Comments of Referee #2
Response to Referee #2

Overall comments:

This study conducts a signal detection analysis for global changes in mean and extreme precipitation using three observational datasets and CMIP6 multi-model outputs. The authors apply a ridge regression (RR) method to construct fingerprints, which helps increase a signal-to-noise ratio of precipitation change patterns. Results show a robust detection of anthropogenic signals in all observations for both mean and extreme precipitation even when removing global mean trends, further supporting the human-induced intensification of global hydrological cycle. I find this paper very well written with sufficient details provided about methods as well as various sensitivity tests and therefore suggest publication after addressing some minor issues.

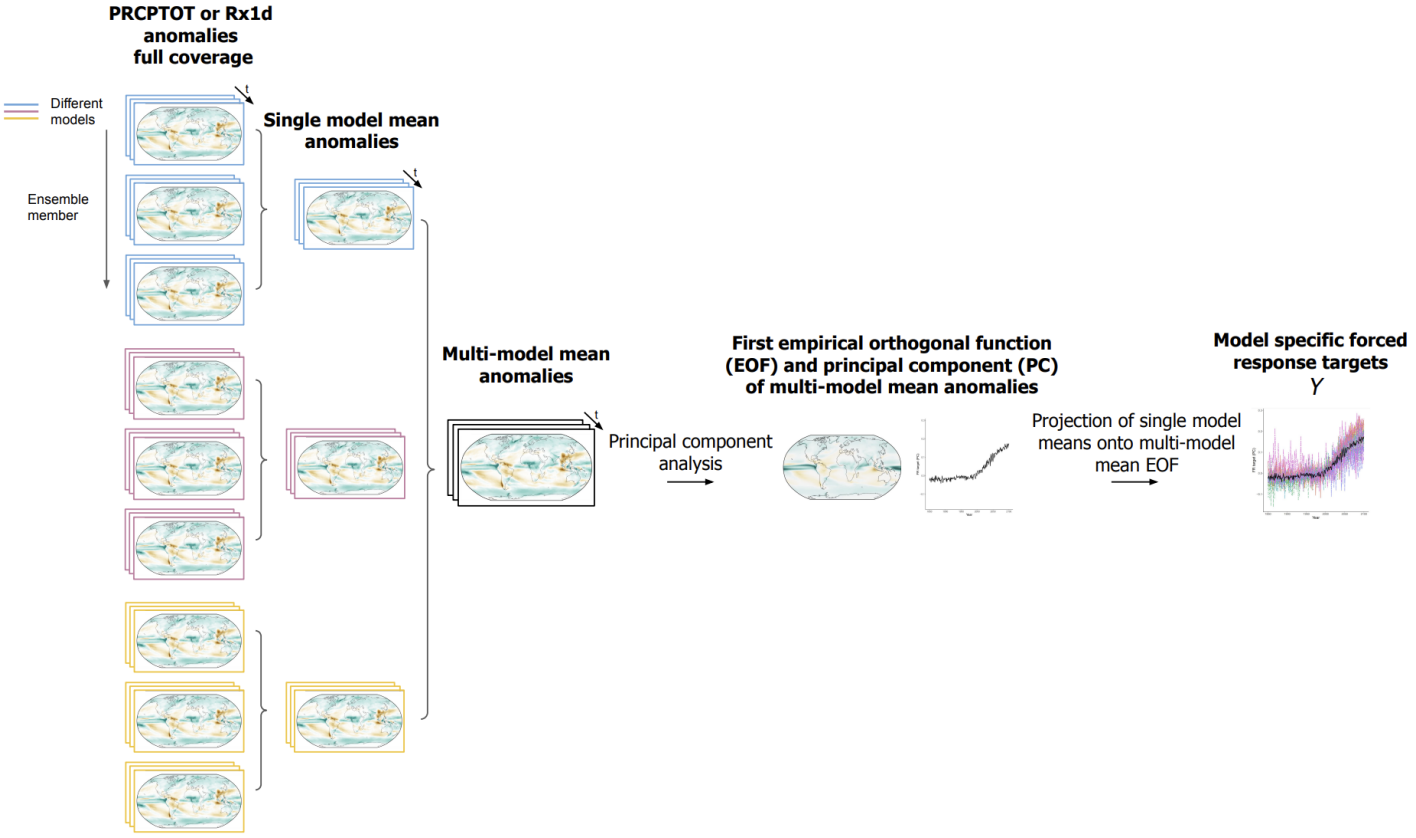
Thanks very much for your kind comments and positive judgment of our manuscript.

Major comments:

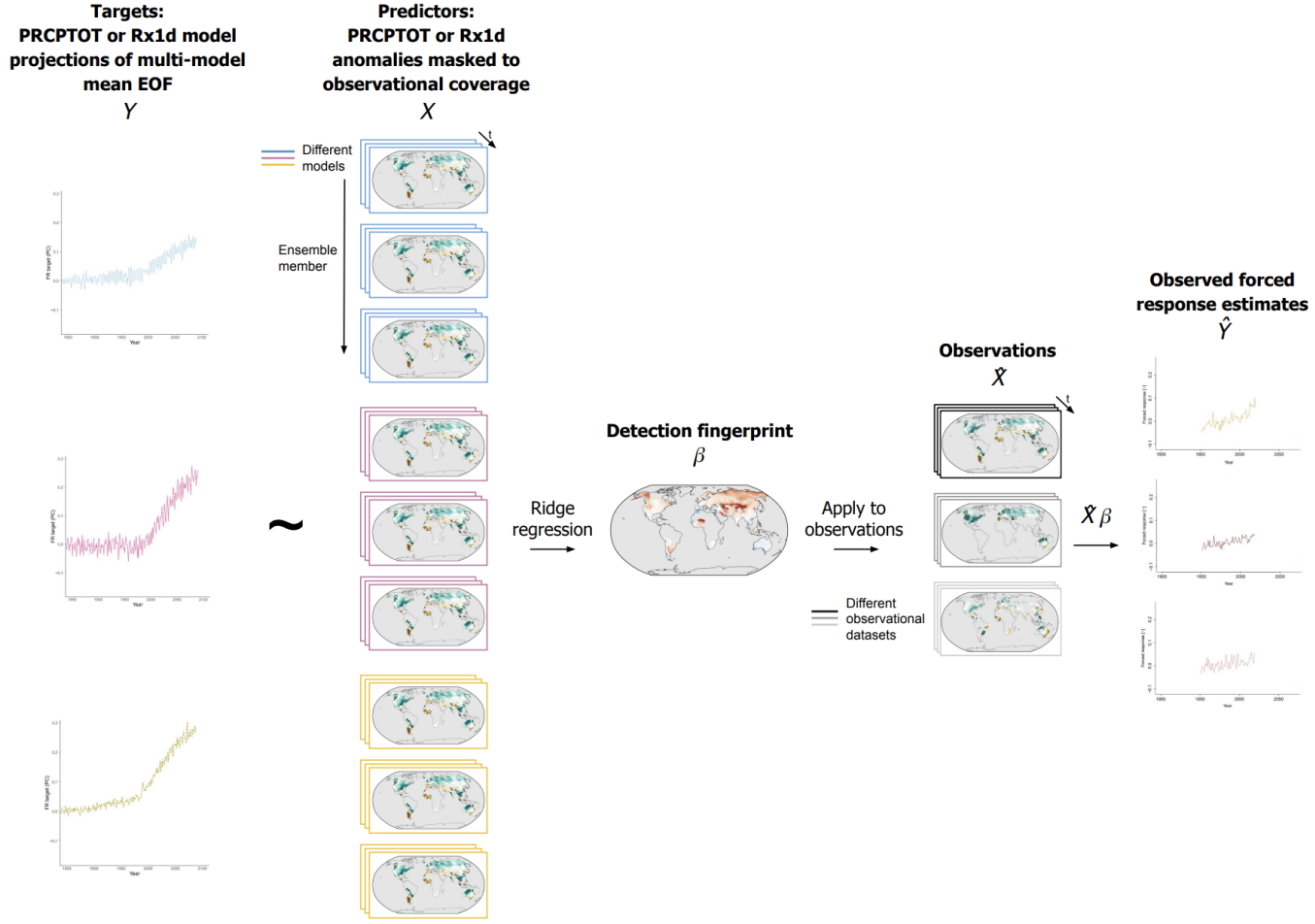
1. Although method details are provided, it would be useful to explain more clearly what are benefits of the attribution approaches employed, including ridge regression, EOF-based metric for target variable, and GMST-based signal estimation. All of these procedures seem to contribute to increase signal-to-noise ratio but how they do and what step is more important. The authors provide some associated results from sensitivity tests but an overall explanation of their method possibly with a schematic would be helpful for readers to understand the contribution of each step to the final signal detection.

We see that the sequence of steps and their relative function with respect to one another can lead to confusion. We like the idea of adding a schematic of the methodology, and we will add this to the supplementary info of a revised paper. We show preliminary drafts of such flowcharts below.

Flowchart part I: Schematic visualisation of determination of ridge regression targets



Flowchart part II: Schematic visualisation of ridge regression procedure and determination of observed forced response estimates



In addition, we add a figure to the supplementary information which allows comparison of the signal-to-noise ratios (SNR) of the procedure performed with

1. EOF based targets (our chosen default) and "optimal" regularisation (λ_sel)
2. Global mean based targets and "optimal" regularisation
3. EOF based targets and minimal regularisation (λ_0).

Comparing 1 and 2 gives an impression of the SNR-effect of using EOF based targets, whereas comparing 1 and 3 shows the SNR-effect of ridge regression (relative to unregularised ordinary least squares).

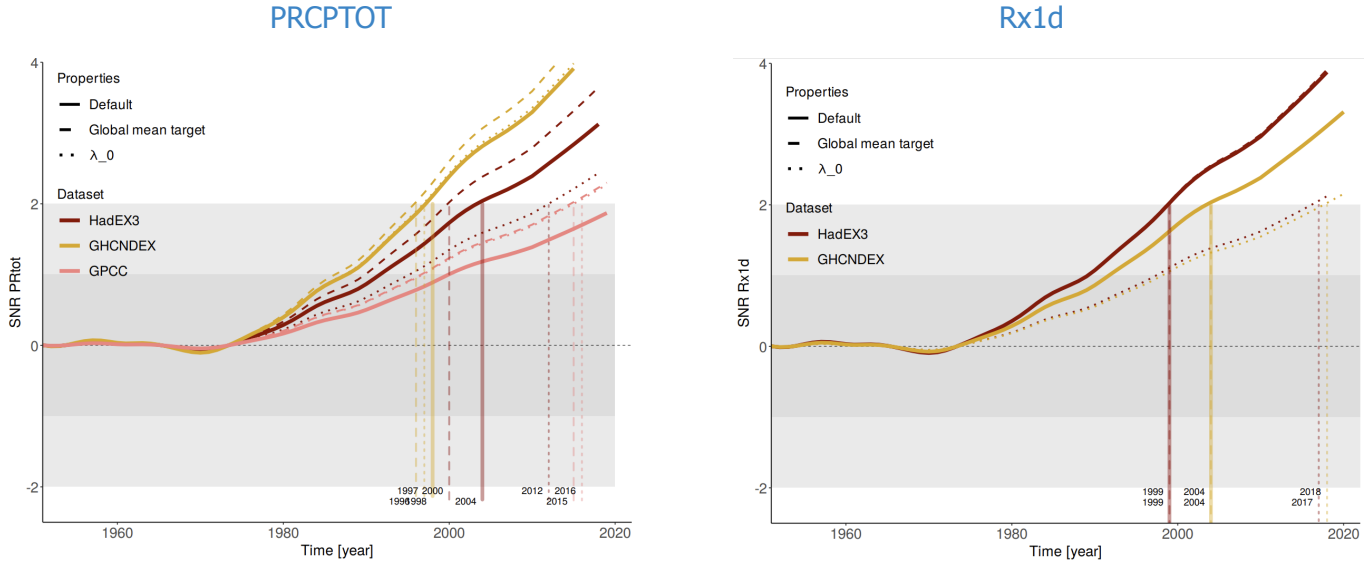


Figure 1: SNR of PRCPTOT (left) and Rx1d (right) forced response estimates from all observational datasets, regressed onto smoothed global mean surfact temperature (GMST) (as in manuscript), for cases 1, 2, and 3 as above.

