Reviewer comments
Replies

First, I would like to thank the authors for their detailed responses to my comments and for the improvements that have been made to the paper, which in my view, is much improved, particularly regarding the presentation of key concepts that are needed to understand the paper and how it is situated in the “D&A” literature.

I do have some additional specific comments that might help to further improve an already very good paper, which I hope you find useful.

Thank you very much for this positive evaluation of our revised manuscript, and we greatly appreciate the effort and conscientiousness you put into reviewing this version so thoroughly once again. Below, we provide point-by-point answers to your additional comments.

60-77: How does this paper help to resolve the issues that are raised here concerning the robustness of findings from D&A studies and the representation of precipitation related processes in models? The text that begins with “Here we show …” (line 70) doesn’t really answer that question.

We updated the text in L70-76 to answer this question more explicitly.

116-119: It seems that you’ve tried to create a very “balanced” climate model dataset (3 runs per model, 450 years of PI-control per model for as many models as possible). A few words on the sampling approach (e.g., what motivates it, what problems are avoided through a balanced approach, and what other problems might arise – see comments below) might be appropriate.

That is true, we added some words on the motivation for the model data selection in L119-126.

134-135: I think the crucial issue is not so much the locations that grid-cells represent, but rather, the quantities that they represent. The question of what quantity they represent is easier to understand in the case of Rx1day. I think we can all agree that the Rx1day values obtained from models can only be interpreted as annual maxima of daily spatial mean (grid box average) precipitation amounts. In contrast, when a product like HadEX3 is produced from station data, Rx1day is first determined at stations and then those point Rx1day values are spatially interpolated to obtain a grid box value. This is a different number from the annual maximum of daily grid box averages of 1-day point precipitation amounts, which is how one would interpret model output. That is, there is an order-of-operations difference that could potentially influence model-observations comparisons. If we had dense observational coverage (e.g., say at least 50 rain gauges in each model grid square), we could presumably first calculate the grid-box average precipitation amount, and then calculate Rx1day from those averages, bringing us conceptually closer to the thing that the models produce.

Indeed, we agree that we did not state this clearly enough, and made this more explicit now. Just to add: it is an inconvenient fact that models output grid cell mean quantities, whereas

observational Rx1d values are not determined from gridded station data (and even if they were, it would still not be the same since gridded observational PRCPTOT data does not reflect the grid cell mean), but from raw station data. We did our best to not make this discrepancy worse by ordering our preprocessing steps on model data in the way that best reflects the observational order of operations. Namely, we first 'extract' the Rx1d values at each individual gridpoint (~equivalent to determining station maxima), and then regrid onto the coarser observational grids (~equivalent to regridding). Fact of the matter is, however, that the maximum of a "small" grid cell on the native model grid still differs greatly from individual station maxima.
We've rewritten this paragraph, L141-151, and also added a sentence or two to section 2.2 to underline the importance of order of operations again, L154-156.

163-165: The paper seems to be getting ahead of itself a bit here. The results show this to be true for the precipitation variables you consider, but at this stage, it isn't known whether we should expect the same for precipitation as for surface air temperature.

That's a valid remark. Given that the linear trendmaps in Fig. 2a/b have high spatial correlation with the first EOF pattern (visible in Fig. 2 and Supp Fig. S2), and global mean timeseries with the first PC (visible in Supp Fig. S3) we can deduce that external forcing is represented in the first EOF. We shuffled the text around a bit to make this point in the correct logical order, see L179-189.

174: It seems clear from Fig. S2 that you have two groups of models with a substantial gap in the range of sensitivities that isn't sampled. It would be worth including some discussion of how that heterogeneity could potentially have affected results.

