

(e) PRCPTOT DJF forced response estimates, extratropical NH

(f) Rx1d DJF forced response estimates, extratropical NH

SI Figure S20: RR fingerprints on all observational coverage masks (a, b, c, d) and forced response estimates (e,f) as in main figure 3 but for winter (DJF) extratropical NH PRCPTOT and Rx1d.



(a) HadEX3-mask PRCPTOT JJA RR fingerprint

(b) HadEX3-mask Rx1d JJA RR fingerprint

(c) GPCC-mask PRCPTOT JJA RR fingerprint

(d) GHCNDEX-mask Rx1d JJA RR fingerprint

(e) PRCPTOT JJA forced response estimates, extratropical NH

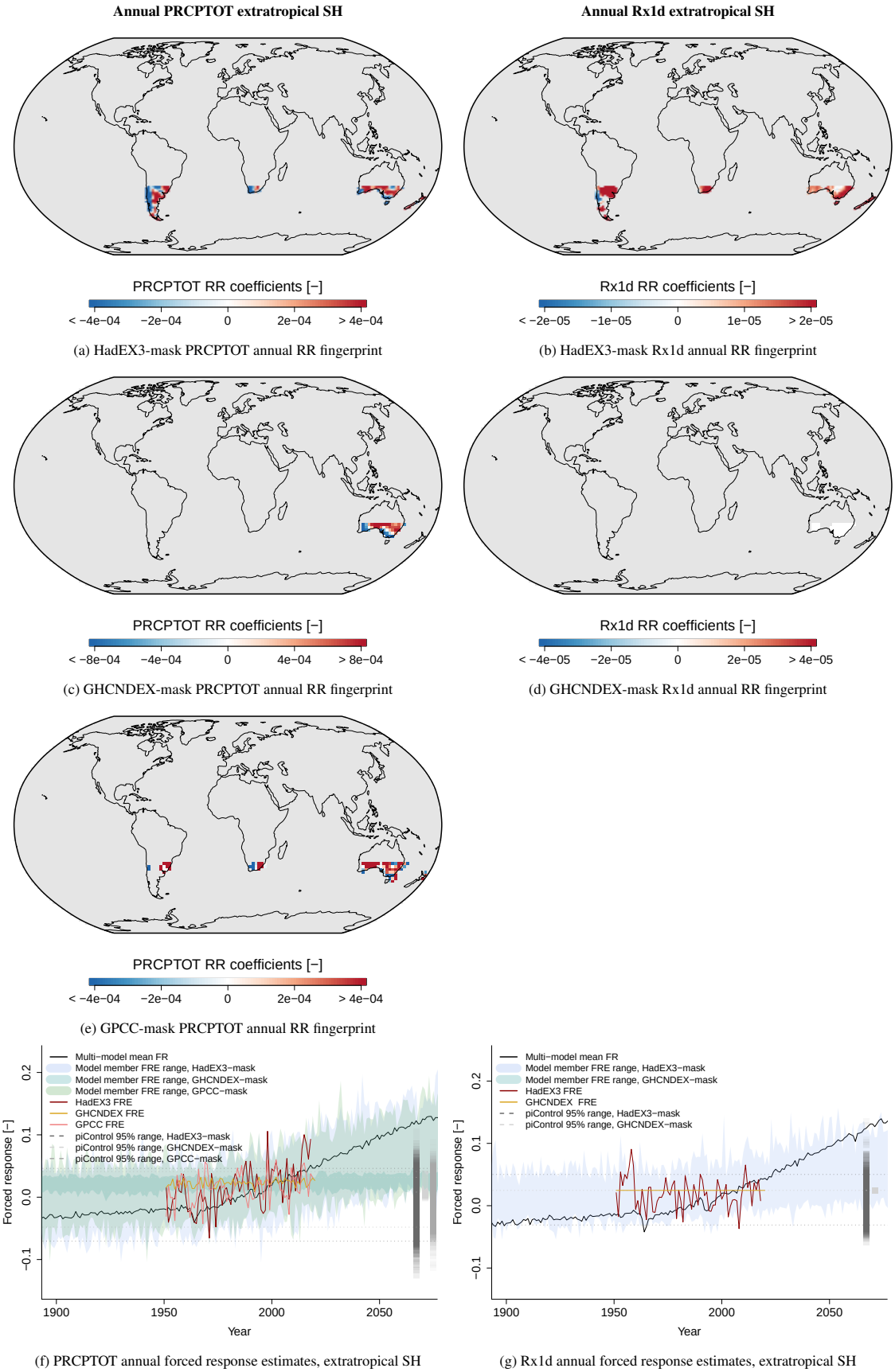
(f) Rx1d JJA forced response estimates, extratropical NH

SI Figure S21: RR fingerprints on all observational coverage masks (a, b, c, d) and forced response estimates (e,f) as in main figure 3 but for summer (JJA) extratropical NH PRCPTOT and Rx1d.

