Reply to Peter Caldwell (Referee)

Reviewer comments are given in **bold**, our answers in red.

This study confronts 11 emergent constraints on ECS with CMIP6 data for the first time. The skill of most of these constraints collapses when faced with new data. All constraints predict higher ECS based on CMIP6 data relative to CMIP5 because manyCMIP6 models have higher ECS yet similar constraint values. Overall I thought this paper was excellent, well-written, and very worthy of publication. I think the statistical significance methodology is inappropriate, however, which will require substantial revision. Otherwise I have a somewhat large number of fairly minor comments.

We thank the reviewer for the helpful and constructive comments. We have now revised our manuscript in light of these and the other reviewer's comments we have received. A pointwise reply is given below.

Major comments (in order of importance)

1. I'm uncomfortable with your bootstrap statistical significance testing method in sect 2.3.

1a. Your definition of statistical significance as "the sensitivity of the regression model to changes in the input data, i.e. the removal or addition of datasets" seems overly narrow. I think of statistical significance in this case as the probability of obtaining a correlation magnitude of at least r under the null hypothesis that no real correlation exists. By using such a narrow definition, I think you've avoided thinking about whether your methodology fully captures all sources of uncertainty. Your test is also weird in that it lacks any sense of a null hypothesis that there is no real correlation. Instead, you seem to just be taking the correlation obtained with all models and creating a histogram of possible values for it by recomputing correlations with models added or removed. I don't think this is appropriate.

1b. I suspect your bootstrapping results are strongly dependent on arbitrary sampling design choices: removing a model or adding multiple copies of a model makes a huge difference to your regression when you only have ~30 models in the CMIP archive with data for a particular constraint. Thus I expect the number of random samples you draw to have a big impact on your histogram of bootstrapped correlations. To disprove my complaint, you could create histograms with M-2, M-1, M, M+1, and M+2 models. If these all look the same, then my criticism is misplaced.

1c. It seems more defensible - or at least complimentary - to use a T-test as described in https://atmos.uw.edu/~dennis/552_Notes_3.pdf. If you did use the T-test, would you get similar results?

Following this comment and the review comment by Thorsten Mauritsen (2nd referee) we decided to remove the bootstrap significance testing from the paper. We agree that our original definition and implementation of statistical significance was not optimal and therefore replaced it with the *t*-test on the correlation coefficient as proposed. The null hypothesis is that no correlation exists between the predictor and ECS. In the new revised version of the manuscript, we now give *p*-values of the emergent relationships that correspond to the probability that the absolute correlation is larger than $|r|$ even though the null hypothesis is true, i.e. the true underlying correlation is zero. In addition, we do not use the *p*-values anymore to specify "absolute" significance (our categories "highly significant", "barely significant", etc. were arguably arbitrary), but only use these *p*-values to specify "relative" significance, i.e. to indicate whether the statistical significance changes when moving from CMIP5 to CMIP6. Consistent with our original bootstrapping approach, the *t*-test also shows that except for the ZHA and SHS constraints, all emergent relationships show a higher significance for the CMIP5 ensemble than for the CMIP6 ensemble.

2. You don't mention any of the limitations of your methodology until the conclusions section. This left me reading through the methodology and results sections under the impression that you were unaware of the possible flaws with what you were doing. It would be helpful for readers if you describe potential problems with the methodology in the methodology section so readers won't traverse the paper thinking you don't know what you're doing and so they can interpret your results with an appropriate level of skepticism.

We moved the discussion of the limitations of our methodology to the methods sections. In the conclusions sections we now only refer briefly to the limitations that are discussed in detail in the methods section:

"Our analysis makes a number of simplifying assumptions common to other studies, such as model independence, discussed in sections 2.1 and 2.2. These assumptions affect the significance of emergent relationships and the PDFs of ECS based on a constraint. However, they do not affect our main conclusions here, which concern the change in performance on CMIP6 relative to CMIP5 and the implications for robustness and future use of emergent constraints."

2a. In addition to the problems with your methodology you currently mention in the conclusions, it is probably worth also mentioning that giving the models which agree worst with the observed constraint value equal weight in determining the regression is probably a bad idea. This issue is explained nicely in Brient (2020; https://link.springer.com/article/10.1007%2Fs00376-019-9140-8).

We added the following discussion of this issue and the reference to the limitations paragraph at the end of the methods section:

"Moreover, our approach assigns equal model weights without taking model performance into account, i.e. agreement with observations. This issue is discussed in detail by Brient (2020)."

3. Caldwell et al (2018; https://journals.ametsoc.org/jcli/article/31/10/3921/94898/Evaluating-Emergent-Constraints-on-Equilibrium) tested 5 constraints trained on earlier CMIP ensembles on CMIP5 data and found that 4 of these constraints (Covey, Trenberth, and Fasullo D and M) also failed when confronted with out-of-sample data. In that context, I see your paper as a follow-up to the Caldwell paper. I think it would be worth mentioning this around L60. It is interesting that Volodin was trained on CMIP3 data but also holds up for CMIP5 and CMIP6 data.

We added your suggestion to the introduction:

"In addition, Caldwell et al. (2018) performed out-of-sample tests on five emergent constraints originally trained on older CMIP versions, by applying them to the CMIP5 ensemble. They found that out only one of the five passed this test. In this paper, we follow up on the work of Caldwell et al. (2018) by analyzing 11 published emergent constraints on ECS […]".

4. On a related note, I felt you undersold the importance of Zhai failing when confronted with CMIP5 data which it wasn't originally trained on. A similar thing happened in Caldwell et al 2018 with the Qu constraint. Such sensitivity to sampling details seems to me an important indicator that the number of models we have in the CMIP archive is insufficient for making robust conclusions about the credibility - or lack of credibility - of the constraints we propose.

We added a short paragraph to section 5 (summary) that picks up on the failing ZHA constraint by referencing the corresponding Figure 6 and discussing that this result suggests that the credibility of the all other analyzed emergent constraints might be impaired:

"Moreover, our more detailed analysis of the ZHA constraint (see Figure 6) showed that this emergent constraint is very sensitive to outliers and the subset of the climate model ensemble used to fit the emergent relationship. Such a behavior might not be unique to the ZHA constraint but could apply to other emergent constraints as well. This in turn suggests that the number of climate models commonly used for emergent constraints might be too low leading to non-robust relationships."

5. Your introduction argument that ECS hasn't changed in 40 yrs feels dated in light of Sherwood et al (2020; https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019RG000678). I know you didn't mention this study because it wasn't accepted when you submitted the paper, but should be cited in the revision.

We added the following sentence to the introduction:

"A new assessment using this evidence has narrowed the 66% range (17–83%) to 2.6–3.9 K (Sherwood et al., 2020), but in the meantime CMIP6 models are displaying a wider range of ECSs (see below)."

6. L116: I'm pretty sure P(y|x)*P(x) can be written as a Gaussian function and therefore evaluated analytically rather than numerically integrated. You might have to use the fact that e^{x + C} = e^x * e^C for some constant C in conjunction with completing the squares to manage this. This comment isn't a big deal - numerical integration is fine - but analytic integration is more elegant.

Here, $P(y|x)P(x)$ cannot be written as a simple Gaussian function in x since $P(y|x)$ is not a Gaussian function in x itself (only in y when x is held constant): the variance in $P(y|x)$ non-trivially depends on x (equations (3) and (4)). Thus, the integration over x of $P(y|x)P(x)$ cannot be done analytically (to the best of our knowledge) and must be done numerically.

7. L118: I don't understand why you need to assume P(y|y0)=P(y0|y) in eq 6 and therefore that the prior is uniform. Perhaps you could explain this in more detail. As I see it, you are just assuming y has a Gaussian distribution with mean \hat{y} and variance sigma_{x_0}. These are definitely big assumptions, but don't imply a prior.

We changed the manuscript accordingly to explain this in more detail:

"[…]. This distribution can be interpreted as the posterior distribution in the regression model based on climate model output but constrained by matching the observable x. However, the observation of x (referred to as x_0) is not error free and has uncertainties associated with it. Assuming again an unbiased Gaussian, the resulting observational probability density for observing x_0 given the true value x is given by

$$P(x_0|x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{(x_0 - x)^2}{2\sigma_x^2}\right),$$

where σ_x^2 is the variance of the observation about the true value. Assuming a uniform prior $P(x) \propto 1$ with no cut-offs (i.e. the cut-offs are positive and negative infinity, which forms an "improper" prior) and using Bayes' theorem implies $P(x \,|\, x_0) = P(x_0 \,|x)$. In a final step, numerical integration is used to calculate the marginal probability density for the constrained prediction of the target variable y dependent on the observation x_0:

$$P(y|x_0) = \int_{-\infty}^{+\infty} P(y|x)P(x|x_0)dx.$$

By assuming a uniform prior in x we also assume a uniform prior in y (the true ECS), since x and y are linearly related (see equation (1)) – in other words, that an ECS near 8 K would be deemed just as probable as one near 4 K if both are equally consistent with the observational best estimate x_0. We do this for simplicity. The PDFs would shift somewhat lower with a broad prior on processes instead (see Sherwood et al. (2020)), but we are concerned here with how outcomes compare using CMIP5 vs. CMIP6 data, rather than the exact ranges obtained. Such comparisons are not sensitive to the prior."

