Supplementary information

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

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.

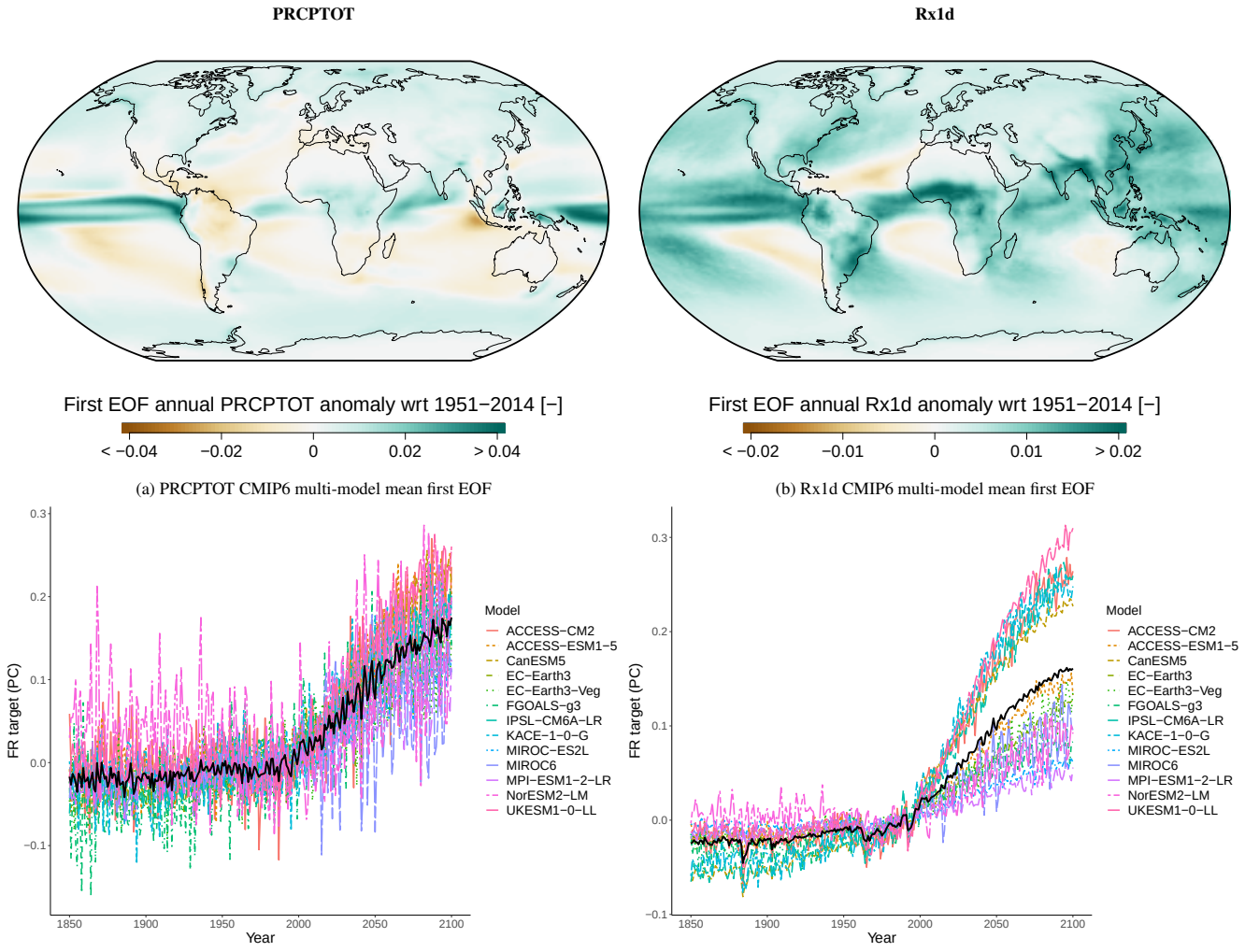
SI Table S1: CMIP6 models and members used for RR model training and model FR estimation. Models for which 450 years of unforced piControl data was available are indicated.

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 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 S2. 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.



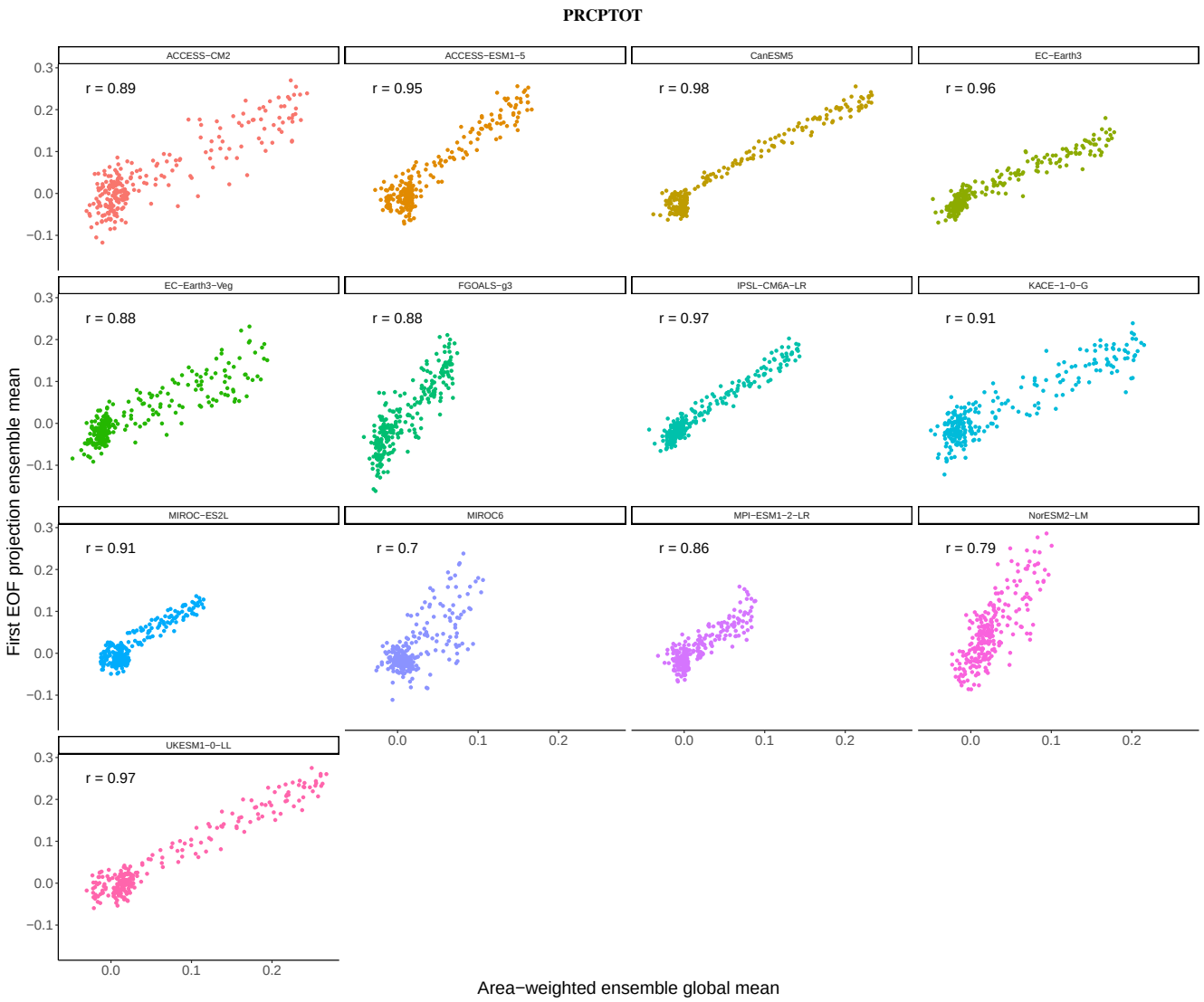
(a) PRCPTOT CMIP6 multi-model mean first EOF