Extratropical Southern Hemisphere, annual and seasonal

Annual PRCPTOT forced response estimates in the SH (figure S22f) for HadEX3 and GPCC show a trend that aligns with the (SH-specific) forced response target, which is impressive given the low coverage. GHCNDEXm however, does not contain enough information to estimate an forced response estimate, and for the other two datasets, variability around the trend is significantly larger than for global and NH annual forced response estimates. For Rx1d (figure S22g), neither of the two datasets contains enough information to estimate the simulated forced trend. Both natural variability masking the trend in observations and the low observational coverage can contribute to the lack of observed forced trend.

For both PRCPTOT and RX1d, and both DJF and JJA in the Southern extratropics the coverage (and perhaps also simply the landmass) is too low to construct RR fingerprints that can predict the forced response; both forced response estimates from models as well as from observations do not capture the multi-model forced response best estimate (target). The multi-model forced response best estimate does in fact show a clear long term increasing trend, meaning that forced changes in PRCPTOT and Rx1d in the SH are expected (to be present already), however, these may be apparent over oceans primarily. The very low coverage of GHCNDEX leads to an RR model without nonzero coefficients, and only an intercept to approach the time mean.



(a) HadEX3-mask PRCPTOT annual RR fingerprint

(b) HadEX3-mask Rx1d annual RR fingerprint

(c) GHCNDEX-mask PRCPTOT annual RR fingerprint

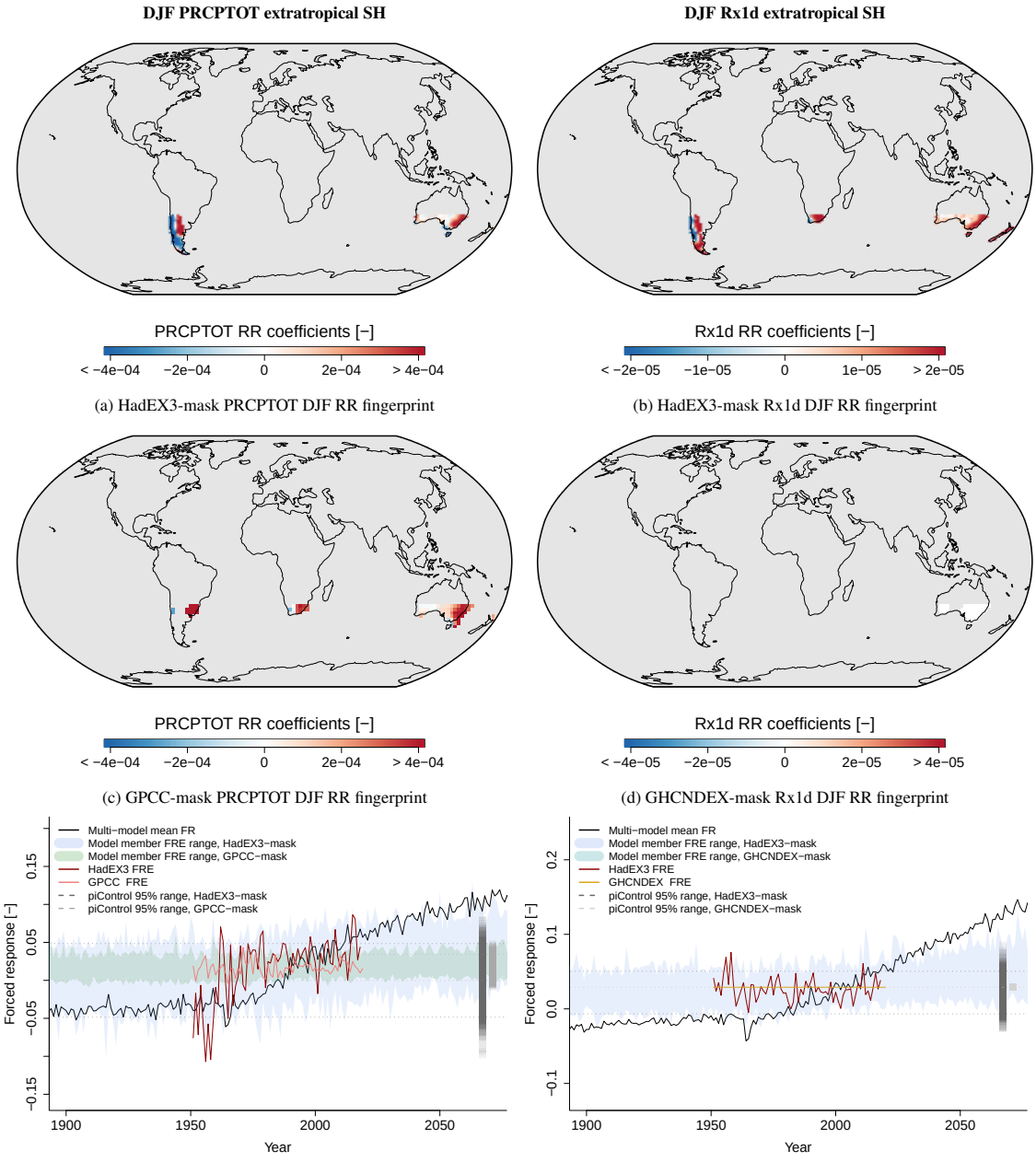
(d) GHCNDEX-mask Rx1d annual RR fingerprint