8. I got a bit lost regarding which of your results depend on the Gaussian approach of sect 2.2 for what results use the bootstrapping of sect 2.3. Am I correct that the left panels of Fig 2-5 use linear regression and the standard error, the middle panels of these same figures use the Gaussian approach and everything else is based on bootstrapping? It would be useful to mention at the end of sect 2.2 and 2.3 what figures use the methodology just described.

Yes, you are correct. Since we removed the bootstrapping approach in the revised version of the manuscript (this includes the right panels in figures 2 to 5), we think that this should be less confusing now. To further clarify things, we added a short paragraph to section 3 that relates the left and right columns to the corresponding equations:

"The left columns in these figures show the emergent relationships including the uncertainty of the linear regressions (blue and orange shaded areas; see equation (3)) and the uncertainty in the observations (gray shaded area, see equation (5)). The right columns show the probability distributions of ECS in the original model ensemble (histogram) and the constrained distribution given by the emergent constraints (blue and orange line; see equation (6))."

9. L156: how did you choose the 11 constraints you evaluate? Readers may think you cherry picked the constraints that behaved poorly if you don't say explicitly why you chose the ones you did.

We choses these 11 emergent constraints based on their availability in the ESMValTool. We added this to section 2.2:

"We chose these particular emergent constraints since these were already implemented in the ESMValTool (see section 2.4) at the time of writing this study, which greatly facilitated this analysis."

10. L276 says Volodin was the first emergent constraint on ECS, which isn't true. Covey et al 2000 and Knutti et al 2006 provide earlier emergent constraints.

We rephrased the sentence and removed the statement that Volodin (2008) was the first emergent constraint on ECS. Thank you for spotting this.

11. L433: Bretherton and Caldwell (2020; https://journals.ametsoc.org/jcli/article/33/17/7413/348548/Combining-Emergent-Constraints-for-Climate) provide a multivariate technique for combining constraints on ECS. Doing so provided less conceptual insight than I expected -having most constraints predict high ECS led to the combined estimate also having high ECS with narrower uncertainty... which seems obvious in retrospect.

Thank you for the interesting reference which we added to the summary section.

Minor comments

1. L18 "which stem the major source" is wrong. I think you mean "which is the major source"?

Changed in the manuscript.

2. You often say things like "the emergent-constrained best estimate". "Emergent-constrained" doesn't make sense. I think you mean the "emergent-constraint-constrained".

Changed to “emergently-constrained” in the manuscript.

3. L66: you already gave the range of ECS in the previous line, so saying CMIP6 models exceed 5K is redundant / unnecessary.

Removed redundant part of the sentence.

4. Eq 3: x should either include or exclude "_m" on *both* sides of the equation.

We added the index *m* to the function argument *x* on the left side of equation (1) (former equation (3)).

5. Eq 8 uses P(y|x) from eq 6, which says it is an equation for P(y|x_0). I think eq 6 is really true for all x rather than just the observed value x_0. I suggest you remove mention of x_0 everywhere before eq 6.

We replaced x_0 by x everywhere except for equation (5) (former equation (7)).

6. Sect 2.3: does it really take 100,000 samples to characterize uncertainty in a correlation between the 20-50 samples you’re getting from the CMIP archive? I would guess 1000 iterations would be sufficient.

We removed the bootstrapping testing in the revised version of the manuscript.

7. L167: "Temperature (ERSST) is used": I’m confused because I thought you said you used HadISST on L164. Are you saying that the Brient + Schneider used ERSST?

Yes, that is correct. We rephrased the sentence to clarify this.

8. L351:’ve never seen the "(here: ...)" nomenclature you use. Do you mean "(e.g....)"?

Replaced “here:” by “ZHA: …; BRI: …” in the first appearance and removed “here:” altogether in the second appearance.

Reply to Thorsten Mauritsen (Referee)

Reviewer comments are marked in **bold**, our answers in red.

Review of "Emergent constraints on Equilibrium climate sensitivity in CMIP5: do they hold for CMIP6?" by Manuel Schlund and co-authors.

In this study a series of mostly process-oriented emergent constraints that were developed on earlier model ensembles are applied to the latest CMIP6 ensemble. This is a very welcome attempt and in a broad sense testing scientific reproducibility. My major concern is with the main conclusions drawn, or perhaps not drawn, from the results. The fact that estimated ECS based on these constraints increases roughly in proportion to the mean ECS increase from CMIP5 to CMIP6 suggests that these constraints are in not actually constraints on ECS, rather, at best they are constraints on the feedback processes they target. I develop argumentation this below, along with providing some more technical comments. I sign this review such that should the authors have any issues understanding my point they can contact me directly.

Sincerely,

Thorsten Mauritsen

We thank the reviewer for the helpful and constructive comments. We have revised our manuscript in light of these and the other reviewer's comments we have received. A pointwise reply is given below.

Major comments

Climate sensitivity is inversely proportional to the feedback parameter (lambda in equation 2) which in turn is a sum of a series of processes (sum of lambda_i). My take on the situation is that many of the emergent constraints (except COX) were successful on CMIP5 because inter-model spread in ECS in that ensemble was dominated by spread in low-level cloud feedbacks in the tropics. However, if any other feedback, e.g. water vapour feedback or any other cloud feedback, is biased in the ensemble as a whole then these kinds of process-oriented emergent constraints will necessarily be biased in their estimates of ECS. Probably even collectively since they thrive on the same kind of model spread, and so just because there are many studies that agree doesn't increase our confidence in their quantitative outcome. Likewise, if structural commonalities among models cause an unreasonable low inter model spread in some other feedback process then the emergent constraint is going to be over-confident. All in all, the results suggest that the original studies were overly confident and that changes in feedbacks not constrained by these studies cause them to be biased with a sign that cannot be determined (since CMIP6 probably also contains collectively biased feedback processes). Thus, these process-oriented emergent constraints are perhaps best thought of as constraints on the processes that they target, rather than constraints on ECS, and in extension the original studies have been disproven by the results of this study.

There are alternatives to process-oriented emergent constraints, though, one of them which is included in this study (COX, more about this study and why I think there is a shift below). Emergent constraints that use global temperature change as a predictor of ECS do not suffer from the same problem: even if one feedback is biased in a model ensemble the constraint can in principle still work since both global change and ECS are inversely proportional to the sum of feedbacks. Suggestions of emergent constraints of this kind include Last Glacial Maximum

(Hargreaves et al. 2012 doi:10.1029/2012GL053872), Pliocene warming (Hargreaves and Annan, 2016 doi:10.5194/cp-12-1591-2016), and post-1970s warming (Jimenez-de-la-Cuesta and Mauritsen 2019 doi:10.1038/s41561-019-0463-y). All of these ideas have been tested across ensembles including CMIP6/PMIP4 (Tokarska et al. 2020doi:10.1126/sciadv.aaz9549; Renoult et al. 2020 doi:10.5194/cp-2019-162), finding essentially unchanged results between ensembles. Other studies worth mentioning are Bender et al. (2010, doi:10.1007/s00382-010-0777-3) and Dessler and Forster(2018), although I haven't seen tests of these.

I think all of the above is rather straightforward and fairly easy to understand. I think the authors have everything at hand that they need to draw the conclusion that the process-oriented emergent constraints are not useful for estimating ECS, but rather should be better thought of as ideally constraining part of the cloud feedback. There are several places throughout that needs revising.

We think the reviewer has a very good point. We therefore extended the section 4 adding the following discussion:

"Our findings suggest that the process-oriented emergent constraints (i.e. all of the emergent constraints investigated here except COX) are only successful in constraining ECS as long as the uncertainty in ECS is dominated by the same process or feedback. In the CMIP5 ensemble, cloud feedback is the main contributor to the spread in ECS with low-level clouds in tropical subsidence regions dominating the spread in cloud feedback (e.g. Ceppi et al. (2017)). If any other process or feedback is biased (or missing) in the ensemble as a whole, then these process-oriented emergent constraints will be biased in their estimates of ECS. The appearance of diverse new feedback processes in CMIP6 could explain the reduced skill when applied to CMIP6 data, and a tendency for these to be positive would explain the upward shift in the model ECS distribution that is not captures by the CMIP5-trained constraints. Process-oriented emergent constraints are therefore perhaps best thought of as constraints on the processes that they target, rather than constraints on ECS.