(b) Rx1d CMIP6 multi-model mean first EOF

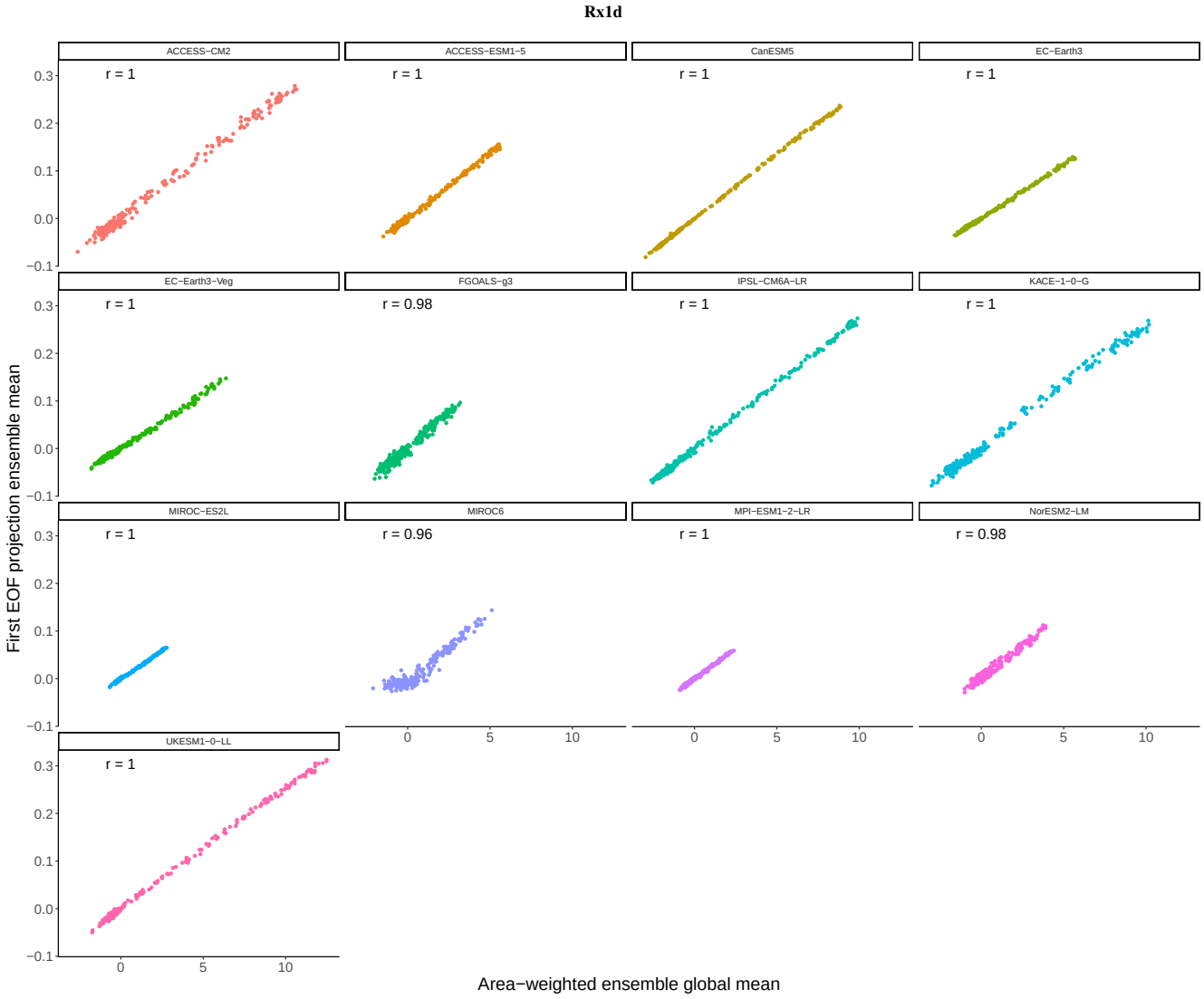
(c) PRCPTOT multi-model mean first principal component and projections onto ensemble means

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

SI Figure S1: 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 S2: 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 four 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:

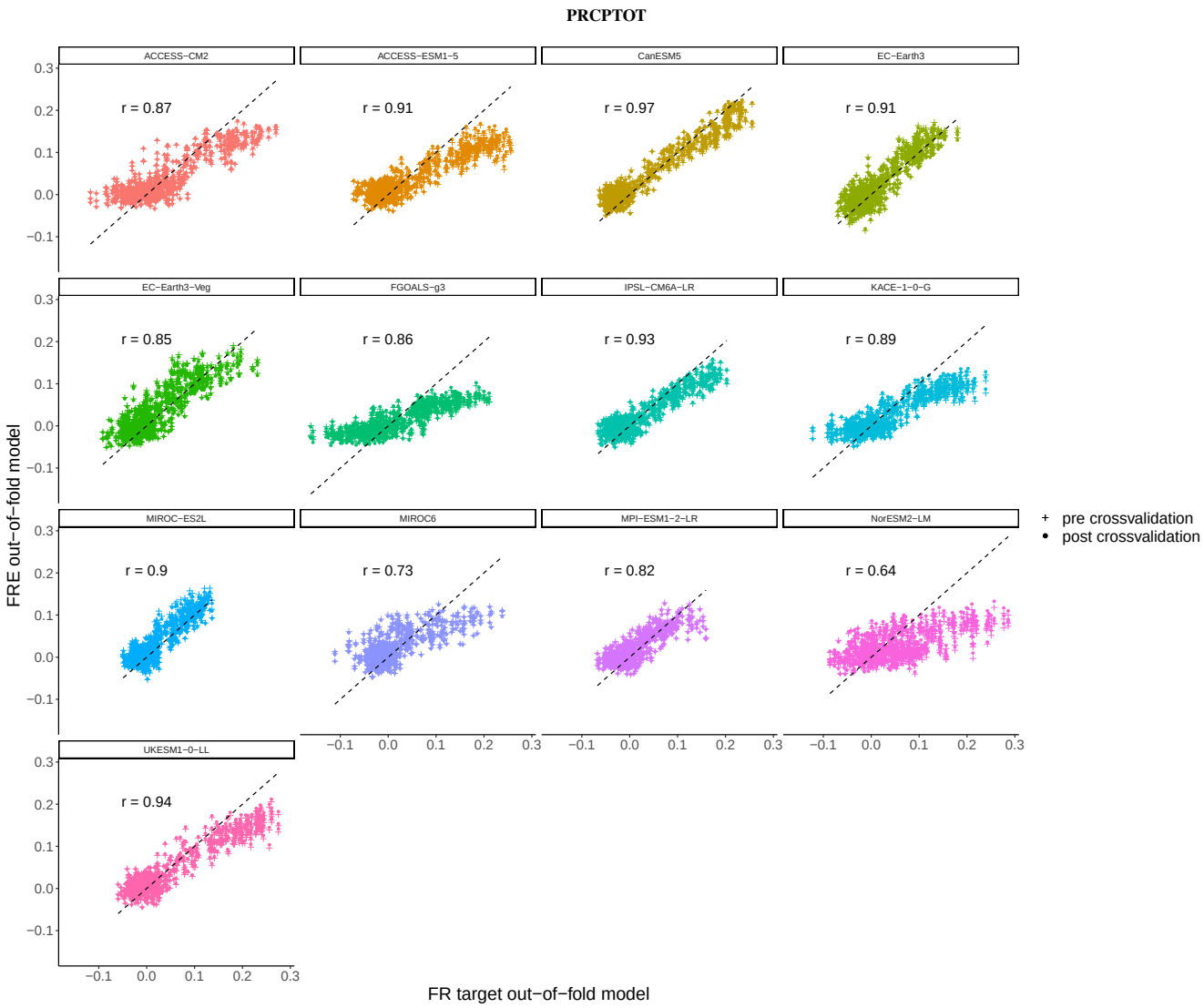
$$\sum_{i=1}^{N} \frac{(\hat{Y} - \text{FRBE})^2}{N} \tag{1}$$

The CVE represents the mean squared error of out-of-fold predictions – i.e. the mean squared error of the FR 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. 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 FRBE. 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}.