(e) GPCC-mask PRCPTOT annual RR fingerprint

(f) PRCPTOT annual forced response estimates, extratropical SH

(g) Rx1d annual forced response estimates, extratropical SH

SI Figure S22: RR fingerprints on all observational coverage masks (a, b, c, d) and forced response estimates (e,f) as in main figure 3 but for annual extratropical SH PRCPTOT and Rx1d.



(a) HadEX3-mask PRCPTOT DJF RR fingerprint

(b) HadEX3-mask Rx1d DJF RR fingerprint

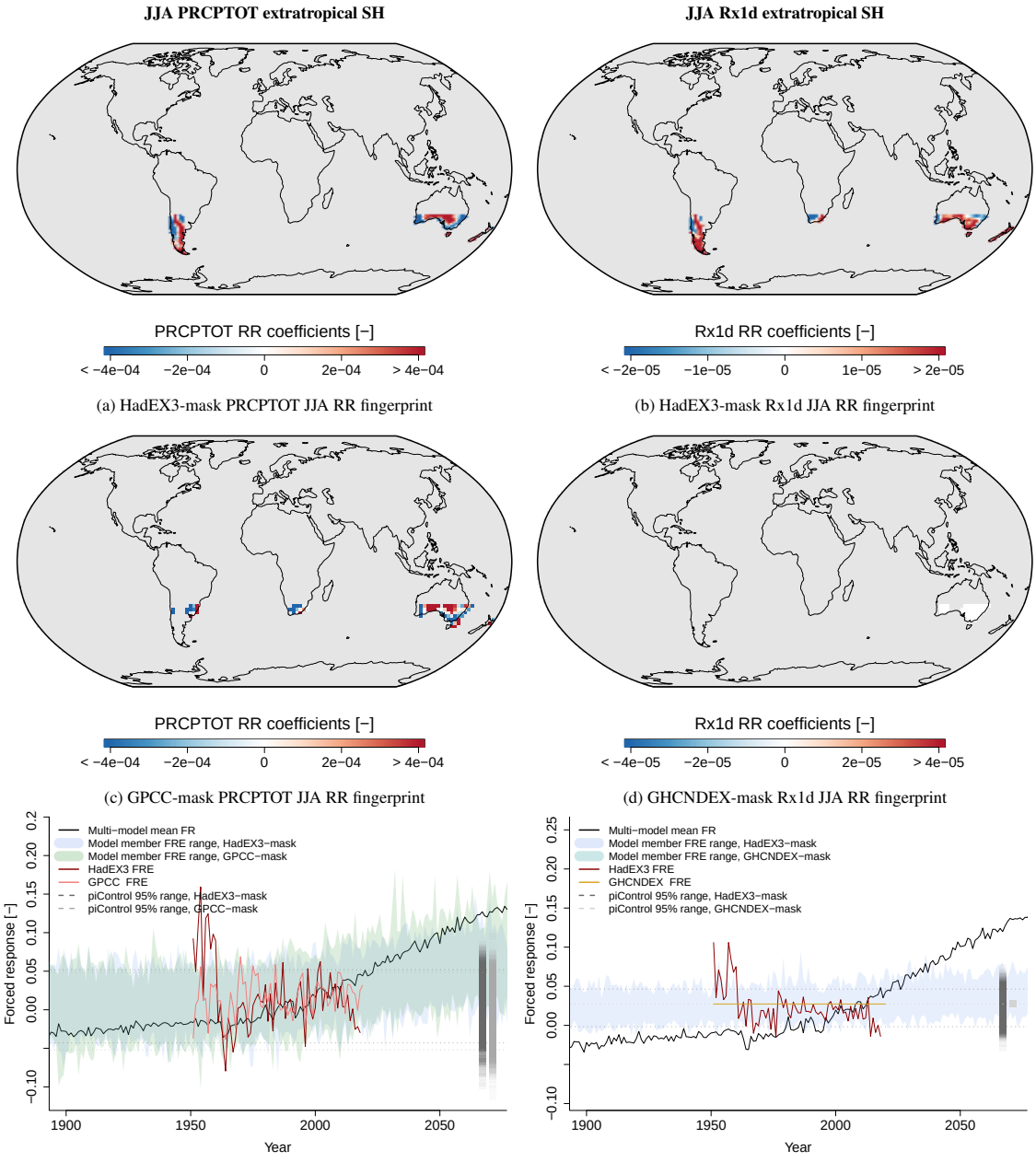
(c) GPCC-mask PRCPTOT DJF RR fingerprint