Emergent constraints that use global temperature change as a way of constraining ECS could in principle not suffer from the same problem. If one feedback is biased in an ensemble the constraint might still work as both, global temperature change and ECS, might similarly reflect the sum of all feedbacks. Emergent constraints of this kind include e.g. the tropical temperature during the Last Glacial Maximum (Hargreaves et al., 2012), tropical temperature anomalies during the mid-Pliocene Warm Period (Hargreaves and Annan, 2016), and post-1970s warming (Jimenez-de-la-Cuesta and Mauritsen, 2019). This seems to be supported by the findings of Tokarska et al. (2020), who tested an emergent constraint for the transient climate response based on recent global warming trends on the CMIP5 and CMIP6 ensembles with similar results for both model ensembles. However, these temperature-based estimates are sensitive to assumptions about forcings and unforced decadal temperature variations, which could also be incorrect, as could model-predicted relationships between feedbacks on short and long time scales that are implicit in most such measures. Indeed, the significance of the COX constraint heavily dropped from CMIP5 to CMIP6 (p=0.0032 to p=0.5415), similar to other constraints in this study."

I think it is not reasonable to provide best estimates of ECS in the abstract and summary based on this study for the following reasons:

1) The above issue.

2) Because the study does not apply the latest observations to the constraints, rather opts for using the original observations. This is a perfectly fine choice given the scope of the paper, but it does mean the constraints are not up to date.

3) Because the study uses an implicit flat prior which in case of weak data automatically leads to high-biased results.

I would instead suggest the authors cite percentage increases which is anyway all that is relevant here.

We agree and therefore removed the best estimates of ECS from the abstract and the summary section and replaced them with percentage increases as suggested. The best estimates are now only given in the results section (3).

Regarding the Cox et al. 2018 constraint, section 3.2, this is built on the Hasselmann (1976) single heat capacity model. In this model there is a linear relationship between Psi and ECS, however, and despite what they claim if you add a deep ocean to the model you obtain a non-linear relationship wherein the relationship is weaker for higher ECS, see Annan et al. (2020 doi:10.5194/esd-11-709-2020), their figure 5 (note flipped axes). If you look at how the CMIP6 models are distributed they are simply situated in the flatter part of the expected curve, and if you fit a straight line to it you will obtain different slopes than for CMIP5.

The reviewer has a good point. We therefore added the following sentence to section 3.2:

"For example, Annan et al. (2020) showed that the assumed linear relationship between Ψ and ECS does not hold when adding a deep ocean to the model."

In this regard, and this applies not only to this study but most of these kinds, I am concerned with the general use of linear regression. The most silly example is SU constraint, where despite getting the wrong sign of the slope in CMIP6, you obtain a constraint on ECS. I think studies must be much more smart about their choice of statistical model, and not just use linear regression when non-linear behaviour is expected or other physical constraints can be applied such as a near-zero intercept, examples in Jimenez-de-la-cuesta-Otero and Mauritsen (2019), Annan et al. (2020) and Renoult et al. (2020). In case of process-oriented emergent constraints one could perhaps think of using Equation (2) in the form ECS~a/(b+x) where x is a process-oriented predictor. I am not saying the authors need to change this, but it would be worthwhile acknowledging that using linear regression, heedlessly, can lead to misleading and over-confident results.

We added the limitations of the approach including the references that you mention to the methods section that introduces emergent constraint methodology (section 2.2) and also briefly refer to this in the summary section:

"Another limitation of our approach is the statistical model itself. Similar to many other emergent constraint studies, we use an ordinary least squares linear regression model for each emergent constraint. However, in some cases this might not be appropriate, e.g. when we expect non-linear behavior or when physical constraints can be used to derive further constraints for the regression model like a zero intercept (Annan et al., 2020; Jimenez-de-la-Cuesta and Mauritsen, 2019; Renoult et al., 2020)."

I found the discussion of statistical significance somewhat disturbing. The chosen thresholds seem purely subjective, as far as I can tell. I would suggest to delete this whole discussion which seem rather pretentious.

Following this comment and a similar comment by Peter Caldwell (2nd reviewer) we decided to remove the whole bootstrap significance testing from the paper. Following the reviewers' suggestions, we replaced the bootstrapping method with a *t*-test on the correlation coefficient. The null hypothesis of this *t*-test is that no correlation exists between the predictor and ECS. In the revised version of the manuscript, we now give *p*-values of the emergent relationships that correspond to the probability that the absolute correlation is larger than $|r|$ even though the null hypothesis is true, i.e. the true underlying correlation coefficient is zero. Moreover, we do not use the *p*-values anymore to specify absolute significance (as you noted, our categories "highly significant", "barely significant", etc. were rather subjective), but only use them to specify relative significance, i.e. to indicate whether the statistical significance changes when moving from CMIP5 to CMIP6. Similar to our original bootstrapping approach, the new approach using the *t*-test shows that except for the ZHA and SHS constraints, all emergent relationships show a higher significance for the CMIP5 ensemble than for the CMIP6 ensemble.

I found Sections 4 and 5 rather long and repetitive. I would suggest revising and sharpening.

In order to address the two reviewers' comments, we changed sections 4 and 5 substantially. Section 4 (discussion) now discusses possible reasons for the change in skill in the emergent constraints when moving from CMIP5 to CMIP6. Section 5 (summary) now gives a summary and discusses limitations of our study, some of which are now described in the methods section.

<u>Minor comments</u>

19, 'of spread in ECS among models'

Replaced "source of uncertainty in ECS" by "source of spread in ECS among models".

36, 'concentration over pre-industrial levels'

Added suggestion.

69, I am not sure Forster et al. 2020 is correctly cited here

We removed the reference to Forster et al. (2020).

85-92, perhaps drop Delta from F, and when using a specific forcing in equation 2 write F_{4x} or something?

We greatly simplified section 2.1. In the revised version, the former equation (2) does not appear anymore.

90-93. did you account for model energy leakage and drift? Concerning drift, some do account for this, but is not always obvious if there is a best way, nevertheless you must document what you did.

We accounted for possible model drift when calculating ECS by subtracting a linear fit of the pre-industrial control simulation from the abrupt4xCO_2 experiment. Other than that, no explicit corrections for drift or energy leakage have been done. We added this information to section 2.1 of the revised manuscript:

"In this calculation, the linear fit of a corresponding pre-industrial control run is subtracted from the 4xCO2 run to remove any model drift that is present in the control climate without adding noise (Andrews et al., 2012). Other than that, we do not explicitly account for other problems such as energy leakage."

120-123, there are also an intermediate option, e.g. the Cauchy prior used in Annan and Hargreaves (2011, Climatic Change). Regarding the uniform prior, please specify which cut-off you use.

We use an "improper" prior with cut-offs at positive and negative infinity. and Following this comment and a similar comment by Peter Caldwell (2nd reviewer) we made the following changes to the manuscript:

"[…]. This distribution can be interpreted as the posterior distribution in the regression model based on climate model output but constrained by matching the observable x. However, the observation of x (referred to as x_0) is not error free and has uncertainties associated with it. Assuming again an unbiased Gaussian, the resulting observational probability density for observing x_0 given the true value x is given by

$$P(x_0|x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{(x_0 - x)^2}{2\sigma_x^2}\right),$$

where σ_x^2 is the variance of the observation about the true value. Assuming a uniform prior $P(x) \propto 1$ with no cut-offs (i.e. the cut-offs are positive and negative infinity, which forms an "improper" prior) and using Bayes' theorem implies $P(x \,|\, x_0) = P(x_0 \,|x)$. In a final step, numerical integration is used to calculate the marginal probability density for the constrained prediction of the target variable y dependent on the observation x_0:

$$P(y|x_0) = \int_{-\infty}^{+\infty} P(y|x)P(x|x_0)dx.$$

By assuming a uniform prior in x we also assume a uniform prior in y (the true ECS), since x and y are linearly related (see equation (1)) – in other words, that an ECS near 8 K would be deemed just as probable as one near 4 K if both are equally consistent with the observational best estimate x_0. We do this for simplicity. The PDFs would shift somewhat lower with a broad prior on processes instead (see Sherwood et al. (2020)), but we are concerned here with how outcomes compare using CMIP5 vs. CMIP6 data, rather than the exact ranges obtained. Such comparisons are not sensitive to the prior."