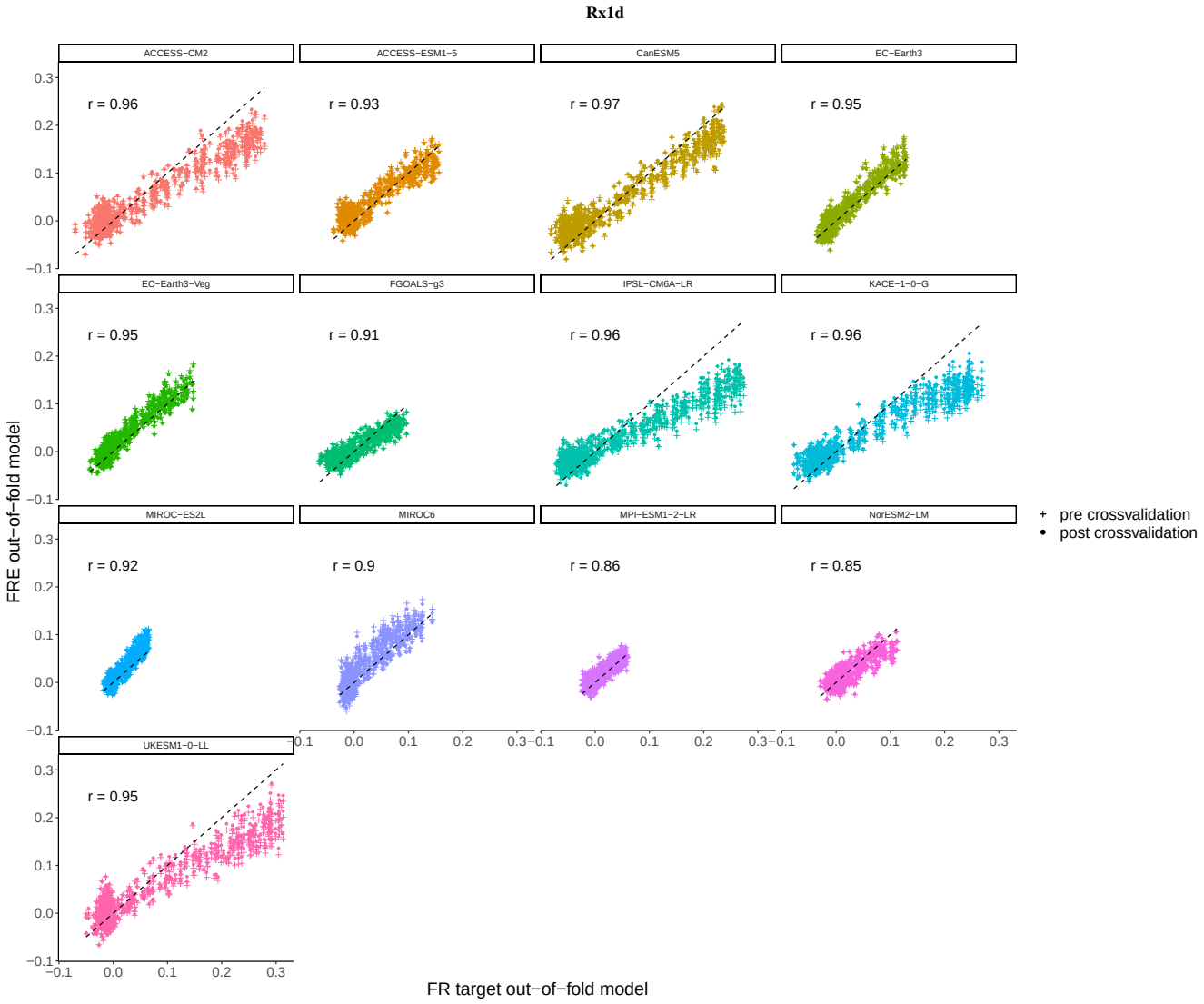
We reason that the most regularised RR model with good performance is a good choice for the detection model, as mentioned in the main text. As both λ_{1se} and λ_{MM} lead to generalisable models and perform well, we select the highest of these two (this differs per case) as our default λ_{sel}.

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 (FREs), and therefore does not jeopardise the relevance of the model FREs, we show model specific correlation plots in figure S3 that include both the pre-crossvalidated FREs (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 FR 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 S1c and S1d. Nonetheless, correlations are high, indicating good performance and generality of the RR models for model FR prediction, although a few individual models have particularly high target spread and/or trends and therefore lower correlations.

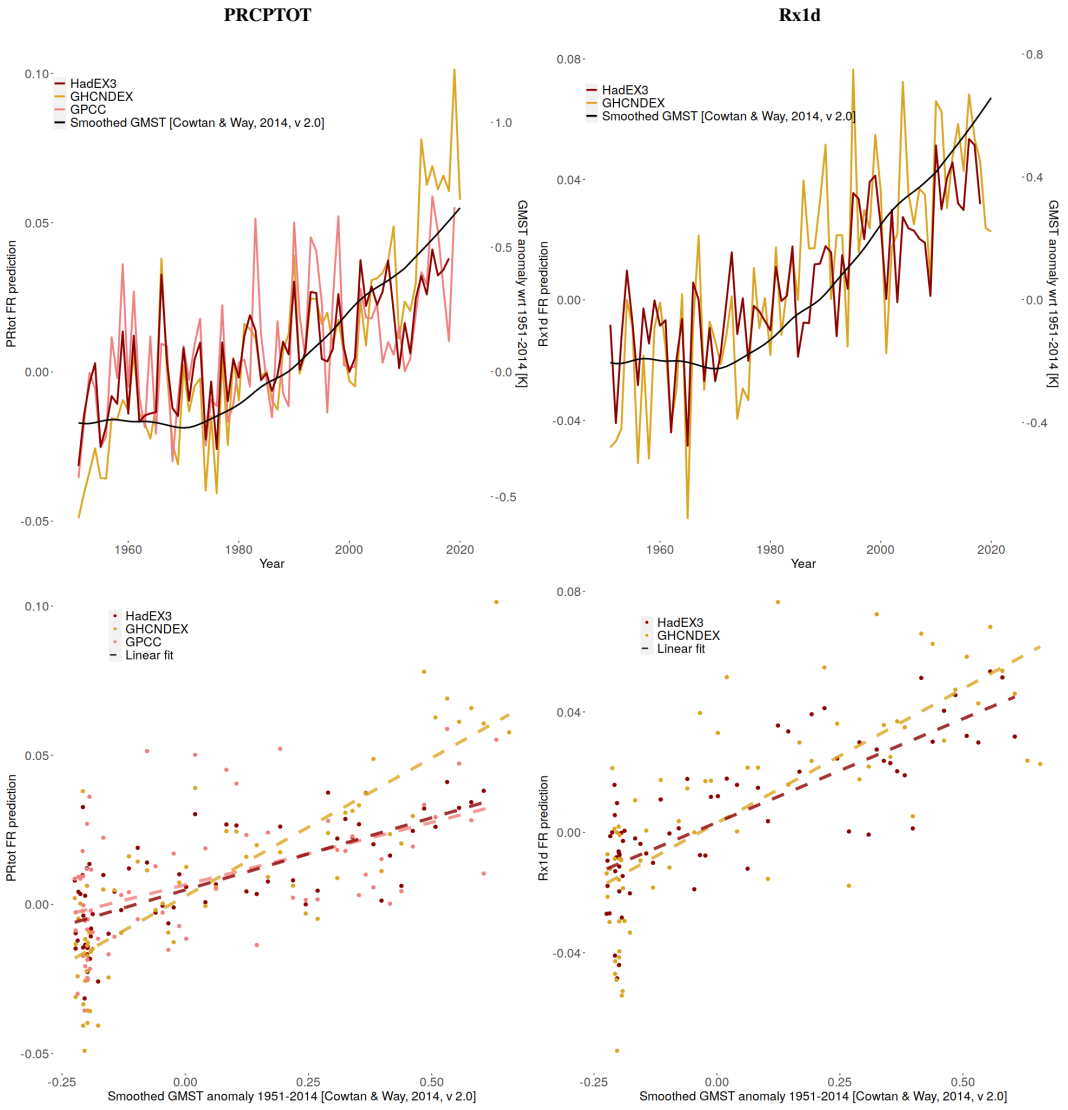


(a) PRCPTOT correlation of EOF-based target with prediction



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

SI Figure S3: Correlations of model-specific EOF-based targets and the FREs obtained from applying the RR model to individual model realisations. The FREs 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 FREs