Comparing cases 1 and 2: As can be seen in figure 1, the SNR does not necessarily increase by using the EOF based target instead of the global mean target; for PRCPTOT, 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 EOF of mean precipitation, and models are more in agreement on global mean precipitation change. However, since forced changes in mean precipitation behave according to a pattern of wetting and drying regions (e.g. Held & 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 first principal component is thus a reflection of the forced pattern strength (e.g. Marvel & Bonfils, 2013), meaning the forced response in all regions is somewhat reflected in this timeseries, 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. We argue that including the uncertainties in the forced response, reflected by uncertainties in the first principal component, has preference and may prevent overconfident detection of a signal. We argue this is a more balanced reflection of the forced response.

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. We therefore use the more conservative estimates, which implicitly include pattern information and uncertainties, by default. We point out, however, that the main conclusions, which are detection of a forced response but disagreement among observational datasets on the observed forced response relative to the simulated forced response, are insensitive to the choice of target metric.

Comparing cases 1 and 3: This comparison 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 used in the training procedure, and in all cases at least two orders of magnitude smaller than λ_sel. 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 is a trade-off between the increased variability and the increased trend. For Rx1d, we see that a smaller λ 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 smaller λ, 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 L283 of the manuscript) is somewhat dampened by larger λs. When λ is minimised, 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 this λ, however, physical consistency of the fingerprints is strongly impaired, as can be seen below, comparing λ_sel and λ_0.

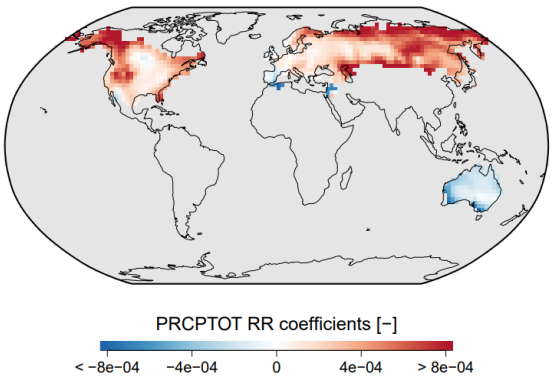


Figure 3a: GHCNDEX detection fingerprint for λ_sel

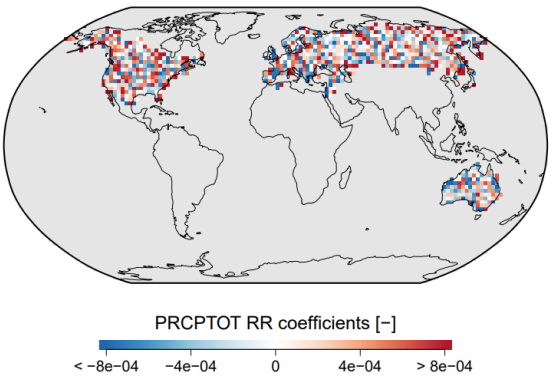


Figure 3b: GHCNDEX detection fingerprint for λ_0

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

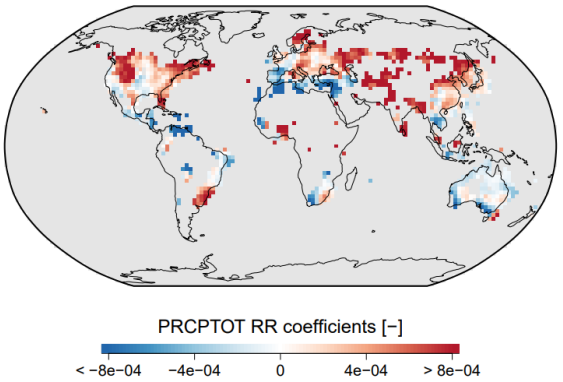


Figure 4a: GPCC detection fingerprint for λ_sel

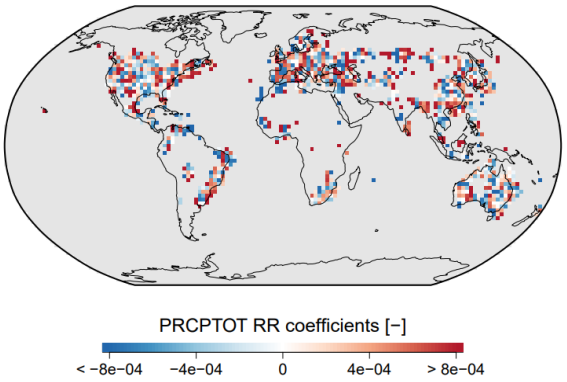


Figure 4b: GPCC detection fingerprint for λ_0

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 & 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.

2. An important motivation of considering different periods and datasets is opposing conclusions by previous studies about model overestimation or underestimation of the observed trends. I am wondering if the authors can go further and compare their results with some previous studies. For instance, if studies based on the latter half of 20th century trends find model underestimation, the authors can assess their model trends for the same/similar periods. Another point here is that the present study uses absolute units of precipitation while most of previous studies considered relative changes or aggregated values. It would be good to discuss possible influences of this difference.

Thanks for these very valid comments. We identify two main comments here - one being the comparison with previous studies and the other being the comparison between absolute versus relative units of precipitation. We address these two issues separately below, in reverse order.

Comparison of different precipitation metrics

To address this, we intend to add a section to the supplement with a concise description of the comparison presented below.

Some studies define precipitation change as a percentage change relative to climatological precipitation levels per degree of global temperature change. One can determine relative precipitation changes at the gridpoint level (normalised w.r.t. climatological gridpoint-mean precipitation) or at the global level (most common - underlying numbers such as ~2%/K and ~7%/K for PRCPTOT and Rx1d). Determining relative precipitation changes at the gridpoint level ensures that e.g. the tropics - a region with high absolute precipitation changes due to high climatological precipitation (Clausius-Clapeyron) - do not dominate the overall response. However, it could also lead to inflation of trends at grid points with very low climatological precipitation levels (e.g. desert areas into which precipitating bands shift, where local relative precipitation metrics approach infinity due to dividing by close-to-zero climatological levels), which is why we do not use gridpoint-level relative precipitation change.