We emphasize that we focus on the relative differences in the constrained ECS distribution between CMIP5 and CMIP6, which does not depend on the choice of the prior.

154-155, this is a misinterpretation, the IPCC 'likely' statements refer to 66-100 percent probability.

Thanks for clarifying this! We replaced "IPCC likely" by "66% confidence interval (17–83%)".

158-159, I felt this statement could be made more informative by explaining that it is the covariance of clouds with surface temperature anomalies.

As suggested, we added this extra information to the beginning of sect 3.1.

180, 'results to choices made in the analysis'

Added suggestion to manuscript.

236-263, these constraints seem to have some legacy with Fasullo and Trenberth (2012, Science), perhaps worth mentioning if the authors agree?

We mentioned the Fasullo and Trenberth (2012) constraint in the sections describing the SU constraint (3.7) and the TIH constraint (3.8).

276-277, it is incorrect that Volodin (2008) was the first emergent constraint on ECS, there is Covey et al. 2000 and Knutti et al. 2006 before then.

We rewrote the sentence and removed the wrong part that stated that Volodin (2008) was the first emergent constraint on ECS. Thank you for spotting this.

320, I would suggest deleting 'describing the real world'

Removed "describing the real world".

354-355, The idea and strength of an emergent constraint is that you use something you can observe to predict ECS. It really shouldn't matter if a process is slightly different in the warmer 2xCO2 world.

We agree with the reviewer that it probably does not matter if a process is different in the future climate as long as the ESMs know about it. But there may be processes that are unimportant in the ESMs and hence not captured by the emergent constraints but that are important in reality. We clarified this by rephrasing the corresponding paragraph in section 4 of the revised manuscript:

"[…]. While this assumption seems to make sense, we do not know whether the ESMs cover all relevant processes of the real Earth system. For example, it may be possible that there exist processes that are unimportant in the ESMs (and hence are not captured by the emergent constraints) but are actually important in reality. This lack of relevant processes may lead to an overconfident constraint. Thus, the more complex ESMs of the CMIP6 ensemble are more likely to capture relevant processes of the true climate which leads to weaker emergent relationships. On the other hand, emergent constraints on the less complex CMIP3 and CMIP5 ensemble may be overconfident."

357-359, same applies here

See answer above.

360-361, this statement goes further than Zelinka et al. 2020. I would suggest replacing 'dominated by' -> 'to some extent associated with'

Changed as requested by the reviewer.

402-403, as per my above argumentation, I would be very careful with making this statement.

Since we removed the bootstrapping testing from the paper, we also removed this whole paragraph.

409, this might also have been shown by Klocke et al. (2011), check.

In Klocke et al. (2011), they state: „This suggests that model weighting based on statistical relationships alone is unfounded and perhaps that climate model errors are still large enough that model weighting is not sensible.”. In our opinion, this does not really fit into our argumentation, since we do not consider model interdependence and model skill at this point.

421-426. the very same paper also shows that the ECS estimated from 4xCO2 runs is higher than twice that in 2xCO2 runs, and that the bias of the same order of magnitude.

We did not find any reference to the reviewers point in Klocke et al. (2011) (https://doi.org/10.1175/2011JCLI4193.1).

Emergent constraints on Equilibrium Climate Sensitivity in CMIP5: do they hold for CMIP6?

Manuel Schlund[1], Axel Lauer[1], Pierre Gentine[2,3], Steven C. Sherwood[4], Veronika Eyring[1,5]

[1]Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany.

[2]Department of Earth and Environmental Engineering, Columbia University, New York, NY 10027.

[3]Earth Institute and Data Science Institute, Columbia University, New York, NY 10027.

[4]Climate Change Research Centre and ARC Centre of Excellence for Climate System Science, University of New South Wales, Sydney 2052, Australia.

[5]University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany.

Correspondence to: Manuel Schlund (manuel.schlund@dlr.de)

Abstract. An important metric for temperature projections is the equilibrium climate sensitivity (ECS) which is defined as the global mean surface air temperature change caused by a doubling of the atmospheric CO_2 concentration. The range for ECS assessed by the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report is between 1.5 and 4.5 K and has not decreased over the last decades. Among other methods, emergent constraints are potentially promising approaches to reduce the range of ECS by combining observations and output from Earth System Models (ESMs). In this study, we systematically analyze 11 published emergent constraints on ECS that have mostly been derived from models participating in the Coupled Model Intercomparison Project Phase 5 (CMIP5) project. These emergent constraints are – except for one that is based on temperature variability – all directly or indirectly based on cloud processes, which ~~stem~~are the major source of ~~uncertainty~~spread in ECS among current models. The focus of the study is on testing if these emergent constraints hold for ESMs participating in the new Phase 6 (CMIP6). Since none of the emergent constraints considered here has been derived using the CMIP6 ensemble, CMIP6 can be used for cross-checking of the emergent constraints on a new model ensemble. The application of the emergent constraints to CMIP6 data shows a ~~large~~ decrease in skill and statistical significance of the ~~correlation coefficients for most~~ emergent ~~relationships, indicating that~~ relationship for nearly all constraints, with this decrease being large in many cases. Consequently, the size of the ~~constraints are less skillful in predicting~~constrained ECS ~~than they were in~~ ranges (66% confidence intervals) increased by 51% on average (using the arithmetic mean of all emergent constraints) in CMIP6 compared to CMIP5. ~~Many do not appear sufficiently skillful to be useful in constraining ECS, and several of them do not pass a significance test.~~ This is likely because of changes in the representation of cloud processes from CMIP5 to CMIP6, but may in some cases also be due to spurious statistical relationships or a too small number of models in the ensemble the emergent constraint was originally derived from. The ~~emergent~~emergently-constrained best ~~estimate~~estimates of ECS also

increased from CMIP5 to CMIP6~~, with a best estimate range of 2.97 – 3.88 K for CMIP5 and 3.41 – 4.36 K for CMIP6.~~ by 12% on average. This can be at least partly explained by the increased number of high-ECS (above 4.5K) models in CMIP6 ~~with an ECS above 4.5 K~~ without a corresponding change in the constraint predictors, suggesting the emergence of new feedback processes rather than changes in strength of those previously dominant. Our results support previous studies concluding that emergent constraints should be based on an independently verifiable physical mechanism, and that emergent constraints focusing on specific processes contributing to ECS ~~are~~may be more ~~promising~~robust than ~~emergent constraints based on statistical model analysis~~those attempting to constrain the total.

1 Introduction

A bulk measure of the sensitivity of the climate system to carbon dioxide in the atmosphere (CO_2) is commonly expressed by the equilibrium climate sensitivity (ECS), an idealized metric defined as the mean global surface air temperature change that results from a doubling of the atmospheric CO_2 concentration over pre-industrial levels once the climate system reached equilibrium. In 1979, the Charney report determined an ECS range of 1.5 to 4.5 K for the Earth system (Charney et al., 1979). ~~This range has not changed substantially over time (Meehl et al., 2020). In the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (AR5), the reported likely range of ECS based on multiple lines of evidence is still between 1.5 and 4.5 K, whereas ECS calculated from climate and~~This range had not changed substantially by the time of the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (AR5) (Collins et al., 2013) and is close to the range of the Earth system models participating in the Coupled Model Intercomparison Project Phase 5 (CMIP5, Taylor et al. (2012)~~) supporting AR5 is ranging between 2.1 and 4.7 K (Collins et al., 2013).~~).

This large range of ~~global~~model climate sensitivity values can be largely attributed to differences in cloud feedbacks (Boucher et al., 2013). In particular, model differences in the change in shortwave reflection of low-level clouds changes in response to climate change dominate the uncertainties in the global warming projections, particularly in the tropics but also in mid-latitudes (Brient and Schneider, 2016; Vial et al., 2013). Over the years, various lines of evidence have been exploited to constrain the range of ECS, including paleoclimate data and analysis of the current observed warming trend (Knutti et al., 2017a). A new assessment using this evidence has narrowed the 66% range (17–83%) to 2.6–3.9 K (Sherwood et al., 2020), but in the meantime CMIP6 models are displaying a wider range of ECSs (see below).