(b) Rx1d correspondence of smoothed GMST and FREs

SI Figure S4: Top panel: correspondence in shape between observed FREs (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 FREs 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 FREs for the default PRCPTOT and Rx1d cases are given in figure S4. Here, the top panel shows qualitatively how the GMST and the PRCPTOT and Rx1d FREs 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 FRE fits: the FRE 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

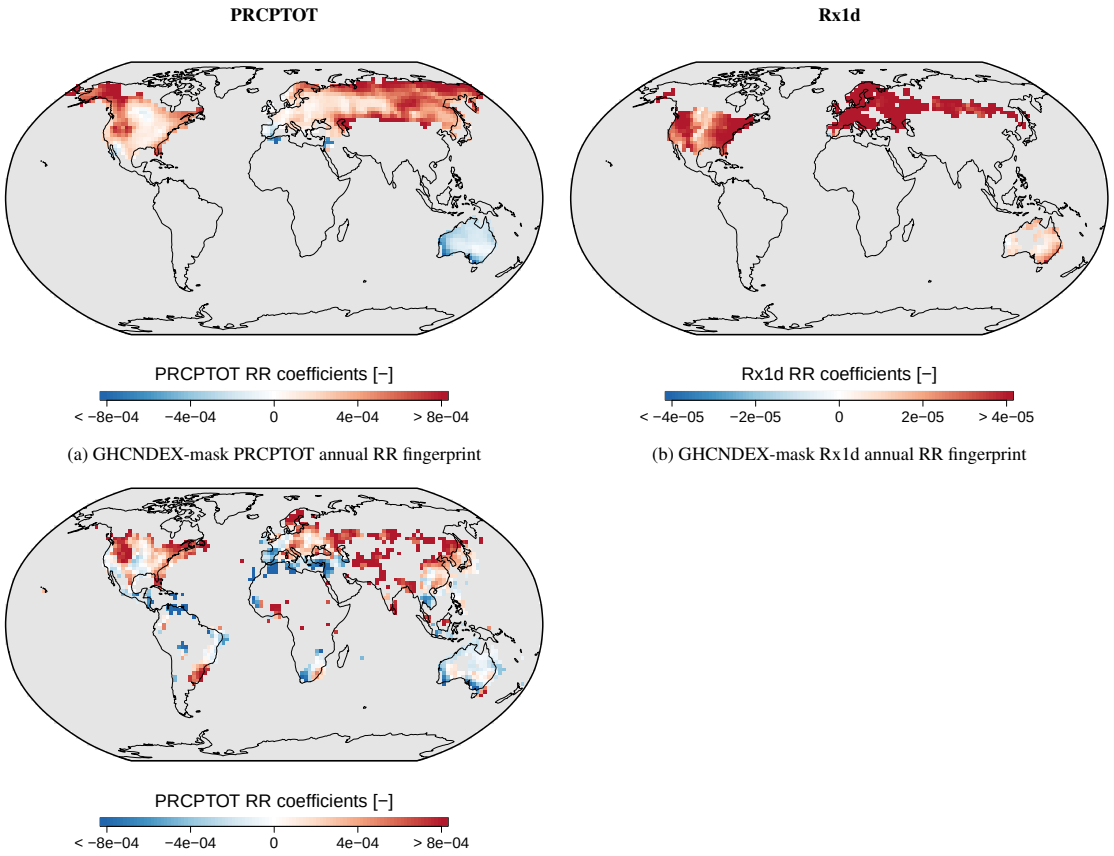
S2.1 Observational dataset and residual consistency

Default case

Figure S5 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 S6 shows the standard deviations of the residuals of the linear trend fits to the FREs over the full period 1951-2014. The standard deviation of the residuals for the observational datasets are shown as vertical lines. For the model FREs, 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 FRE residuals are higher than for HadEX3, and model FRE predictions on the GHCNDEX coverage mask also show larger residuals than model FREs on the HadEX3 coverage mask.

Generally, the residuals of model FREs and observed FREs 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 FRE residuals for piControl compared to forced model FREs, whereas Rx1d FRE residuals of piControl and forced FREs are more consistent. This potentially results from an already measurable increase in variability in PRCPTOT in the forced simulations.



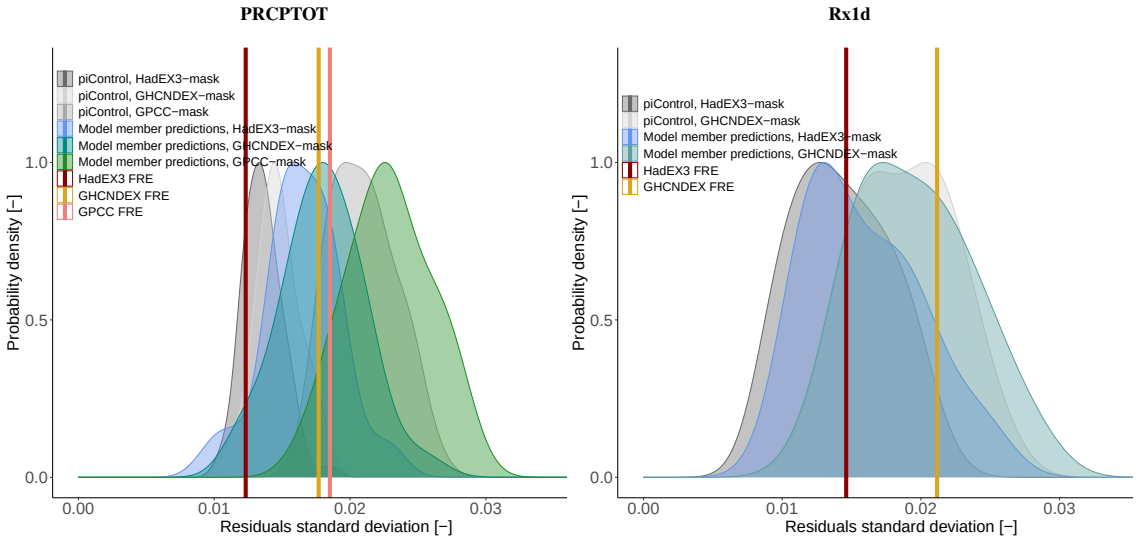
(a) GHCNDEX-mask PRCPTOT annual RR fingerprint

(b) GHCNDEX-mask Rx1d annual RR fingerprint

(c) GPCC-mask PRCPTOT annual RR fingerprint

(d)

SI Figure S5: RR fingerprints for PRCPTOT (a, c), and Rx1d (b) as in main figure 2, for resolution and coverage masks of GHCNDEX and GPCC



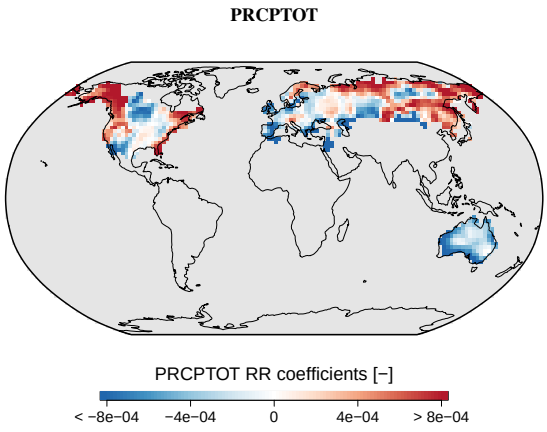
(a) PRCPTOT standard deviation of residuals

(b) Rx1d standard deviation of residuals

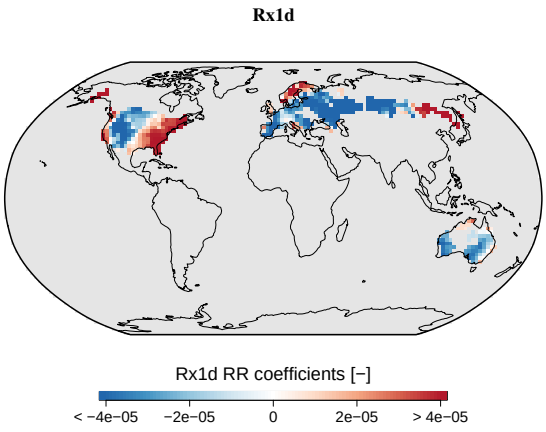
SI Figure S6: Slightly smoothed distribution of standard deviations of the residuals of the linear trend fits to the FREs over the full period 1951-2014 for FREs determined from piControl and forced model simulations on all observational masks (shaded density plots, corresponding FREs are those shown in main figure 2). Standard deviation of the residuals of the linear fit to observed FREs are shown by coloured vertical lines. PRCPTOT (a) and Rx1d (b).