Whereas we thus use absolute units in our predictors, our forced response metric (the model projections onto the first multi-model mean EOF) does not have meaningful physical units, but reflects a time series that includes pattern information, as mentioned above (it is a linear transformation of the raw data in original units). Implicitly, this already partially accounts for the regional differences in the expected absolute trends, since the pattern has higher loading in regions where precipitation is climatologically high. Note also that our forced response estimates do not have meaningful physical units in terms of mm/s, and reflect the strength of the forced response pattern, rather than the absolute change in precipitation in mm. Nonetheless, differences in overall precipitation level between different models and observations, which can affect the found strength of the forced pattern, are not accounted for.

Hence, to allow comparison with studies that use *global* relative precipitation change in % (which do not suffer from the approach toward division-by-zero that can occur at some gridpoints when *local* relative change is used), we have normalised our forced response estimate trends and model target trends with respect to their corresponding global mean precipitation levels (average over the gridpoints in the observational masks). We assess the model forced response *targets* (EOF-based) and the observational forced response *estimates*, since this allows assessing whether the answer to the question "do models over- or underestimate observed forced change?" depends on the unit of precipitation (absolute vs. normalised). Note that we normalise our forced response estimate trends, which are unitless. The resulting trend unit is thus mm^{-1}.

Figure 5 shows the results. Note that these plots represent three points (start years 1951, 1971, and 1991, from left to right) in Figure 2 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, 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 (e.g. 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.

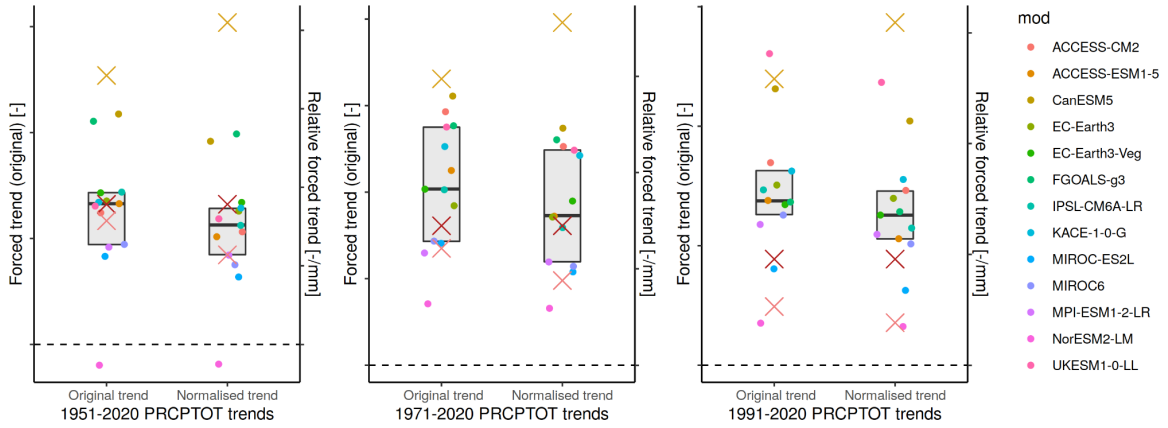


Figure 5a: Comparison of original PRCPTOT trends (as in manuscript) and trends normalised by the model's/observation's corresponding climatological PRCPTOT level; the 1951-2014 mean, averaged over the observational masks. Trends of single-model targets (points and corresponding boxplot indicating the interquartile range), and observed forced response estimates (X-marks). Non-physical units, black dashed line indicates 0.

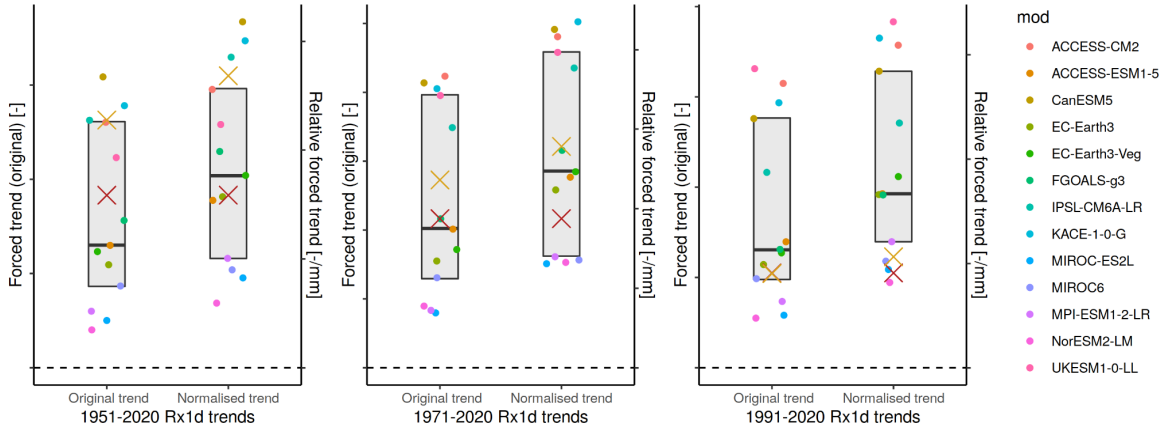


Figure 5b: As 5a but for Rx1d

For Rx1d (Figure 5b), 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 (e.g. Sillmann et al., 2013, Bador et al., 2020). Nonetheless, again, 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 & Knutti (2016), who suggest PRCPTOT changes are overestimated by models, whereas Rx1d changes are underestimated.

Some studies assess precipitation change as a function of global mean temperature change, e.g. in %/K. Given the relationship between temperature and specific humidity/saturation vapour pressure (Clausius-Clapeyron), this can in fact make a large difference, since different models, as well as observations, warm at different rates (different climate sensitivity). Although our forced response metrics, as said above, represent strength of forcing, we can still normalise the strength of forcing w.r.t. global mean warming to account for differences in climate sensitivity.