(d) GHCNDEX-mask Rx1d DJF RR fingerprint

(e) PRCPTOT DJF forced response estimates, extratropical SH

(f) Rx1d DJF forced response estimates, extratropical SH

SI Figure S23: RR fingerprints on all observational coverage masks (a, b, c, d) and forced response estimates (e,f) as in main figure 3 but for summer (DJF) extratropical SH PRCPTOT and Rx1d.



(a) HadEX3-mask PRCPTOT JJA RR fingerprint

(b) HadEX3-mask Rx1d JJA RR fingerprint

(c) GPCC-mask PRCPTOT JJA RR fingerprint

(d) GHCNDEX-mask Rx1d JJA RR fingerprint

(e) PRCPTOT JJA forced response estimates, extratropical SH

(f) Rx1d JJA forced response estimates, extratropical SH

SI Figure S24: RR fingerprints on all observational coverage masks (a, b, c, d) and forced response estimates (e,f) as in main figure 3 but for winter (JJA) extratropical SH PRCPTOT and Rx1d.

Code and data availability. Data and code used can be found in the main publication, as well CMIP6 model data and observational datasets and their sources.

References

Allan, R. P. and Soden, B. J.: Atmospheric warming and the amplification of precipitation extremes, Science, 321, 1481–1484, https://doi.org/https://10.1126/science.1160787, 2008.

Bador, M., Boé, J., Terray, L., Alexander, L. V., Baker, A., Bellucci, A., Haarsma, R., Koenigk, T., Moine, M.-P., Lohmann, K., Putrasahan, D. A., Roberts, C., Roberts, M., Scoccimarro, E., Schiemann, R., Seddon, J., Senan, R., Valcke, S., and Vanniere, B.: Impact of Higher Spatial Atmospheric Resolution on Precipitation Extremes Over Land in Global Climate Models, Journal of Geophysical Research: Atmospheres, 125, e2019JD032 184, https://doi.org/10.1029/2019JD032184, 2020.

Borodina, A., Fischer, E. M., and Knutti, R.: Models are likely to underestimate increase in heavy rainfall in the extratropical regions with high rainfall intensity, Geophysical Research Letters, 44, 7401–7409, https://doi.org/10.1002/2017GL074530, 2017.

Cowtan, K. and Way, R. G.: Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends, Quarterly Journal of the Royal Meteorological Society, 140, 1935–1944, https://doi.org/10.1002/qj.2297, 2014.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geoscientific Model Development, 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

Fischer, E. M. and Knutti, R.: Detection of spatially aggregated changes in temperature and precipitation extremes, Geophysical Research Letters, 41, 547–554, https://doi.org/10.1002/2013GL058499, 2014.