The use of emergent constraints is another promising approach to reduce the uncertainty in ECS (Eyring et al., 2019). Originally applied to the hydrological cycle and the snow-albedo feedback (Allen and Ingram, 2002; Hall and Qu, 2006), emergent constraints offer the possibility to constrain future projections of Earth system model (ESM) ensembles with observations. Their theoretical basis is an emergent relationship between an observable quantity in the past or present-day climate and a quantity related to the future climate (such as for example ECS). Typically, the observable quantity is related to a climate feedback allowing the emergent relationship to be physically motivated by some key processes driving this feedback. Such a physical mechanism is a crucial prerequisite for the plausibility of an emergent constraint: due to large number of possible observables and small number of models, spurious emergent relationships are possible just by chance, which was shown by statistical tests (Caldwell et al., 2014). Caldwell et al. (2018) evaluated the credibility of several published emergent constraints on ECS. Using a feedback decomposition analysis, they assessed whether the published emergent relationship could be explained by the proposed mechanism. Out of 19 emergent constraints on ECS, only four of them were considered credible, while the rest of them were considered as either not plausible or could not be tested using this approach. In addition, Caldwell et al. (2018) performed out-of-sample tests on five emergent constraints originally trained on older CMIP versions, by applying them to the CMIP5 ensemble. They found that out only one of the five passed this test.

~~In this paper, we analyze 11 published emergent constraints on ECS which are summarized in Table 1 and assess~~In this paper, we follow up on the work of Caldwell et al. (2018) by analyzing 11 published emergent constraints on ECS which are summarized in Table 1, and assessing whether they still hold for the new CMIP6 model ensemble (Eyring et al., 2016). We first calculate these emergent constraints for the most recent ensemble ~~that was~~ used to derive ~~the majority~~all but one of them – CMIP5 (Taylor et al., 2012) – and then test whether they still hold in the CMIP6 ensemble. The ~~only~~one exception ~~here~~ is ~~an~~the emergent constraint ~~by~~of Volodin (2008)~~ that~~, which was derived on CMIP3 data. While the model range of ECS in CMIP5 is between 2.1 ~~to~~and 4.7 K, the CMIP6 model range is considerably larger~~ (~~, 1.8 to 5.6~~ K) with some models showing ECS values above 5~~ K (Meehl et al., 2020), see Figure 1. Possible reasons for this increased ECS range are changes in the extratropical cloud parametrizations and microphysics in the CMIP6 models (Zelinka et al., 2020). However, despite including more detailed cloud physical processes, further analyses suggest that the high sensitivity models might overestimate the future warming trend ~~(Forster et al., 2020; Tokarska et al., 2020)~~(Tokarska et al., 2020). The large ECS range in CMIP6 emphasizes the need for reliable methods to constrain the uncertainty range of future climate projections with observations. The CMIP6 ensemble can ~~therefore~~ be used for an independent testing of the constraints on previously unknown data. If the proposed underlying physical mechanisms are robust, i.e. targeting a key feedback mechanism controlling most of the observed CMIP6 spread, the emergent constraints would be expected to hold when applied to CMIP6 data. In this analysis we thus test the robustness of the constraints to new models and models with advances in model design over time.~~ In addition, we test the statistical significance of the relationships found in either model ensemble using a bootstrap resampling approach.~~

Section 2 provides an overview of the data and methods used. Section 3 gives a discussion of the 11 emergent constraints on ECS and their results when derived from the CMIP5 or CMIP6 ensemble including ~~a detailed~~an analysis of their statistical significance. The paper ends with a discussion and summary in sections 4 and 5.

2 Data and methods

2.1 Equilibrium Climate Sensitivity (ECS)

In this study we use the output from climate models participating in CMIP5 and CMIP6, see Table 2 and Table 3, respectively. Traditionally, ECS is defined as the global mean surface air temperature change after an instantaneous doubling of the atmospheric CO_2 concentration once the climate system reaches radiative equilibrium. Since running a fully-coupled ESM into equilibrium is computationally expensive (this would require thousands of model years, see Rugenstein et al. (2020)), ECS is typically approximated by a so-called “effective climate sensitivity~~”. This quantity~~”, which is ~~calculated by applying~~derived from the ~~Gregory~~first 150 years that follow an instantaneous quadrupling of the atmospheric CO_2 concentration (4xCO_2 run). Since the ESMs are not in radiative equilibrium during these 150 years, a regression ~~method~~of the top-of-atmosphere net downward radiation N versus the global mean surface air temperature change ΔT extrapolated to $N = 0$ gives an estimate of the equilibrium warming (Gregory et al., 2004)~~, based on an assumed relationship between forcing and response~~. In this paper, we use the term “ECS” to denote this effective climate sensitivity derived from the Gregory regression method. In this calculation, the linear fit of a corresponding pre-industrial control run is subtracted from the 4xCO_2 run to remove any model drift that is present in the control climate without adding noise (Andrews et al., 2012). Other than that, we do not explicitly account for other problems such as energy leakage.

$$\cancel{N = \Delta F + \lambda \Delta T} \qquad \cancel{(1)}$$

~~with the top-of-atmosphere (TOA) net downward radiation N equal to the heat storage flux in the system, the external greenhouse gas (here CO_2 only) forcing ΔF and a global mean surface air temperature change ΔT with λ the feedback parameter. This equation forms a linear regression model for $N(\Delta T)$ with slope λ and intercept ΔF. If these regression parameters are calculated from annual means of the quantities N and ΔT from a 150-year simulation where the atmospheric CO_2 concentration is abruptly quadrupled relative to a pre-industrial simulation and assume radiative equilibrium i.e. negligible heat storage flux ($N = 0$), the ECS can be evaluated as~~

$$\cancel{ECS = -\frac{F}{2\lambda}.} \qquad \cancel{(2)}$$

~~The factor 2 in the denominator accounts for the fact that we use an abrupt four times CO_2 simulation, whereas ECS is defined for a doubling of the atmospheric CO_2 concentration.~~

Even though widely used in literature, this Gregory regression method is known to be only an approximation of the true climate sensitivity. As shown by (Sherwood et al., 2020), the effective climate sensitivity is 6% lower than the best estimate of the true equilibrium warming obtained from integrating the climate models until a new equilibrium is reached. However, only a few ESMs provide simulations long enough to assess the true climate sensitivity. The CMIP endorsed LongRunMIP (Rugenstein et al., 2019) could be a promising way to estimate the true climate sensitivity that can then be used to reevaluate emergent constraints and their proposed underlying physical mechanisms.

2.2 Calculation of emergent constraints on ECS

An overview of the 11 emergent constraints analyzed in this study including the variables required for their calculations is given in Table 1 and the following section. ~~In all cases~~We chose these particular emergent constraints since these were already implemented in the ESMValTool (see section 2.4) at the time of writing this study, which greatly facilitated this analysis. For all emergent constraints, we use the historical simulations of CMIP5 and CMIP6 in order to ensure maximum agreement with the observational data. If necessary, the historical simulation of CMIP5 is extended after its final year 2005 with data from the

RCP 8.5 scenario (Riahi et al., 2011). Note that we only use data through 2014, during which time all RCP scenarios behave similarly and the choice of the scenario is not expected to affect results considerably. Such an extension is not needed for CMIP6 models as their historical simulations cover a longer time period until 2014.

To evaluate the resulting constrained probability distribution of ECS, we use the following nomenclature: let x_m be the x-axis variable (i.e. the observable, constraining variable) of climate model m and y_m its corresponding target variable (ECS in our case). Following Cox et al. (2018), we use ordinary least squares regression to fit the linear model

$$\hat{y}_m\cancel{(x) =}(x_m) = a + bx_m, \tag{1}$$

where $\hat{y}_m$ is the predicted target variable for predictor x_m, a the intercept of the linear regression line and b the slope of the linear regression line. Fitting the regression model includes minimizing the standard error s of the estimate

$$s^2 = \frac{1}{M-2}\sum_{m=1}^{M}(y_m - \hat{y}_m)^2, \tag{2}$$

where M is the total number of climate models.