Therefore, we further normalise the relative trends shown above, by dividing by the temperature change over the trend period. This results in trends that are independent of the model's and observations' differences in climatological precipitation levels and warming rate (climate sensitivity).

Model targets are normalised w.r.t. their specific model's mean global mean surface air temperature (GSAT) change over the corresponding trend period, and observational forced responses are normalised w.r.t. the GMST change from the Cowtan & Way (2014) temperature dataset. We determine GMST change by simply computing the difference between the 2020 value and the values in 1951, 1971, and 1991 of the 21-year LOWESS-smoother GMST.

The comparison between original trends, as in the manuscript, and relative GMST-normalised trends (in $mm^{-1}K^{-1}$) is shown in the figures below for PRCPTOT (6a) and Rx1d (6b). Comparing the left and right column in each panel shows that normalising the forced relative trends from Figure 5 w.r.t. their corresponding temperature change reduces model spread, which is to be expected. For PRCPTOT (Figure 6a), 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 (Figure 6b), GMST-normalisation reduces model trends as well, and offsets some of the effect of normalising w.r.t climatological Rx1d levels seen in figure 5b.

However, more importantly, figure 5 and 6 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.

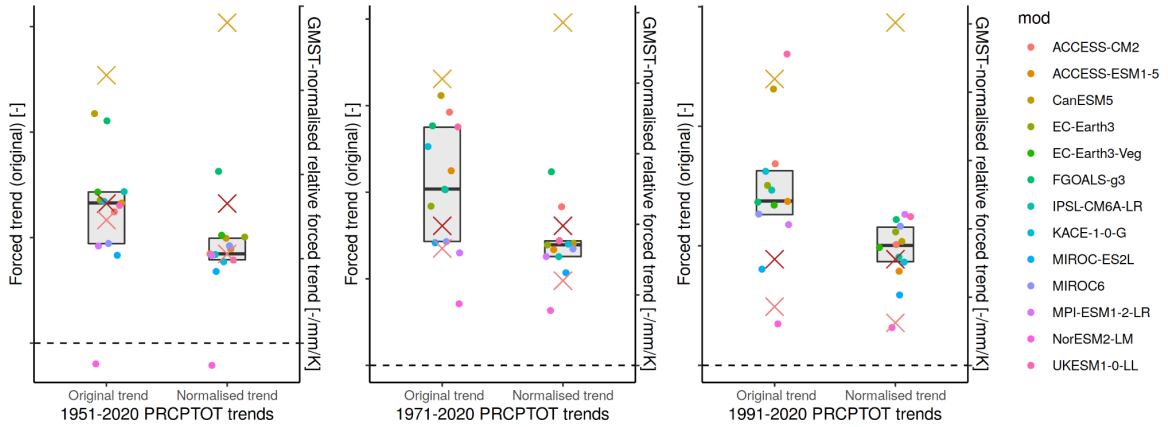


Figure 6a: Comparison of original PRCPTOT trends (as in manuscript) and trends normalised by the model's/observation's corresponding climatological PRCPTOT level; the 1951-2014 mean, averaged over the observational masks. Trends of single-model targets (points and corresponding boxplot indicating the interquartile range), and observed forced response estimates (X-marks). Non-physical units, black dashed line indicates 0.

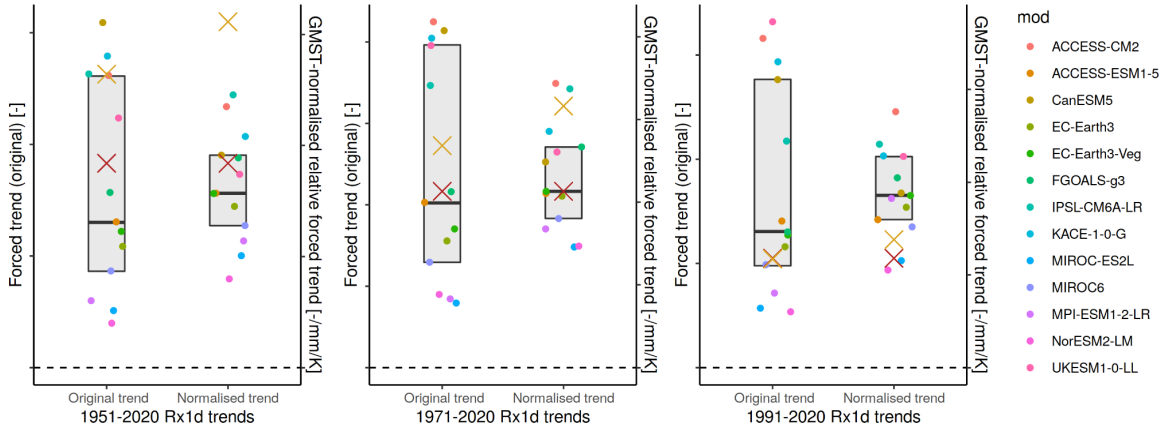


Figure 6b: As 6a but for Rx1d

The table below contains an overview of whether the model trend interquartile range lies higher than (+) lower than (-), or contains (0) each observational time series. As can be seen, the full range from under- to overestimation of observed trends by models is covered across trend periods and observational datasets, both for original as well as normalised trends.

		PRCPTOT		Rx1d	
Obs dataset	**Trend period**	**original [-]**	**norm [-/mm/K]**	**original [-]**	**norm [-/mm/K]**
HadEX3	1951-2020	0	-	0	0
	1971-2020	0	-	0	0
	1991-2020	+	0	0	+
GHCNDEX	1951-2020	-	-	-	-
	1971-2020	-	-	0	-
	1991-2020	-	-	0	+
GPCC	1951-2020	0	0		
	1971-2020	+	+		
	1991-2020	+	+		

In addition, to complete the comparison, we assessed trends in %/K (i.e. physical units), which are obtained by using the forced response estimates based on the ridge model with the *global mean* target (supplementary information section S2.2). The main conclusion still does not change qualitatively - 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 (e.g. Allan & Soden, 2008, Kotz et al. (preprint)). 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. (Note - these numbers only apply to global mean changes, not to local gridpoint changes.)