Detrended case

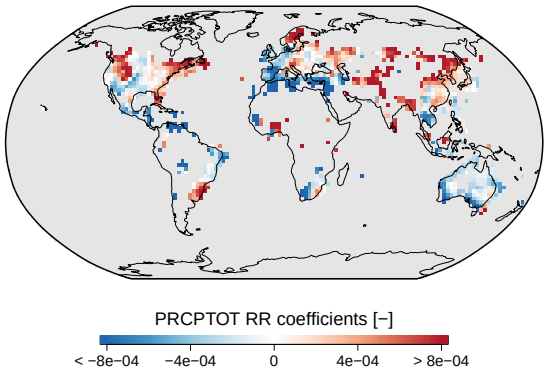
Also for the mean removed case, figure S7, the similarities show. For GHCNDEX Rx1d, however, it can also be seen that the missing coverage in South-East Asia, South America and South Africa, as compared to HadEX3, is detrimental for the RR model's ability to estimate the FR (compare to main figure 3d). The residuals for the detrended case are consistent across models and observations too, as figure S8 shows, apart from GHCNDEX PRCPTOT. The fact that GHCNDEX PRCPTOT FREs show considerably higher variance than model PRCPTOT FREs, implies that the very high FRE trend seen in this case is unreliable.



(a) GHCNDEX-mask PRCPTOT annual RR fingerprint



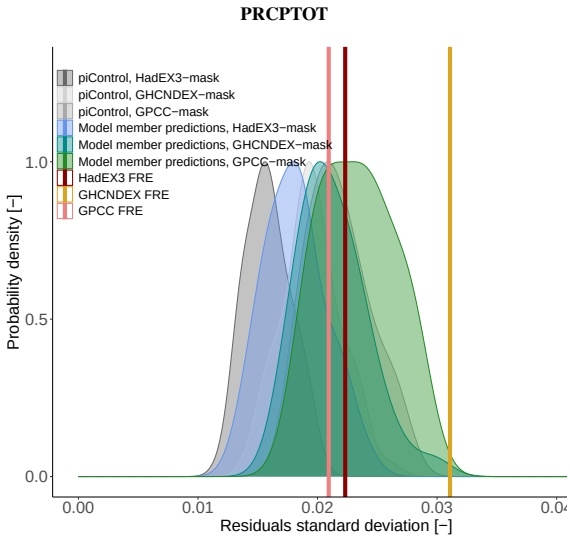
(b) GHCNDEX-mask Rx1d annual RR fingerprint



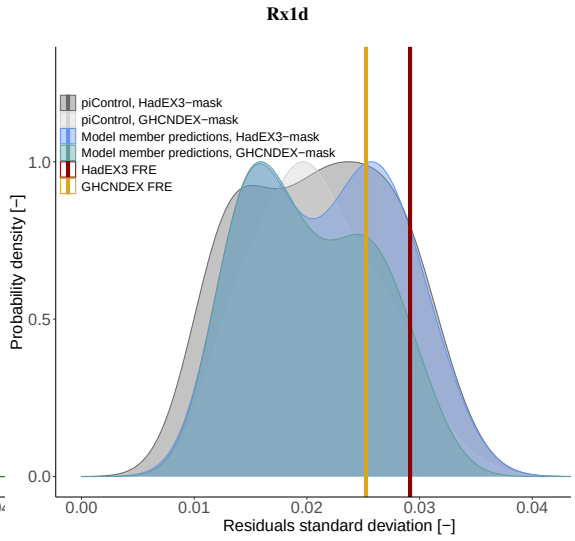
(c) GPCC-mask PRCPTOT annual RR fingerprint

(d)

SI Figure S7: 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

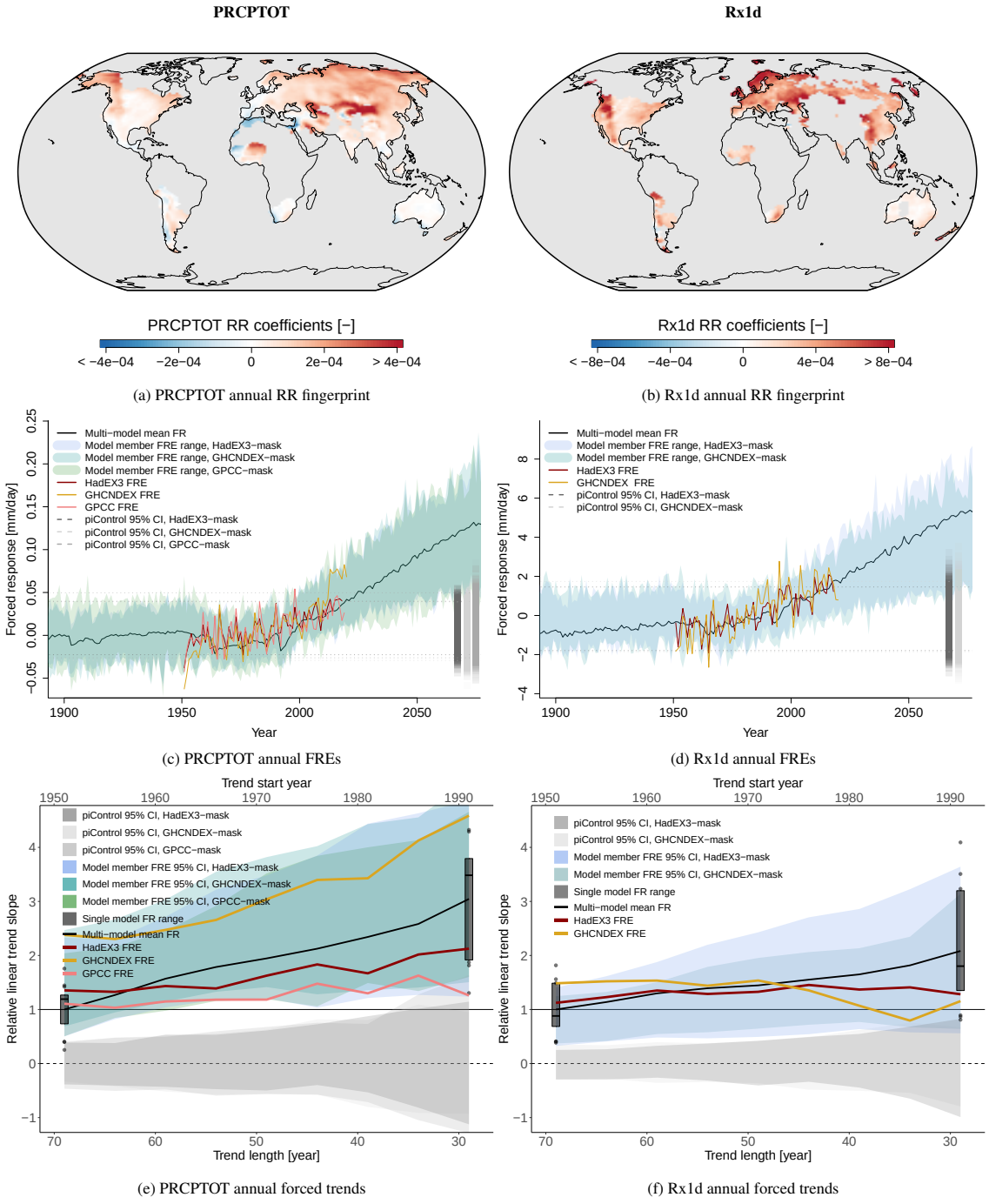


(b) Rx1d standard deviation of residuals, detrended case

SI Figure S8: As figure S6 but for RR models trained on data from which the global mean is subtracted (detrended). Corresponding FREs are those shown in main figure 3.