In the standard emergent constraint approach, the constrained ~~best estimate for the~~ target variable y (here ECS) is given by the regression $\hat{y}(\cancel{x_0})x)$ evaluated at an observed or observationally based (in case of using reanalysis data) value $\cancel{x_0.}x$ that has not been used to fit the regression line. In that case, the corresponding uncertainty in ~~predicting that best estimate~~the prediction of $\hat{y}$ is given by the standard prediction error

$$\sigma_{\hat{y}}^2\cancel{(x_0) = s^2\left(1 + \frac{1}{M} + \frac{(x_0 - \bar{x})^2}{\sum_{m=1}^{M}(x_m - \bar{x})^2}\right)}(x) = s^2\left(1 + \frac{1}{M} + \frac{(x - \bar{x})^2}{\sum_{m=1}^{M}(x_m - \bar{x})^2}\right). \tag{3}$$

Here, $\bar{x}$ is the arithmetic mean of x over all models. Assuming Gaussian errors, this equation defines the conditional probability density function (PDF) for predicting a value of y given $\cancel{x_0,}x,$ i.e. the posterior distribution:

$$\cancel{P(y|x_0) = \frac{1}{\sqrt{2\pi\sigma_{\hat{y}}^2(x_0)}}\exp\left(-\frac{\left(y - \hat{y}(x_0)\right)^2}{2\sigma_{\hat{y}}^2(x_0)}\right).}P(y|x) = \frac{1}{\sqrt{2\pi\sigma_{\hat{y}}^2(x)}}\exp\left(-\frac{\left(y - \hat{y}(x)\right)^2}{2\sigma_{\hat{y}}^2(x)}\right). \tag{4}$$

This distribution can be interpreted as the posterior distribution in the regression model based on climate model output but constrained by matching the ~~observation x_0.~~observable x. However, the observation of x (referred to as x_0) is not error free and has uncertainties associated with it. Assuming again an unbiased Gaussian ~~uncertainties~~, the resulting observational probability density for ~~x~~ observing x_0 given the ~~observation x_0~~true value x is given by

$$\cancel{P(x) = \frac{1}{\sqrt{2\pi\sigma_{x_0}^2}}\exp\left(-\frac{(x - x_0)^2}{2\sigma_{x_0}^2}\right),}P(x_0|x) = \frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\left(-\frac{(x_0 - x)^2}{2\sigma_x^2}\right), \tag{5}$$

where $\cancel{x_0}\sigma_x^2$ is the ~~best estimate because the error is assumed unbiased, and $\sigma_{x_0}^2$ the~~ variance of the observation about the true value. Assuming a uniform prior $P(x) \propto 1$ with no cut-offs (i.e. the cut-offs are positive and negative infinity, which forms an "improper" prior) and using Bayes' theorem implies $P(x|x_0) = P(x_0|x)$. In a final step, numerical integration is used to calculate the marginal probability density for the constrained prediction of the target variable y~~:~~ dependent on the observation x_0:

$$\cancel{P(y) = \int_{-\infty}^{+\infty} P(y|x)P(x)dx}\; P(y|x_0) = \int_{-\infty}^{+\infty} P(y|x)P(x|x_0)dx. \tag{6}$$

By assuming a uniform prior in x we also assume a uniform prior in y (the true ECS), since x and y are linearly related (see equation (1)) – in other words, that an ECS near 8 K would be deemed just as probable as one near 4 K if both are equally consistent with the observational best estimate x_0. We do this for simplicity. The PDFs would shift somewhat lower with a broad prior on processes instead (see Sherwood et al. (2020)), but we are concerned here with how outcomes compare using CMIP5 vs. CMIP6 data, rather than the exact ranges obtained. Such comparisons are not sensitive to the prior.

As typically done in other studies proposing a single emergent constraint on ECS, we do not explicitly take model interdependency into account when applying the linear regression model (see equation (1)) to the model ensemble data. We simply assume that the individual data points (i.e. climate models) are independent. As some modeling groups provide output from multiple ESMs and some ESMs from different modeling groups share components and code, this is clearly not the case. Duplicated code in multiple models is expected to lead to an overestimation of the sample size of a model ensemble and may result in spurious correlations (Sanderson et al., 2015). Possible approaches could be to stop treating all models equally by either applying a model weighting based on a model's interdependence with the other models or by simply reducing the ensemble size considering models only that are above a given (yet to be defined) interdependence score. Promising approaches to quantify the model interdependency that could be followed include, for example, the studies of Sanderson et al. (2015); Sanderson et al. (2017) and Knutti et al. (2017b).

Moreover, our approach assigns equal model weights without taking model performance into account, i.e. agreement with observations. This issue is discussed in detail by Brient (2020). Another limitation of our approach is the statistical model itself. Similar to many other emergent constraint studies, we use an ordinary least squares linear regression model for each emergent constraint. However, in some cases this might not be appropriate, e.g. when we expect non-linear behavior or when physical constraints can be used to derive further constraints for the regression model like a zero intercept (Annan et al., 2020; Jimenez-de-la-Cuesta and Mauritsen, 2019; Renoult et al., 2020).

We further note that also observational uncertainties can play a role as using different observational datasets for a given variable as a proxy for observational uncertainty might lead to different emergent constraints. As this study uses only one combination of observational dataset(s) to calculate the emergent constraints as in the original published emergent constraint studies, the error estimations given by our analysis are expected to underestimate the true error. This could be investigated by systematic tests using different observational datasets and/or combinations of thereof as a proxy for observational uncertainty. Where available, additionally observational uncertainty estimates could be used to give better estimates of the constrained range of ECS. A major challenge associated with this is, however, to determine how observational uncertainties propagate to the space-time scales represented by the models because of the typically not well-known correlation of observational errors in space and time (e.g. Bellprat et al. (2017)).

~~In assigning probabilities via equation (6) we have assumed that $P(y|y_0) = P(y_0|y)$ (y_0 is the unknown true value of y) and thereby have implicitly assumed a uniform prior probability density in ECS — in other words, that an ECS near 8 K is just as probable as one near 4 K if both are equally consistent with the observation x_0. We do this for simplicity. Note that if we had instead applied this procedure to the climate radiative feedback parameter, which is arguably more physical than applying it to ECS (e.g. Roe and Baker (2007)), the resulting ECS PDFs would have non-symmetric shapes and different means. The conclusions of this study regarding changes from CMIP5 to CMIP6 would be unchanged however in either case. Even though it is also possible to use previous information (e.g., the range seen in GCMs) to inform a prior resulting in a narrower posterior PDF with the emergent constraint added, here we focus on the constraining power of emergent constraints on their own. To assess the effectiveness of a constraint, the emergent-constrained PDFs are compared with published ECS ranges based on expert judgements and other information: if the constraint-based PDF is wider than the one from previous estimates, the constraint brings little added value, even if its validity has been shown.~~

2.3 Statistical significance of emergent ~~constraints~~relationships

~~In this study, we evaluate the statistical significance of the different emergent constraints on ECS. The term "statistical significance" refers to the sensitivity of the regression model to changes in the input data, i.e. the removal or addition of datasets. The basis for this evaluation is a non-parametric bootstrapping approach (similar to the one used by Zhai et al. (2015)). For this, we generate 100,000 bootstrap samples of size M for every emergent constraint by randomly drawing from the original sample $\{(x_m, y_m)\}$ with replacement (M is the total number of climate models for which all data required for the considered emergent constraint are available). For each bootstrap sample, the linear (Pearson) correlation coefficient r is calculated. Using the probability distribution of the bootstrap samples of r, we define a p-value as the probability that r exhibits the opposite sign as originally expected from the emergent relationship (see shaded areas in the right columns of Figure 2 to Figure 5). In other words, $p \coloneqq \mathrm{CDF}(0)$ ($p \coloneqq 1 - \mathrm{CDF}(0)$) for an expected positive (negative) emergent relationship where CDF refers to the cumulative distribution function of the bootstrap samples of r. In this context we introduce the following nomenclature: an emergent relationship is called "highly significant" if $p < 0.02$, "barely significant" if $0.02 \leq p < 0.05$, "almost significant" if $0.05 \leq p < 0.1$, and "far from significant" if $p \geq 0.1$.~~

To evaluate the statistical significance of the different emergent relationships, we use a two-sided *t*-test to determine how likely the correlation found between the predictors and ECS would be to appear by chance. The null hypothesis for this test is that the predictor and ECS are not linearly correlated, i.e. the true underlying Pearson correlation coefficient of the population is zero. In that case, the variable

$$t = \frac{r\sqrt{M-2}}{\sqrt{1-r^2}} \tag{7}$$

has a Student's *t*-distribution with $M-2$ degrees of freedom. r is the Pearson correlation coefficient evaluated on the sample $\{(x_m, y_m)\}$. In this study, we indicate the statistical significance with the *p*-value, which describes the probability of obtaining an absolute sample Pearson correlation coefficient greater than $|r|$ if the null hypothesis is true, i.e. the predictor and ECS are not linearly correlated. Smaller *p*-values indicate higher significance. Although threshold values such as $p < 0.05$ are often used to declare "significance," here we focus mainly on how *p*-values are affected by the change from CMIP5 to CMIP6, noting that they may anyway be biased low by our assumptions discussed in section 2.2.

2.4 ESMValTool

All figures in this paper are produced with the Earth System Model Evaluation Tool (ESMValTool) version 2~~.0 (v2.0)~~ (Eyring et al., 2020; Lauer et al., 2020; Righi et al., 2020). The ESMValTool is an open-source community diagnostics and performance metrics tool for the evaluation of Earth system models (https://www.esmvaltool.org/). An ESMValTool recipe (configuration file defining input data, preprocessing steps and diagnostics to be applied) is available that can be used to reproduce all figures in this paper. This also allows redoing the analysis presented in this study once new model simulations from CMIP6 or other model ensembles become available.