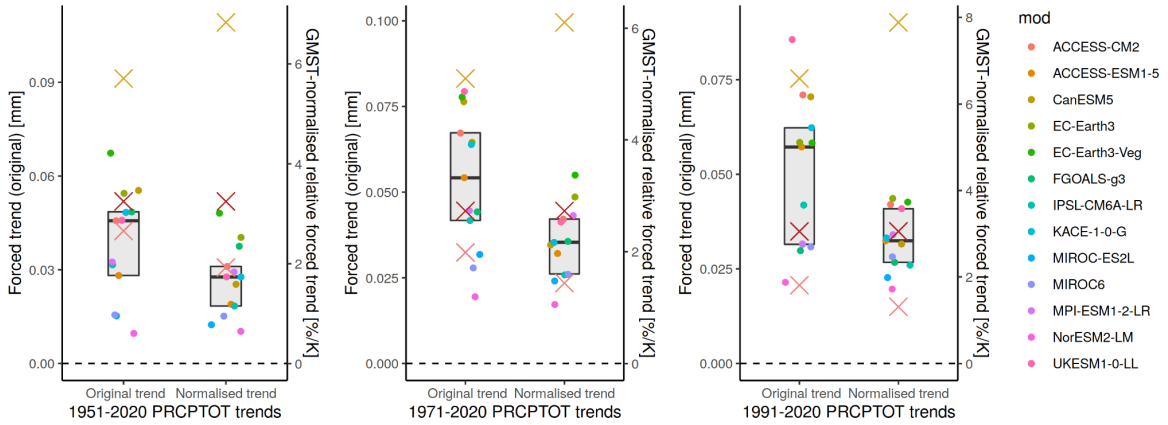


Figure 7a: Comparison of original PRCPTOT trends (as in manuscript) and trends normalised by the model's/observation's corresponding baseline (1951-2014) PRCPTOT level and global mean surface temperature (GMST) change for three different periods (1951-2020, 1971-2020, 1991-2020). Target trends (points/boxplot) and forced response estimates are based on the global mean in this case, leading to physical units.

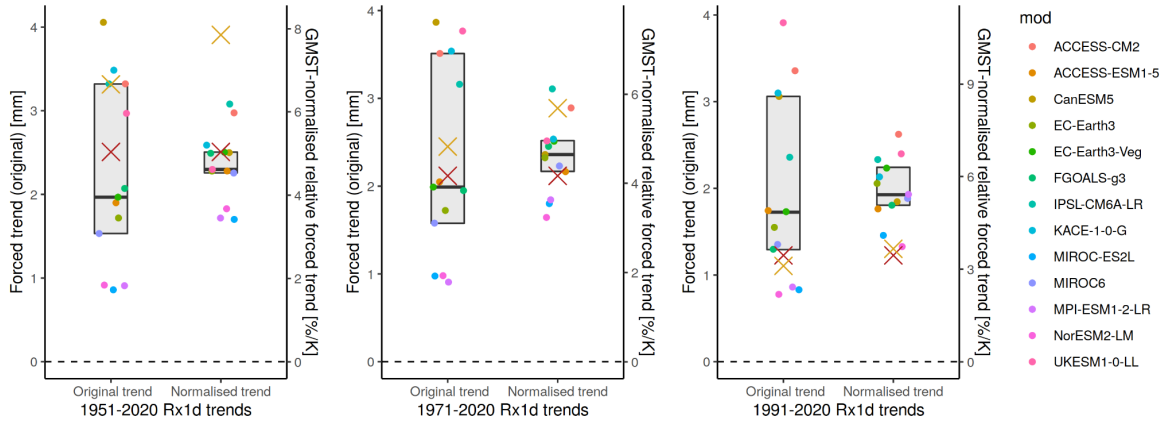


Figure 7b: As 7a but for Rx1d.

Comparison to previous studies

We suggest addressing this by adding a few relevant references to previous studies in the discussion of the results, as well as by adding a table to the supplementary information. This table contains an overview of a set of recent papers that have attempted detection of either PCRPTOT or extreme indices (Rx1d or e.g. Rx5d) and that have made assertions on model vs. observational trends. This provides an overview of the (dis)agreement regarding over- or underestimation of observed forced changes in precipitation by models. See a preliminary version of the table below.

From this table, several interesting comparisons result. First of all, scaling factors obtained through optimal fingerprinting for PRCPTOT often lead to the result that models underestimate observed change. Assessing the spatial distribution of trends, however, shows that models in fact produce positive PRCPTOT trends over a larger land fraction than observations do. In our study, models also underestimate PRCPTOT trends in GHCNDEX, but agree better with, or overestimate, HadEX3. HadEX3 has considerably higher coverage and resolution than GHCNDEX, and also than HadEX2. Also GPCC estimates much lower global forced response trends than model projections, whereas Knutson & Zeng (2018) find higher local trends in GPCC compared to CMIP5 models. Model underestimation of local internal variability in mean precipitation may partly cause this. This also aligns well with the findings of Fischer & Knutti (2014), since they assess trends in units of local standard deviation. The higher PRCPTOT trends in models may be an artefact of underestimation of local PRCPTOT variability (standard deviation) in models.

For Rx1d, optimal fingerprinting studies often use a probability index (PI), meaning effectively that trends in percentiles are assessed (an increasing prevalence of Rx1d values that lie further to the right on the local GEV-distribution of Rx1d values). An interesting finding is that this approach leads to the conclusion that models overestimate Rx1d changes based on scaling factors, whereas normalising changes by warming rate and computing trends, leads to the conclusion that models underestimate trends in Rx1d with warming (Paik et al, 2020). We showed above that our method does not lead to fundamentally different results depending on the metric used (non-normalised changes or relative changes as a function of warming). Primarily, we can conclude that the observational dataset seems to have a large influence on the results. Where e.g. GPCC did not allow detection in any of the assessed cases in Polson et al. (2013), GHCNDEX often seems to suggest models underestimate observations. HadEX3, with highest coverage and resolution, lies in between these two extremes in our study. Overall, these comparisons suggest that observational uncertainty is still large and may be highly relevant as to whether models over- or underestimate precipitation trends. This is consistent with Bador et al. (2020), who find that observational uncertainty can be partly as large as uncertainty across climate models.