S2.2 Alternative FR target: global mean

In figure S9, the results of the RR procedure with area-weighted annual global means (model ensemble means) as targets is shown. Comparing figure S9 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 FRBE 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 FREs

(d) Rx1d annual FREs

(e) PRCPTOT annual forced trends

(f) Rx1d annual forced trends

SI Figure S9: As main figure 2 but for RR models trained with area-weighted global mean PRCPTOT and Rx1d as FR target

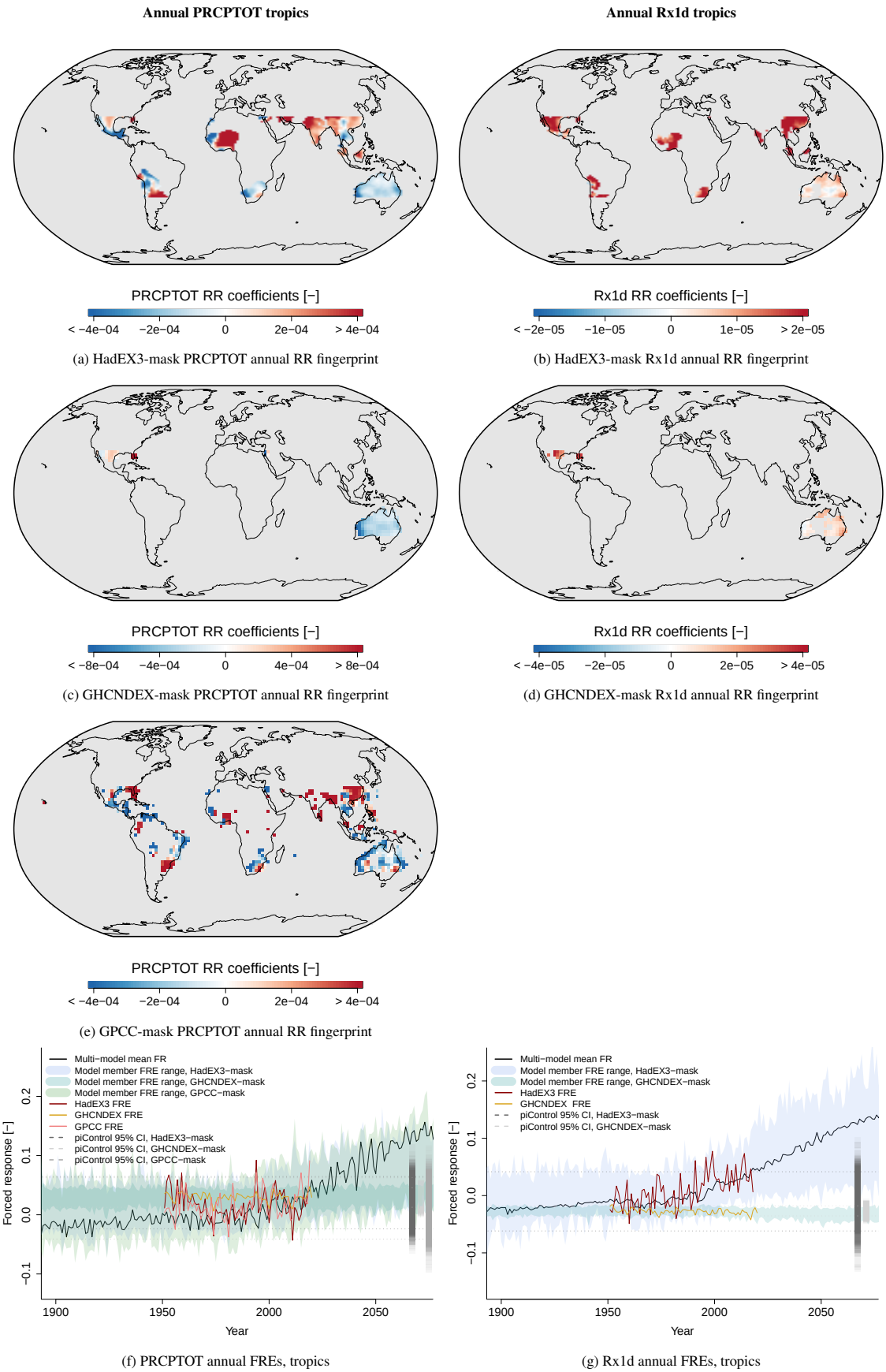
S3 Regional and seasonal analysis

As mentioned in the main text, the Northern Hemisphere (NH) signals make up the largest contribution to primarily to the total FRE. To exemplify this we show fingerprints and FREs 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. 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 assess December-January-February (DJF) and June-July-August (JJA) in the extratropical regions.

For the figures shown, the FR 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 main section 2. Separate EOFs and corresponding FR targets were determined for each region, to capture the region-specific FR in the target.

Tropics, annual

Figure S10 shows the fingerprints for the annual tropics case, all observational coverage masks are shown for comparison, as well as the corresponding FREs. 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 (S10g). 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