Fischer, E. M. and Knutti, R.: Observed heavy precipitation increase confirms theory and early models, Nature Climate Change, 6, 986–991, https://doi.org/10.1038/nclimate3110, 2016.

Friedman, J. H., Hastie, T., and Tibshirani, R.: Regularization Paths for Generalized Linear Models via Coordinate Descent, Journal of Statistical Software, 33, 1–22, https://doi.org/10.18637/jss.v033.i01, 2010.

Held, I. M. and Soden, B. J.: Robust Responses of the Hydrological Cycle to Global Warming, Journal of Climate, 19, 5686 – 5699, https://doi.org/10.1175/JCLI3990.1, 2006.

Knutson, T. R. and Zeng, F.: Model Assessment of Observed Precipitation Trends over Land Regions: Detectable Human Influences and Possible Low Bias in Model Trends, Journal of Climate, 31, 4617 – 4637, https://doi.org/10.1175/JCLI-D-17-0672.1, 2018.

Kotz, M., Wenz, L., Lange, S., and Levermann, A.: Changes in mean and extreme precipitation scale universally with global mean temperature across and within climate models, EarthArXiv [preprint], https://doi.org/10.31223/X5C631, 2022.

Min, S.-K., Zhang, X., Zwiers, F. W., and Hegerl, G. C.: Human contribution to more-intense precipitation extremes, Nature, 470, 378–381, https://doi.org/10.1038/nature09763, 2011.

Noake, K., Polson, D., Hegerl, G., and Zhang, X.: Changes in seasonal land precipitation during the latter twentieth-century, Geophysical Research Letters, 39, https://doi.org/10.1029/2011GL050405, 2012.

Paik, S., Min, S.-K., Zhang, X., Donat, M. G., King, A. D., and Sun, Q.: Determining the Anthropogenic Greenhouse Gas Contribution to the Observed Intensification of Extreme Precipitation, Geophysical Research Letters, 47, e2019GL086 875, https://doi.org/10.1029/2019GL086875, 2020.

Polson, D., Hegerl, G. C., Zhang, X., and Osborn, T. J.: Causes of Robust Seasonal Land Precipitation Changes, Journal of Climate, 26, 6679 – 6697, https://doi.org/10.1175/JCLI-D-12-00474.1, 2013.

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.

Simon, N., Friedman, J. H., Hastie, T., and Tibshirani, R.: Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent, Journal of Statistical Software, 39, 1–13, https://doi.org/10.18637/jss.v039.i05, 2011.

Stephens, G. L., L'Ecuyer, T., Forbes, R., Gettelmen, A., Golaz, J.-C., Bodas-Salcedo, A., Suzuki, K., Gabriel, P., and Haynes, J.: Dreary state of precipitation in global models, Journal of Geophysical Research: Atmospheres, 115, https://doi.org/10.1029/2010JD014532, 2010.

Stjern, C. W. and Kristjánsson, J. E.: Contrasting Influences of Recent Aerosol Changes on Clouds and Precipitation in Europe and East Asia, Journal of Climate, 28, 8770 – 8790, https://doi.org/10.1175/JCLI-D-14-00837.1, 2015.

Sun, Q., Zwiers, F., Zhang, X., and Yan, J.: Quantifying the Human Influence on the Intensity of Extreme 1- and 5-Day Precipitation Amounts at Global, Continental, and Regional Scales, Journal of Climate, 35, 195 – 210, https://doi.org/10.1175/JCLI-D-21-0028.1, 2022.

Undorf, S., Bollasina, M. A., and Hegerl, G. C.: Impacts of the 1900–74 Increase in Anthropogenic Aerosol Emissions from North America and Europe on Eurasian Summer Climate, Journal of Climate, 31, 8381 – 8399, https://doi.org/10.1175/JCLI-D-17-0850.1, 2018.

Wu, P., Christidis, N., and Stott, P.: Anthropogenic impact on Earth's hydrological cycle, Nature Climate Change, 3, 807–810, https://doi.org/10.1038/nclimate1932, 2013.

Zhang, X., Wan, H., Zwiers, F. W., Hegerl, G. C., and Min, S.-K.: Attributing intensification of precipitation extremes to human influence, Geophysical Research Letters, 40, 5252–5257, https://doi.org/10.1002/grl.51010, 2013.