Paper	Model archive	Obs dataset	Spatial region	Variable	Method	Trend periods	Models w.r.t. 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 values of daily precipitation	Probability distribution comparison	1951-1980 and 1981-2013 distributions	-	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. Scaling factors	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")

Table 1: 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

3. The lower detectability in GHCNDEX observations are suggested to be due to the poorer spatial coverage. Regarding this issue, I would suggest using Rx5d. As I understand, Rx5d has larger spatial coverage than Rx1d and comparison with Rx1d-based results may provide a way to support the authors' interpretation. Another way would be to compare detection results from using a selected model run but with different spatial coverages applied.

We assessed the effect of coverage by masking HadEX3 as GHCNDEX and running the ridge regression and detection procedure. This resulted in higher consistency between HadEX3 and GHCNDEX, but did not reconcile the differences fully. Therefore, coverage only explains part of the differences. We will make this more clear in the text (L318-320).

Also, our primary motivation for using PRCPTOT and Rx1d is to assess mean and extreme precipitation separately; Rx5d would not accomplish this goal as well as Rx1d because it is less extreme and thus more similar to PRCPTOT (Pendergrass & Knutti, 2018). Furthermore, to our knowledge, Rx5d does not have higher coverage in GHCNDEX than does Rx1d, see below: all the white cells have no coverage. The difference between 1951-2020 coverage of GHCNDEX for Rx5d versus Rx1d is shown in figure 7c - the dark red cells have coverage for Rx5d but not in Rx1d, yellow cells have coverage for both. Given that the coverage increase is minor and only in areas where there is reasonable coverage already, we anticipate that this would not make much of a difference.

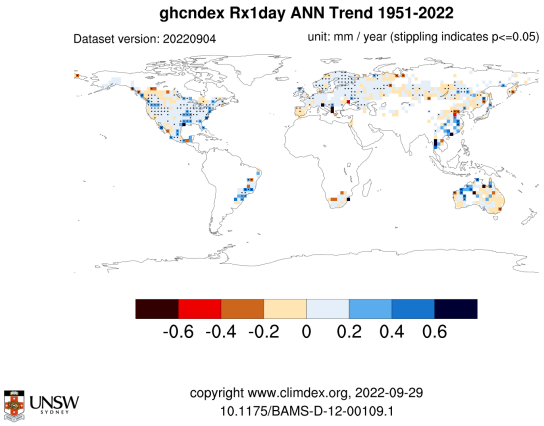


Figure 8a: Rx1d trend from climdex.org, white cells have no coverage

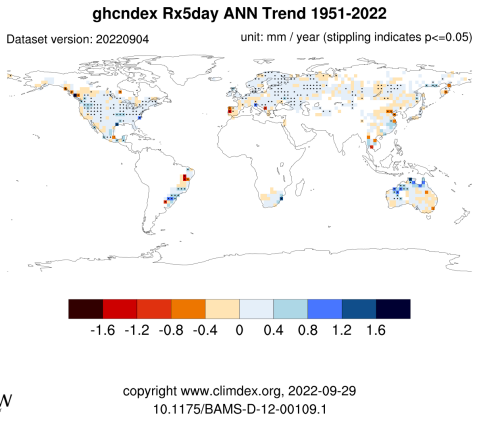


Figure 8b: Rx5d trend from climdex.org, white cells have no coverage

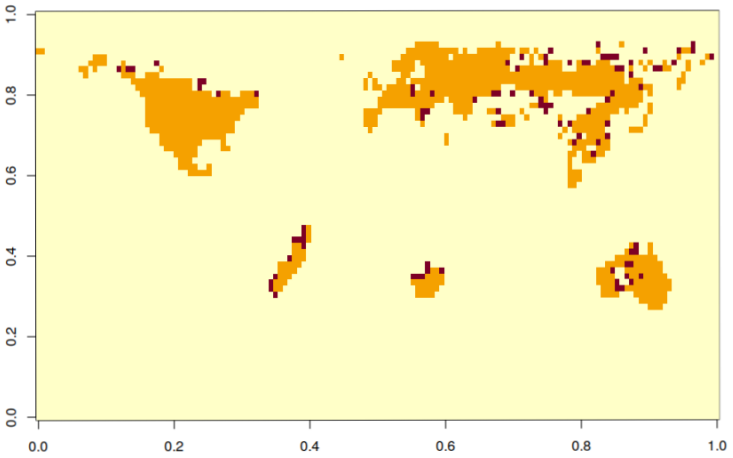


Figure 8c: Coverage differences GHCNDEX Rx5d and Rx1d: red cells are added in Rx5d w.r.t Rx1d.

Minor comments:

L8: Indicating analysis period or trend period with signal detection would be useful here.

In a revised manuscript we will changes this to "[...] to assess the degree of forced change detectable in the real-world climate in the period 1951-2020."

L17-19, L58-64: Better comparisons can be made by applying the same periods as those used in previous studies. See my major comment above.

See reply to major comment above: both previous studies as well as we assess multiple trend periods. Disagreements across studies and observational datasets remain.

L20-21: Is this confirmed by repeating detection analysis using NH-extratropics only?

Yes, see supplementary information.

L34: "discrepancies with respect to observations". Its meaning is unclear.

We hope this will be resolved by changing the sentence to "[...] model representation of the water cycle has also been shown to disagree with observations."

L69-71: Need to explain what the previous studies have found additionally using these "data-science methods". Also, what's the novelty of this study compared with them? Is it detection based on spatial pattern information alone?