This gap is remarkable indeed (for Rx1d) and is related to the different warming rates of the models (clausius-clapeyron). Although there is an unsampled space between the two clusters, the ridge model would still end up somewhere in the middle (which lies in the non-sampled space) to minimise the bias in both directions. Supp. Fig. S4 shows and main text L246 states that the RR models have low sensitivity to any left-out model (pre- versus post-crossvalidated performance is nearly identical). This indicates that the predictors of all-but-1 models can train an RR model that generalises well to the left-out model..
Furthermore, we show in supp sect. S2.3 and state in the main text that the results are not fundamentally affected by differing climate sensitivities, since the results do not change if we normalise each model's forced response by its temperature change.
This leads us to think that the heterogeneity does not affect the results much, given that there is sufficient data on both sides of the gap. We think our referral to the pre- and post crossvalidation results, and the results where the dependence on climate sensitivity is removed by normalising wrt temperature change, provide enough proof of robustness, and we do not add additional discussion on this.

200: Is the cost function minimized separately at each location? In minimizing the cost function, how do you account for the impacts of spatial and temporal dependence?

The cost function is minimised globally across the masked coverage), see eq. 4 in the paper.

There is no location-dimension in the two variables (y and Xbeta) in the cost function which govern the magnitude of the error. However, we fully agree with the reviewer that there is (strong) spatio-temporal depedencies between the predictor variables in the regression formulation. Because of that, and the large amount of predictors and data points, there would be high (perhaps inevitable) risk of overfitting in a standard OLS multiple linear regression framework. Preventing this – overfitting due to spatial dependence and large data matrices – is exactly the goal of regularisation key to ridge regression (see also more detailed description in L232-236). Because of the spatial dependence in the data, predictors are not orthogonal and OLS regression would lead to non physical and high coefficients to, in a way, "force" an approximation of the target. This is overcome by the L2-norm regularisation used in ridge regression, since this forces coefficients to become more homogeneous: due to the squared-sum penalty, high coefficients are penalised most strongly, effectively resulting in moderate coefficients, which, when applied to geophysical data with spatial dependence, leads to spatially coherent maps of coefficients. Hence, regularisation accounts for spatial correlation in the data.
Temporal dependence in the data is induced through long-term trends in the forcing, or decadal variability in the climate system. In the way we applied ridge regression, autocorrelation in the data does not affect overfitting in the same way as spatial correlation does, since every year makes up a separate equation in the linear system. We use a careful cross-validation scheme, in which we strictly separate the fitting of the model (on a set of *k* climate models) from the application to observations, or other, unseen climate models. This has the effect that the algorithm has no knowledge of the temporal variability in the application model or in observation. This ensures that only the temporal structures where the models agree (i.e. those that are not due to internal variability or model biases) are reflected in the coefficient fingerprint.
We mention the role of regularisation in suppression spatial artefactss in L232-236, and the temporal aspects are addressed in L223-228.

269-271: Since the fingerprints are model based, they could presumably be shown for the global domain, with outlines of the regions with observational coverage. But perhaps this suggestion is naïve (see my previous comment).
Even though the fingerprints are model based indeed, and the model has full coverage, we still do not (cannot) determine the fingerprints for the full coverage since our ultimate goal is to apply the fingerprints to observations. We want to predict/estimate the forced response from observations, which means that we need an RR model that maps the *observed* gridcells to a forced response estimate. The RR-model is thus trained only with observed grid cells.
If we would determine the weights based on global coverage (i.e. a coefficient for a predictor at every grid cell), and only use a part of the coefficients when we apply the RR model to observations (thus leaving out a large fraction of the predictors), the result would not predict the forced response anymore. Hence, the masked fingerprints we show, show the full RR model.

If the cost function is minimized "globally" across the observation grid, then a question that emerges would be whether the geometry of the observation grid affects the fingerprints that are obtained, and what implications that might have for detection (or the timing of detection).

As the cost function is minimised across the observational grid, the weights depend on the coverage and details of the grid: unobserved regions can not contribute to the forced response estimate. We refer to the consequences of this several times in the paper, pointing out the lower coverage of GHCNDEX, and the dominance of the Northern hemisphere due to observational coverage (and land mass) being much larger there, e.g. in L517-521.

308-309: I think this is a consequence of having a collection of models that do not sample the model sensitivity spectrum very well. (See also my comment concerning line 174).