3 Comparison of emergent constraints on ECS for CMIP5 and CMIP6

In this section we describe and discuss the 11 emergent constraints on ECS summarized in Table 1 using CMIP5 and CMIP6 data (sections 3.1 to 3.11) and provide a best estimate for ECS and statistical significance of the 11 emergent constraints in section 3.12. While most of these emergent constraints have been derived using data from the CMIP5 and/or CMIP3 ensembles, to our knowledge none of them has been evaluated on the CMIP6 ensemble so far. The results for the individual emergent constraints described in the following are shown in Figure 2 to Figure 5. The left columns in these figures show the emergent relationships including the uncertainty of the linear regressions (blue and orange shaded areas; see equation (3)) and the uncertainty in the observations (gray shaded area, see equation (5)). The right columns show the probability distributions of ECS in the original model ensemble (histogram) and the constrained distribution given by the emergent constraints (blue and orange line; see equation (6)). Table 4 shows corresponding ~~IPCC likely ranges (i.e.~~ 66% confidence ~~interval~~intervals (i.e. 17–83% intervals) of ECS derived from the probability distributions given by equation (6) and the *p*-values used to assess the significance of the emergent relationships.

3.1 Response of shortwave cloud reflectivity to changes in sea surface temperature (BRI)

In this emergent constraint proposed by Brient and Schneider (2016), ECS is correlated with the tropical low-level cloud (TLC) albedo~~.~~, i.e. using the covariance of clouds with changes in sea surface temperatures (SSTs). Differences in the TLC albedo account for more than half of the variance of the ECS in the CMIP5 ensemble. Following Brient and Schneider (2016), TLC regions are defined as grid points that are in the driest quartile of 500 hPa relative humidity of all grid cells over the ocean between 30°S and 30°N. The albedo of the TLC is obtained by calculating the ratio of TOA shortwave cloud radiative forcing and solar insolation averaged over the TLC region. The regression coefficients of deseasonalized variations of TLC shortwave albedo and ~~sea surface temperature~~ SST (in % per K) are then used as an emergent constraint for ECS. Here, we use observational data from HadISST for SST (Rayner et al., 2003), ERA-Interim for 500 hPa relative humidity (Dee et al., 2011) and CERES-EBAF (Loeb et al., 2018) for the TOA radiative fluxes over the time period 2001–2005. ~~With the exception of SST, where data from the Extended Reconstructed Sea Surface Temperature (ERSST) is used~~In the original publication, Brient and Schneider (2016) use similar observation-based datasets with the exception of SST, where they take data from the Extended Reconstructed Sea Surface Temperature (Smith and Reynolds, 2003)~~, the original publication used similar observation-based datasets.~~ as reference instead. Our analysis yields a ~~likely~~66% confidence range for ECS of 3.72 K ± 0.~~56~~59 K for CMIP5 ($R^2 = 0.38$) and 4.~~36~~32 K ± 1.~~16~~07 K for CMIP6, with much lower $R^2 = 0.$~~15~~12. The original publication stated a best estimate of 4.0 K, with a very low likelihood of values below 2.3 K (90% confidence). ~~The PDFs of the Pearson correlation coefficient r obtained by the non-parametric bootstrapping approach (right column in Figure 2) show that the emergent relationship exhibits the expected negative correlation for CMIP5 and CMIP6. The emergent relationship is highly significant for CMIP5 ($p = 0.0002$) and barely significant for the CMIP6 ensemble ($p = 0.0219$).~~The statistical significance of the emergent relationship dropped from $p = 0.0005$ for CMIP5 to $p = 0.0355$ for CMIP6.

3.2 Temperature variability (COX)

The emergent constraint on ECS proposed by Cox et al. (2018)~~ uses~~ uses a temperature variability metric Ψ which is based on the interannual variation of global mean temperature calculated from its variance (in time) and one-year-lag autocorrelation. In contrast to the majority of emergent constraints which focus on cloud-related processes, this constraint is based on the fluctuation-dissipation theorem, which relates the long-term response of the climate system to an external forcing (ECS) to short-term variations of the climate system (climate variability). This arguably places the constraint on a more solid theoretical foundation, although several questions were raised on the robustness of the results to choices made in the analysis (Brown et al., 2018; Po-Chedley et al., 2018; Rypdal et al., 2018)~~.~~. For example, Annan et al. (2020) showed that the assumed linear relationship between Ψ and ECS does not hold when adding a deep ocean to the model. As observational data, here we use the HadCRUT4 dataset (Morice et al., 2012) over the time period 1880~~–~~–2014. ~~For~~Under the COX constraint~~,~~ we thus assess a ~~likely~~66% ECS range of 3.03 K ± 0.~~71~~73 K for CMIP5 ($R^2 = 0.31$) and 3.~~44~~71 K ± 1.~~15~~09 K for CMIP6 ($R^2 = 0.$~~08~~01). Cox et al. (2018)~~ derived a likely range of 2.8 K ± 0.6 K from a different subset of CMIP5 models. For CMIP6, the distribution of r evaluated from the bootstrap samples (right column of Figure 2) shows high probability densities around $r = 0$, which means that many bootstrap samples show a very low correlation coefficient. Moreover, while the majority of bootstrap samples indicate a positive r, there is also a considerably high fraction of bootstrap samples that show a negative correlation (orange shaded area). In contrast to that, the distribution of r for the CMIP5 ensemble supports a clear positive correlation. Consequently, the COX emergent relationship is highly significant for the CMIP5 ensemble ($p = 0.0010$), but only almost significant for the CMIP6 ensemble ($p = 0.0545$).~~ derived a 66% range of 2.8 K ± 0.6 K from a different subset of CMIP5 models but the same observations. When moving from CMIP5 to CMIP6, the significance of the emergent relation drops massively from $p = 0.0032$ to $p = 0.5415$, respectively.

3.3 Southern hemisphere Hadley cell extent (LIP)

The results of Lipat et al. (2017) show that the multi-year average extent of the Hadley cell correlates with ECS in CMIP5 models. The Hadley cell edge is defined as the latitude of the first two grid cells from the equator going south where the zonal average 500 hPa mass stream function calculated from December-January-February means of the meridional wind field changes sign from negative to positive. Lipat et al. (2017) explain this correlation by tying it to the observed correlation of the interannual variability in mid-latitude clouds and their radiative effects with the poleward extent of the Hadley cell. For the calculation of the emergent constraint, we use reanalysis data from ERA-Interim (Dee et al., 2011) for the meridional wind speed over the time period 1980–2005. Our application of this emergent constraint gives ECS ~~likely~~66% ranges of 2.97 K ± 0.~~76~~75 K for CMIP5 ($R^2 = 0.18$) and 3.~~66~~75 K ± 1.~~27~~11 K for CMIP6 (R^2 ~~=~~< 0.~~03~~01). The original publication does not specify an ECS range. ~~The~~For CMIP6, the emergent constraint ~~by Lipat et al. (2017) is highly significant~~shows a much lower statistical significance ($p = 0.6791$) than for CMIP5 (p = 0.~~0043) but far from significant for CMIP6 (p = 0.2039~~0228).

3.4 Large-scale lower-tropospheric mixing (SHD)

Sherwood et al. (2014) proposed that the degree of mixing in the lower troposphere determines the response of boundary-layer clouds and humidity to climate warming, as the associated moisture transport would increase rapidly in a warmer atmosphere due to the Clausius-Clapeyron relationship. The large-scale component D of this mixing is defined as the ratio of shallow to deep overturning. D is calculated from the vertical velocities averaged over two height regions: 850 hPa and 700 hPa for shallow overturning and 600, 500 and 400 hPa for deep overturning. Both quantities are averaged over parts of the tropical ocean region away from the regions of highest SST and strongest mid-level ascent, in particular the region 160°W – 30°E, 30°S – 30°N, wherever air is ascending at low levels. As observationally based data, we use vertical velocities from ERA-Interim (Dee et al., 2011) over the time period 1989–1998 similar to the original publication. We derive ECS ~~likely~~66% confidence ranges of 3.65 K ± 0.~~63~~64 K for CMIP5 ($R^2 = 0.28$) and 3.~~74~~77 K ± 1.~~11~~06 K for CMIP6 (R^2 = 0.~~04~~03). Sherwood et al. (2014) do not give a best estimate for ECS based on the large-scale component of mixing D or its small-scale counterpart S (section 3.5) but for the sum of D+S only (see section 3.6). ~~The evaluation of the bootstrap distribution of r indicates that the SHD constraint is highly significant for the CMIP5 ensemble (p = 0.0006) but far from significant for the CMIP6 ensemble (p = 0.1120~~The regression shows a much lower significance for CMIP6 ($p = 0.2805$) than for CMIP5 ($p = 0.0037$).