As the sentence reads: these studies have detected forced signals. Since the main purpose of the D&A field is to answer the question "can we detect and attribute effects of forcing in observations", this is the finding that matters. Our method fits in these recent data science developments in D&A that move towards mapping multidimensional data onto a one-dimensional detection space. Studies based on neural networks and deep learning for detection and attribution, employ non-linear methods - as opposed to our linear ridge regression method - but use a very similar framework with similar goals. We do not argue that ridge regression is fundamentally better than any of the older or newer methods, but we are convinced that the intuitive, physical outputs combined with high SNR can be valuable for trend detection and attribution. (See also response to referee #1.)

L108-109: "Trend biases due to this structural difference … negligible". But the cited reference considered south-east Australia only?

This is true, however, the study investigates the effects of data operations on time series. The temperature and precipitation time series of course differ per region, but the effects of operations on the long term trends is not expected to differ greatly from region to region. Nonetheless, we will add the reference to Dunn et al. (2020) (also referenced in L107), who also makes this statement more generally.

L201: How to define S when global means are removed?

The definition of S stays the same. To obtain the results for the detrended case, global means are removed from the predictor data in training of the ridge model. The ridge model is however still trained to predict the same forced response target. Therefore, the forced response estimates, based on the detrended observations, still contain the forced trend (if the method works). These forced response estimates are regressed onto GMST to obtain S in the same way as for the default case. We hope this clarifies things.

L212: "CMIP6 ssp245" should be "CMIP6 historical"?

Yes, thanks for noticing this.

L227: "virtually identical". adding spatial correlation would help with this.

That is a good point, we will add the correlation value in a revised manuscript. Pearson correlations between linear trends and EOFs over the full historical-ssp245 period for both PRCPTOT and Rx1d, on both the coarser GHCNDEX and the finer HadEX3 grid, are > 0.99.

L314-316: This suggests possible dependence of Rx1d FRE on temperature, resembling global warming slowdown due to PDO influence?

Potentially, although we do not have enough evidence to claim that the levelling off of the trends is not simply due to shorter trend length and internal variability. Attributing changes in trend slope to lange scale modes of variability is outside the scope of this study.

L331-332: "results … hold when the global mean is used as FR target". Then what are benefits of using EOF-based metric for target variable?

See above discussion of using the EOF-based target metric. We will add a sentence or two to the method section to further justify this choice of target.

L382-383: "accuracy of the CMIP6 climate models in simulating the processes …". It's unclear how the authors get this conclusion. Observation-model agreement in residual variability? More explanation would be useful.

Precisely. As we mention in the manuscript, removing the mean trend from the predictor data implies that only the relative pattern of precipitation can be used by the ridge regression model to predict the forced trend (note: the area-mean trend is removed from every grid point time series, meaning that spatial pattern information (how local precipitation differs from the area-mean) is retained in the predictors). The ridge regression detection model is trained to predict the (model) forced response from simulated spatial precipitation patterns, meaning it finds the *simulated* relationship between spatial precipitation patterns and forced precipitation change. Applying the detection model to observations, results in detection of a forced trend in observations, which implies that the relationship between spatial precipitation patterns and the forced response that the ridge model learnt from models, also holds in observations. This thus implies accuracy of the models in representing the spatial patterns related to the forced response. We hope changing the last sentence and splitting it into two as below solves the confusion:

" Taken together, the above shows, first, detection of forced change in mean and extreme precipitation beyond a global mean trend, second, the power of RR for signal extraction from high-dimensional noisy data, and third, the accuracy of the CMIP6 climate models in simulating the processes relevant to the spatial pattern of forced change in mean and extreme precipitation." → " Taken together, the above shows, first, detection of forced change in mean and extreme precipitation beyond a global mean trend, and second, the power of RR for signal extraction from high-dimensional noisy data. Also, the fact that the relationship between relative spatial precipitation patterns and the forced precipitation trend learnt from climate model simulations by the ridge model holds in observations, suggests accuracy of the CMIP6 climate models in simulating the processes relevant to the spatial pattern of forced change in mean and extreme precipitation."

L394-395: "(not shown)". This looks important and I suggest showing them in the supplement.

We will add these plots to the supplementary information.

L428: "value of RR-based fingerprint construction". What happens in detection or SNR without applying RR? See my major comment above.

See reply to major comment above. Overfitting leads to high variability in forced response estimates, and low SNRs.

References

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.

Marvel, K. and Bonfils, C.: Identifying external influences on global precipitation, Proceedings of the National Academy of Sciences, 110, 19 301–19 306, https://doi.org/10.1073/pnas.1314382110, 2013

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.

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, *J. Geophys. Res.*, 115, D24211, doi:10.1029/2010JD014532, 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, *J. Geophys. Res. Atmos.*, 118, 1716– 1733, doi:10.1002/jgrd.50203, 2013.

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.

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.

Allan, R. P., & Soden, B. J.: Atmospheric warming and the amplification of precipitation extremes. *Science*, *321*(5895), 1481-1484, 2008.

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, https://doi.org/10.31223/X5C631, 2022 (preprint).

Bador, M., Boé, J., Terray, L., Alexander, L. V., Baker, A., Bellucci, A., et al. Impact of higher spatial atmospheric resolution on precipitation extremes over land in global climate models. *Journal of Geophysical Research: Atmospheres*, 125, e2019JD032184. https://doi.org/10.1029/2019JD032184, 2020.

Pendergrass, A. G., & Knutti, R.: The uneven nature of daily precipitation and its change. *Geophysical Research Letters*, 45, 11,980– 11,988. https://doi.org/10.1029/2018GL080298, 2018.

Dunn, R. J. H., Alexander, L. V., Donat, M. G., Zhang, X., Bador, M., et al.: Development of an 520 Updated Global Land In Situ-Based Data Set of Temperature and Precipitation Extremes: HadEX3, Journal of Geophysical Research: Atmospheres, 125, https://doi.org/10.1029/2019JD032263, 2020.

References in table

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.

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.

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.

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

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.

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

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.

Paik, S., Min, S.-K., Zhang, X., Donat, M. G., King, A. D., and Sun, Q.: Determining the Anthropogenic Greenhouse Gas 575 Contribution to the Observed Intensification of Extreme Precipitation, Geophysical Research Letters, 47, e2019GL086 875, https://doi.org/10.1029/2019GL086875, 2020.

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.