The differences in climate sensitivity, and the unsampled climate sensitivities you pointed out above, may contribute to the fact that the ridge model does not perfectly predict the modelled forced response from model data. However, the main reason for the phenomenon these lines refer to – the flattening of forced response estimates relative to the targets – is the use of regularisation. We show this with the plots below:

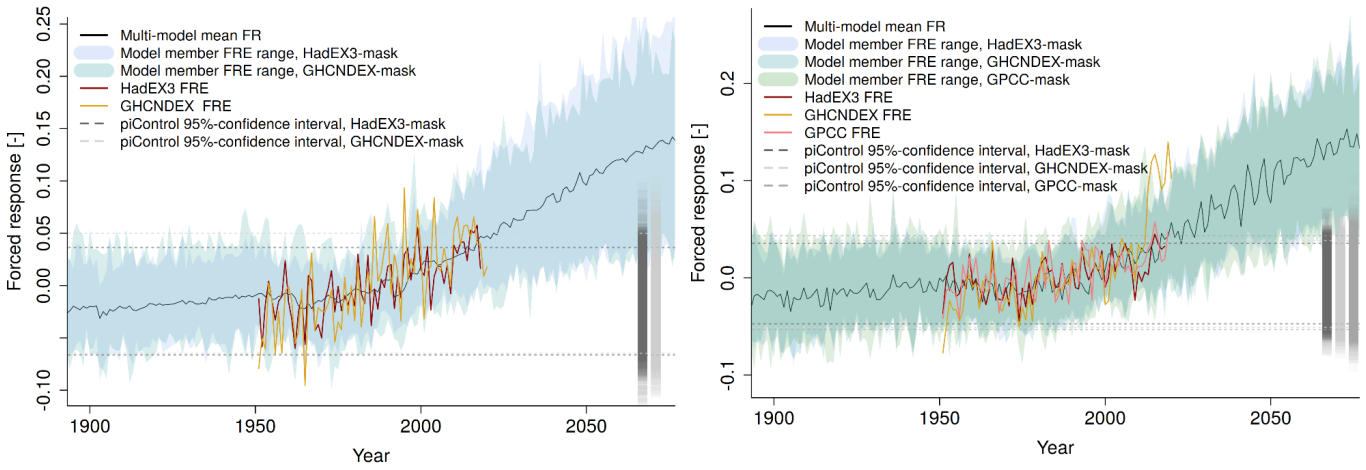

Figure 1: Forced response estimates obtained by applying the RR model **with regularisation parameter lambda_min** to model data (shading) and observations (colours), PRCPTOT (left) and Rx1d (right) – same as Fig. 3c and d in the main paper, apart from regularisation parameter choice.

These plots show the forced response estimates made by a minimally regularised RR model (lambda_min, as described in the supplementary information). Effectively, a very small regularisation parameter means that the overfitting is not much reduced, and therefore the variance in the estimates is still large, and there is little bias. We see that the shading in these plots is centred on the black line, and does not show the “flattening” effect we described in L308-309 (of the previous manuscript, referred to by the reviewer).
The more we regularise, the more we dampen variance, but this comes at the cost of flattening the trend (which is also a source of variance): in the limit, infinite regularisation leads to a ridge model with spatial coefficients of zero, and just an intercept, predicting the mean of the timeseries, as shown below:

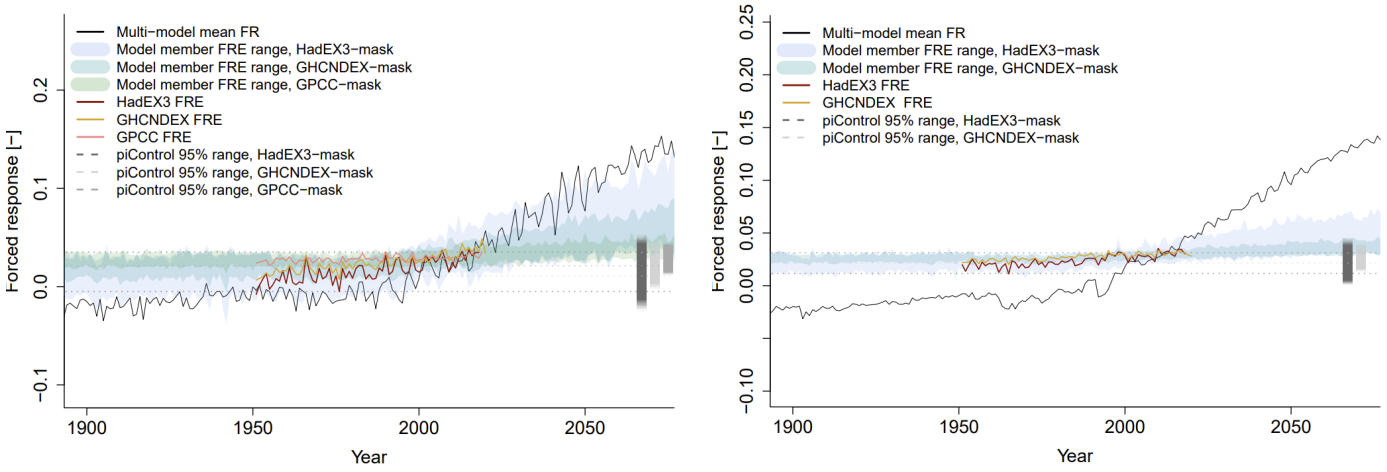

Figure 2: Forced response estimates obtained by applying the RR model **with regularisation parameter very large** to model data (shading) and observations (colours), PRCPTOT (left) and Rx1d (right) – same as Fig. 3c and d in the main paper, apart from regularisation parameter choice.

We see a mild signature of this bias-variance trade-off regularisation effect if we choose the lambda that is optimal for generalisability and interpretability (lambda_sel), which is what we described in L308-309.

360: Confidence in the consistency of results perhaps? I'm not sure what it means for a method to be consistent.

This was sloppily formulated indeed. What we mean is that the method produces consistent results in different contexts and is therefore robust. So we have rephrased this to "This consistency of results increases confidence in the robustness of the method, …", L382-383.

387: I'm not sure I understand what it means for change to be in phase, but of opposite sign.

It means that the correlation is -1 → the temporal evolution is the same, but where the forced response grows in the positive direction over time, the changes in these regions become increasingly strongly negative. We added a clarification in L410.

393: Replace "… coefficients flip sign." with "… coefficients that flip sign."
Done.

495-496: There are certainly papers that use precipitation datasets with greater coverage, but it takes a lot of work and personal contacts to collect that data and organize it for use in a D&A study. But even with that kind of work, coverage over land is still not very good. I wonder if we will get to the stage where we might have enough trust in reanalyzed precipitation to use it in a D&A study?

As you say, station data has certain "irreducible" caveats, such as being point-measurements, sparsely distributed, and sometimes even subject to (geo)political considerations. Also reanalysis data has its caveats, as you hint at. Nonetheless, reanalysis precipitation data might be fit for lower-resolution, longer timescale purposes, for example,

Bonfils et al. (2020) successfully use reanalysis data in a pattern correlation D&A study (assessing temperature, precipitation, and aridity jointly, i.e. not solely precipitation based). In the same way, station data also can be fit for different purposes, despite its imperfections. There is value in different lines of evidence and different types of detection and attribution studies and statements. Besides, we can learn many things from detection efforts that are not set out to detect the strongest signal, but that aim to combine physical understanding of the manifestation of the signal with the signal strength. The question whether forced climate change is detectable in sparse station observations is not solely aimed at making the best detection statement possible, possibly reanalysis would yield much more constrained results. The question is whether this type of very direct, “true” observations can be used in a physically explainable manner to show the effects of external forcing.
Anyway, to prevent getting lost in (interesting!) philosophical contemplations, a short answer to the question: we can probably learn things about climate change using reanalysed precipitation if we ask the right questions for the possibilities reanalysis data provides.

References

Bonfils et al. (2020). https://doi.org/10.1038/s41558-020-0821-1