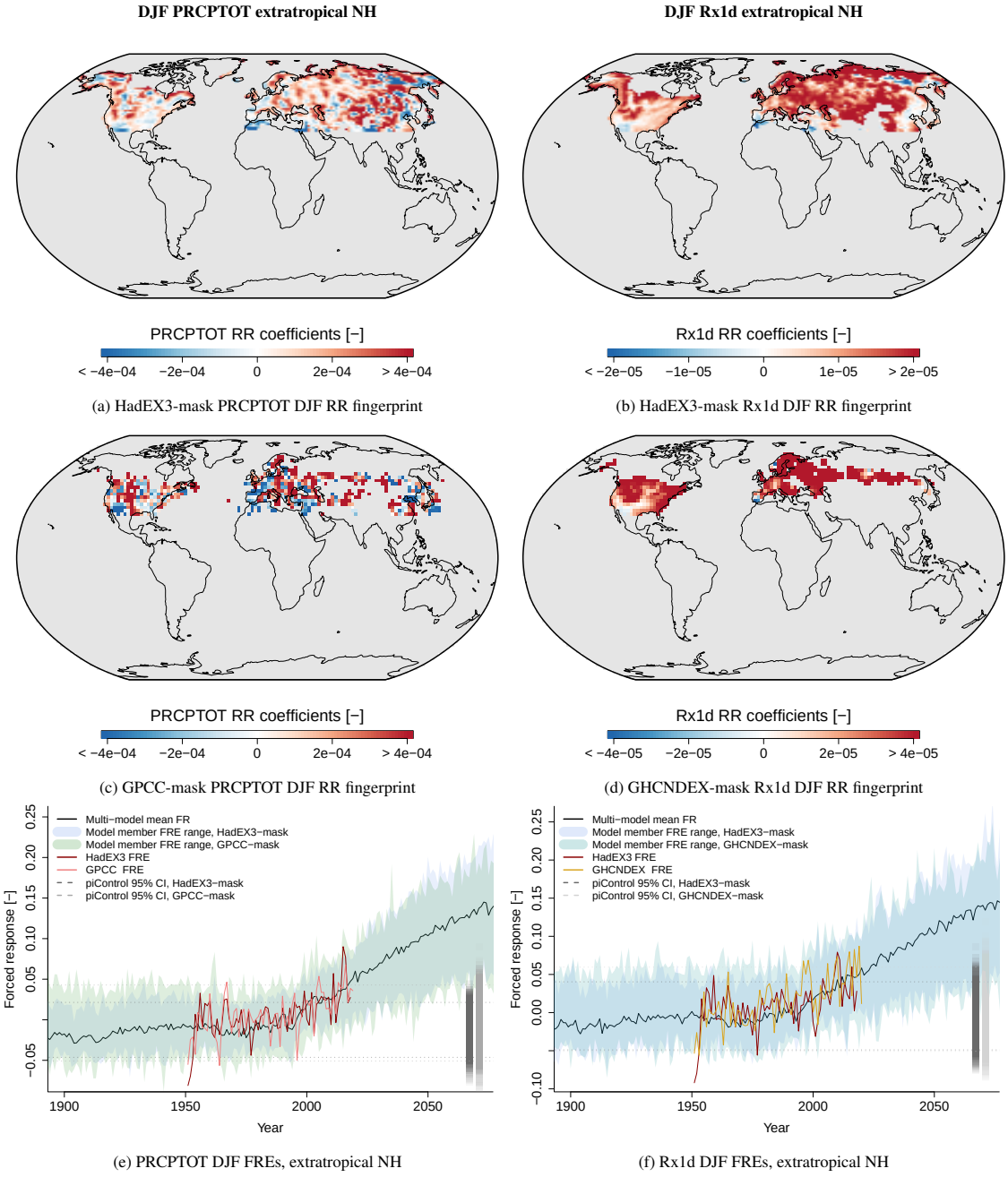
(f) PRCPTOT annual FREs, tropics

(g) Rx1d annual FREs, tropics

SI Figure S10: 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, seasonal

Figure S11 shows the fingerprints and FREs for NH winter (DJF). As mentioned above, the extratropical FR 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 FRE errors and less of an error increase when variance of the FRE 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 FREs (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.



(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

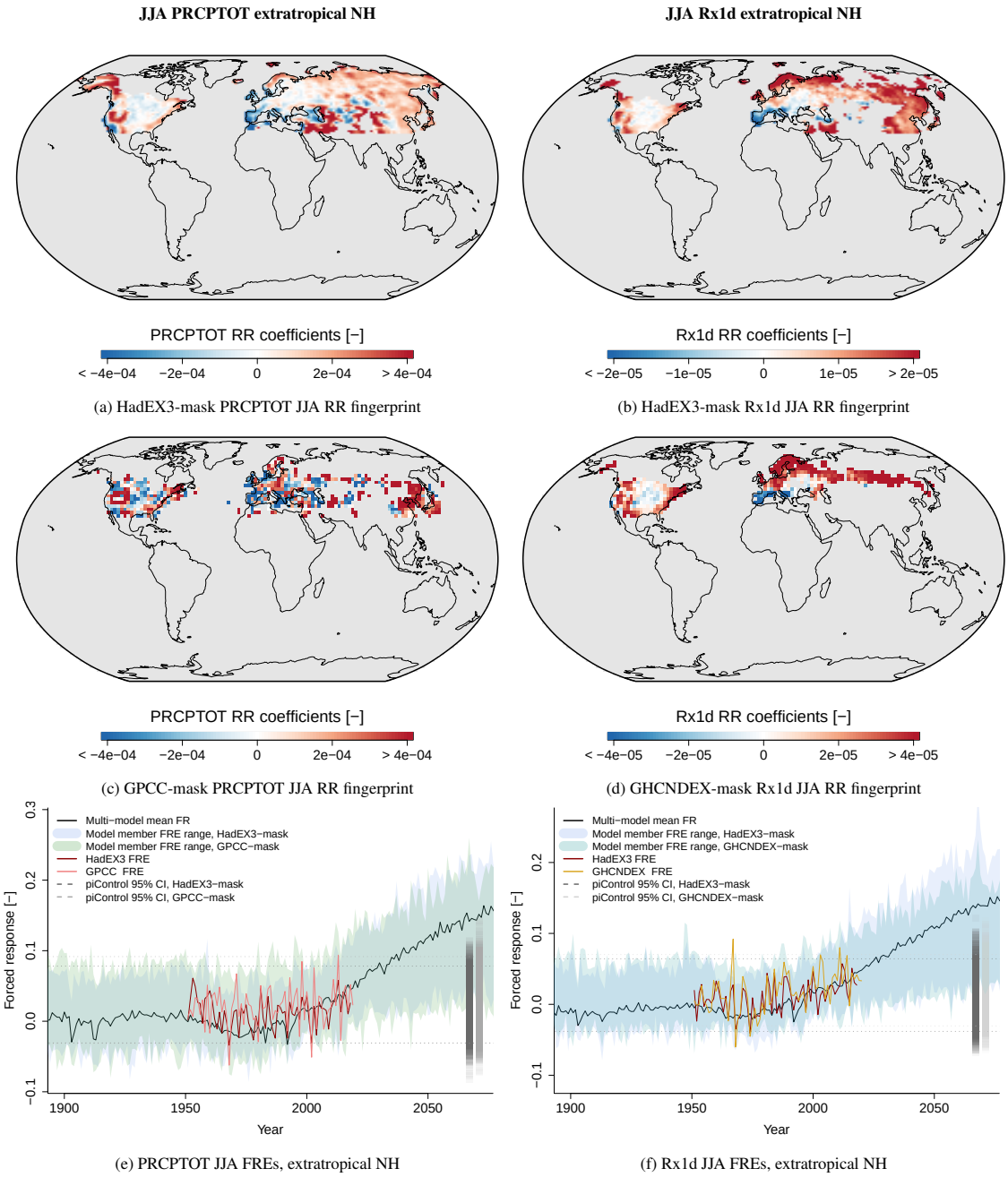
(e) PRCPTOT DJF FREs, extratropical NH

(f) Rx1d DJF FREs, extratropical NH

SI Figure S11: RR fingerprints on all observational coverage masks (a, b, c, d) and forced response estimates (e,f) as in main figure 2 but for winter (DJF) extratropical NH PRCPTOT and Rx1d.

The fingerprints for extratropical NH summer (JJA), figure S12 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 FREs from observations (S12e and S12f) 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 JJA RR fingerprint

(b) HadEX3-mask Rx1d JJA RR fingerprint

(c) GPCC-mask PRCPTOT JJA RR fingerprint

(d) GHCNDEX-mask Rx1d JJA RR fingerprint

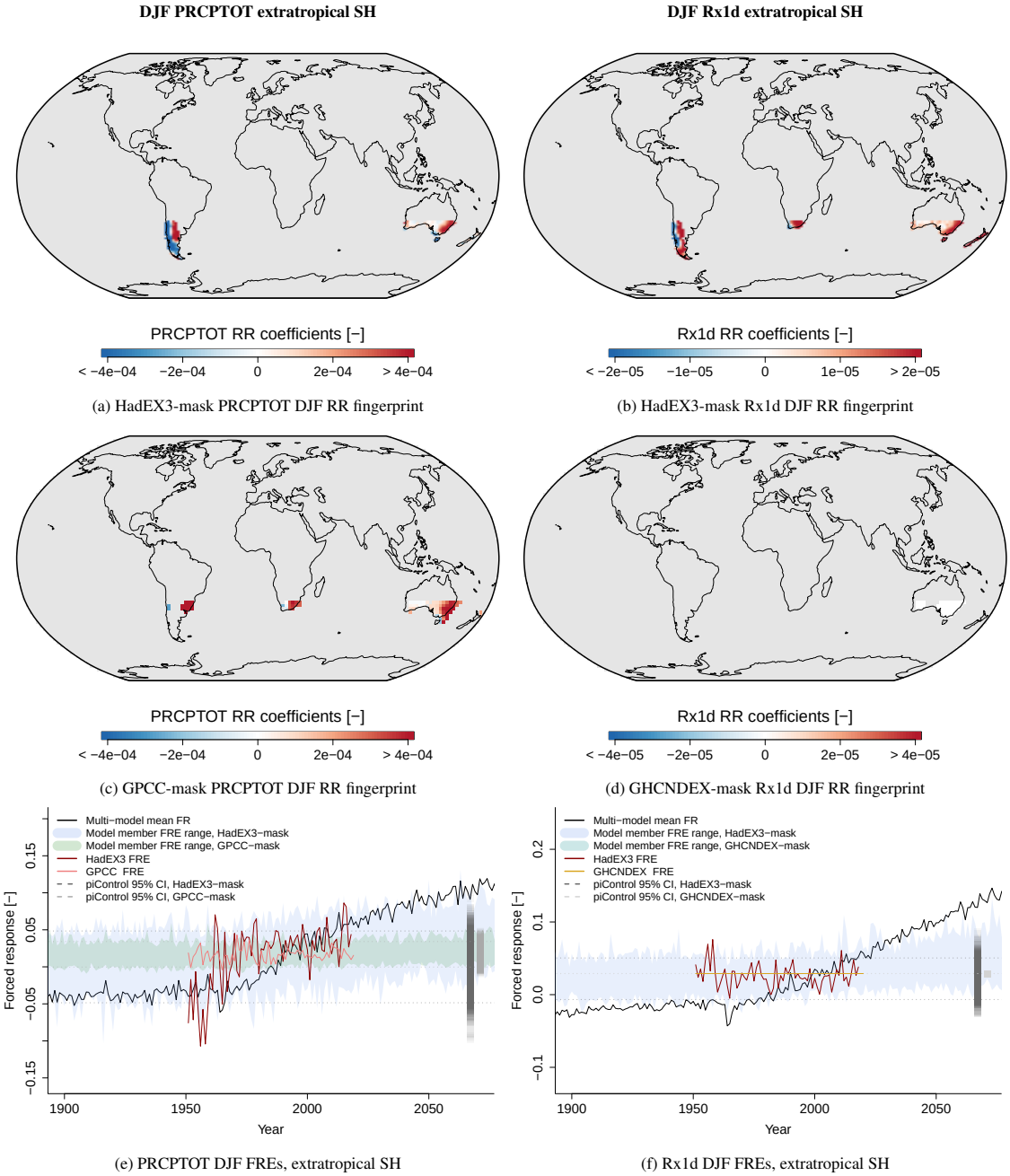
(e) PRCPTOT JJA FREs, extratropical NH

(f) Rx1d JJA FREs, extratropical NH

SI Figure S12: RR fingerprints on all observational coverage masks (a, b, c, d) and forced response estimates (e,f) as in main figure 2 but for summer (JJA) extratropical NH PRCPTOT and Rx1d.

Extratropical Southern Hemisphere, seasonal

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 FR; both FREs from models as well as from observations do not capture the multi-model FRBE (target). The multi-model FRBE 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 DJF RR fingerprint

(b) HadEX3-mask Rx1d DJF RR fingerprint

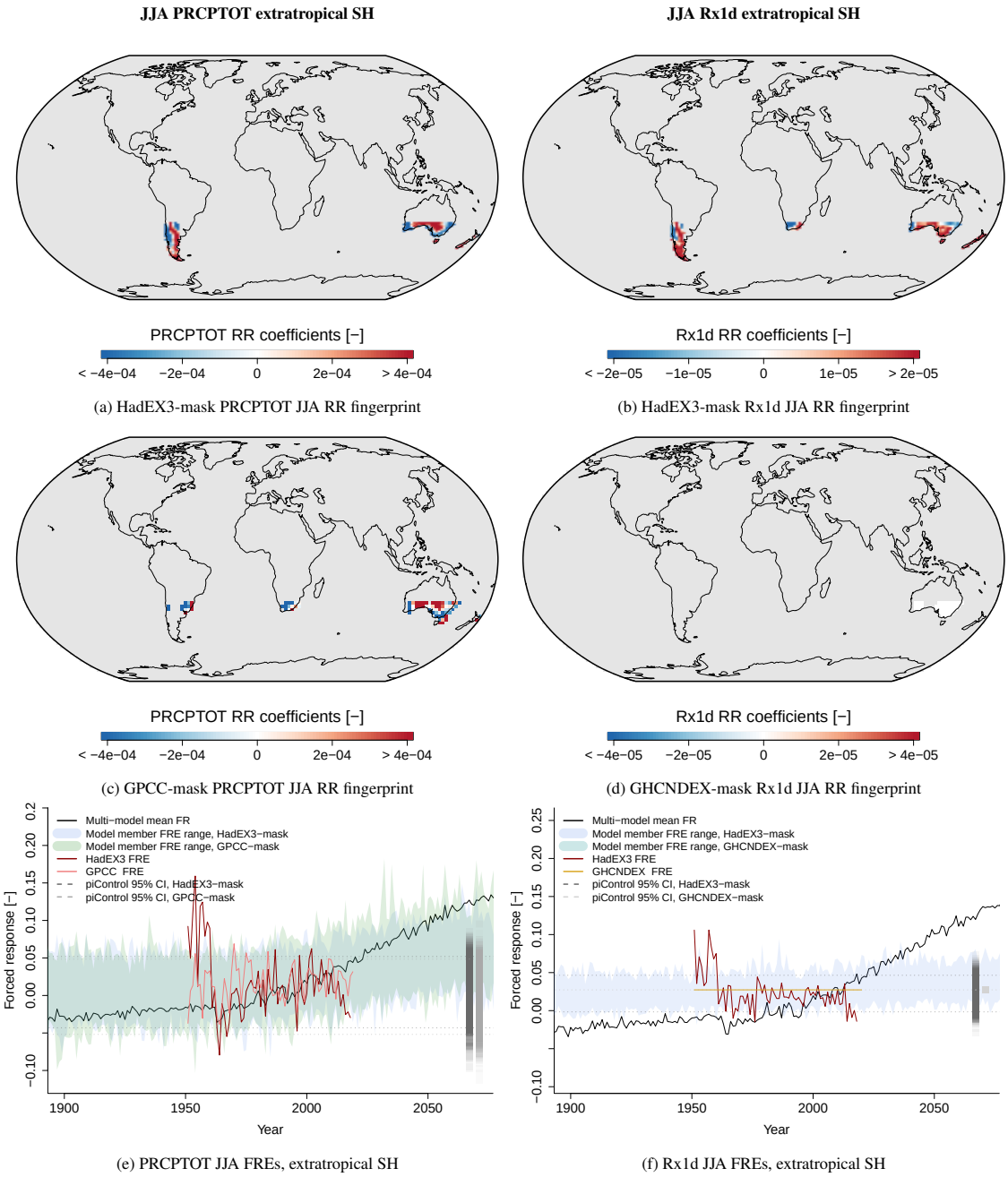
(c) GPCC-mask PRCPTOT DJF RR fingerprint

(d) GHCNDEX-mask Rx1d DJF RR fingerprint

(e) PRCPTOT DJF FREs, extratropical SH

(f) Rx1d DJF FREs, extratropical SH

SI Figure S13: RR fingerprints on all observational coverage masks (a, b, c, d) and forced response estimates (e,f) as in main figure 2 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 FREs, extratropical SH

(f) Rx1d JJA FREs, extratropical SH

SI Figure S14: RR fingerprints on all observational coverage masks (a, b, c, d) and forced response estimates (e,f) as in main figure 2 but for winter (JJA) extratropical SH PRCPTOT and Rx1d.

References

K. Cowtan and R. G. Way. Coverage bias in the hadcrut4 temperature series and its impact on recent temperature trends. Quarterly Journal of the Royal Meteorological Society, 140(683):1935–1944, 2014. doi: 10.1002/qj.2297.

V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. Geoscientific Model Development, 9(5): 1937–1958, 2016. doi: 10.5194/gmd-9-1937-2016.

C. W. Stjern and J. E. Kristjánsson. Contrasting influences of recent aerosol changes on clouds and precipitation in europe and east asia. Journal of Climate, 28(22):8770 – 8790, 2015. doi: 10.1175/JCLI-D-14-00837.1.

S. Undorf, M. A. Bollasina, and G. C. Hegerl. Impacts of the 1900–74 increase in anthropogenic aerosol emissions from north america and europe on eurasian summer climate. Journal of Climate, 31(20):8381 – 8399, 2018. doi: 10.1175/JCLI-D-17-0850.1.