3.5 Small-scale lower-tropospheric mixing (SHS)

The small-scale mixing S (Sherwood et al., 2014) is calculated from the differences in relative humidity and temperature between 700 and 850 hPa. The differences are averaged over all grid cells within the upper quartile of the annual mean 500 hPa ascent rate (within ascending regions) in the tropics. The tropics are defined as region between 30°S and 30°N. In the Cloud Feedback Model Intercomparison Project models (CFMIP, Webb et al. (2017)), for which convective tendencies were available, upward moisture transport by parameterized convection was shown to increase more rapidly with warming for higher values of S. We use reanalysis data from ERA-Interim (Dee et al., 2011) for temperature and relative humidity to calculate the observationally based constraint (1989–1998). Our analysis shows a ~~likely~~66% range of ECS of 3.07 K ± 0.~~68~~73 K for CMIP5 ($R^2 = 0.13$) and 3.~~41~~48 K ± 1.~~13~~07 K for CMIP6 ($R^2 =$ 0.~~15~~12). The correlation of S and ECS ~~is almost significant for the CMIP5 ($p = 0.0581$) and~~shows a slightly higher significance in the CMIP6 ~~($p = 0.0638$)~~ ensemble~~.~~ ($p = 0.0396$) than in the CMIP5 ensemble ($p = 0.0647$). The SHS constraint is one of the two analyzed emergent constraints (ZHA being the other exception) that shows a higher ~~coefficient of determination R^2~~statistical significance for the CMIP6 than for the CMIP5 ensemble.

3.6 Lower tropospheric mixing index (SHL)

The lower tropospheric mixing index (LTMI) formulated by Sherwood et al. (2014) is defined as the sum of the small-scale mixing S (see section 3.5) and the large-scale mixing D (see section 3.4), which are supposed to capture complementary components of the total mixing phenomenon. Sherwood et al. (2014) argue that the increase in dehydration depends on initial mixing linking it to cloud feedbacks and thus also to ECS. For this constraint, we derive an ECS ~~likely~~66% confidence range of 3.42 K ± 0.~~63~~65 K for CMIP5 ($R^2 = 0.41$) and 3.~~65~~67 K ± 1.~~05~~06 K for CMIP6 ($R^2 =$ 0.~~19~~16). Sherwood et al. (2014) give a best estimate of about 4 K with a lower limit of 3 K. ~~As illustrated by the right column of Figure 3, the SHL emergent relationship is highly significant for both considered climate model ensembles, CMIP5 ($p = 0.0001$) and CMIP6 ($p = 0.0114$), although weaker in the newer ensemble.~~Similar to both other constraints by Sherwood et al. (2014), SHD and SHS, the statistical significance of the SHL emergent relation decreased in CMIP6 ($p = 0.0138$) compared to CMIP5 ($p = 0.0002$).

3.7 Error in vertical profile of relative humidity (SU)

Another emergent constraint on ECS that targets uncertainties in cloud feedbacks was proposed by Su et al. (2014). They show that changes in the Hadley circulation are physically connected to changes in tropical clouds and thus ECS. Consequently, the inter-model spread in the change of the Hadley circulation in an ensemble of climate models is well correlated with the corresponding changes in the TOA cloud radiative effect. Moreover, Su et al. (2014) found a correlation between a model's ECS and its ability to represent the present-day Hadley circulation. The latter is calculated from the tropical (45°S – 40°N) zonal-mean vertical profiles of relative humidity from the surface to 100 hPa. These profiles are then used to define the x-axis of the SU constraint by calculating a performance metric based on the slope of the linear regression between a climate model's relative humidity profile and the corresponding observational reference. Similarly to the original publication, we use humidity observations from AIRS (Aumann et al., 2003) for pressure levels greater than 300 hPa and MLS-Aura data (Beer, 2006) for pressure levels of less than 300 hPa. Our analysis yields a constrained ~~likely~~66% range of ECS of 3.30 K ± 0.~~90~~88 K for CMIP5 ($R^2 = 0.08$) and 3.~~69~~77 K ± 1.~~59~~35 K for CMIP6 ($R^2 =$ 0.~~03~~05). The original publication gives a best estimate of 4 K with a lower limit of 3 K. Figure 4 shows that in addition to the low R^2 values, the emergent relationship shows different slopes for CMIP5 and CMIP6. For the CMIP5, the expected positive correlation is found, while for CMIP6, a negative correlation is found. This suggests that the constraint is not working (any more) when applied to the CMIP6 data. ~~Consequently, the SU constraint is almost significant for the CMIP5 ensemble ($p = 0.0919$) and far from significant for the CMIP6 ensemble ($p = 0.8573$).~~Consequently, the SU constraint shows a weaker statistical significance in the CMIP6 ensemble ($p = 0.1935$) than for the CMIP5 ensemble ($p = 0.1676$). The SU constraint is related to an emergent constraint on ECS proposed by Fasullo and Trenberth (2012), who correlated May-August zonal-mean relative humidity against ECS. In contrast to Su et al. (2014), they did not use the entire tropics, but identified two distinct regions with largest correlation.

3.8 Tropical mid-tropospheric humidity asymmetry index (TIH)

Tian (2015) found a link between mid-tropospheric humidity over the tropical Pacific and simulated moisture, precipitation, clouds, and large-scale circulation and thus ECS in CMIP3 and CMIP5 models. The study explains this link with the similarity of mid-tropospheric humidity and precipitation patterns as both are related to the ITCZ. The proposed tropical mid-tropospheric humidity asymmetry index to constrain ECS is defined as relative bias (in percent) in simulated annual mean 500 hPa specific humidity averaged over the Southern Hemisphere (SH) tropical Pacific (30°S – 0°, 120°E – 80°W) minus the bias averaged over the Northern Hemisphere (NH) tropical Pacific (20°N - 0°, 120°E - 80°W) when compared with observations. Similar to the SU constraint, the index proposed by Tian (2015) seems to be related to the emergent constraint by Fasullo and Trenberth (2012), who found correlations between relative humidity of the middle and upper troposphere and ECS. Here, we use humidity observations from AIRS (Aumann et al., 2003) over the time period 2003–2005 as reference dataset. We assess a ~~likely~~66% ECS range of 3.88 K ± 0.~~78~~75 K for CMIP5 ($R^2 = 0.24$) and 4.~~07~~15 K ± 1.~~21~~10 K for CMIP6 ($R^2 = 0.$~~07~~06). Tian (2015) specifies a best estimate of 4.0 K. The significance of the emergent relationship ~~is highly significant for the~~ dropped massively from $p = 0.0089$ in CMIP5 ~~ensemble ($p = 0.0002$) and barely significant for the~~ to $p = 0.1348$ in CMIP6~~ ensemble ($p = 0.0454$).~~.

3.9 Southern ITCZ index (TII)

In addition to the humidity index, Tian (2015) proposed an emergent constraint on ECS based on the southern ITCZ index (Bellucci et al., 2010; Hirota et al., 2011). This index is defined as the climatological annual mean precipitation bias averaged over the south-eastern Pacific (30°S – 0°, 150°W – 100°W). The southern ITCZ index is calculated in mm day^{-1} and dominated by the so-called double ITCZ, a common problem in many CMIP5 climate models. Tian (2015) found a link between double-ITCZ bias and simulated moisture, precipitation, clouds, and large-scale circulation in CMIP3 and CMIP5 models. He argues that this could explain the link found between the double-ITCZ bias and ECS. As reference data, we use observed precipitation data for the years 19862005 from GPCP (Adler et al., 2003). We calculate an ECS ~~likely~~66% confidence range of 3.87 K ± 0.~~70~~67 K for CMIP5 ($R^2 = 0.33$) and ~~4.11~~3.84 K ± 1.~~17~~09 K for CMIP6 (R^2 ~~=~~< 0.~~06~~01). Tian (2015) specifies a best estimate of 4.0 K. ~~As shown by the right column of Figure 4, the TII~~The emergent relationship ~~is highly significant for~~

~~the~~shows a much lower statistical significance in CMIP6 ($p = 0.8236$) than in CMIP5 ~~ensemble ($p = 0.0004$), but only almost significant for the CMIP6 ensemble~~ ($p = 0.$~~0634~~0013).

3.10 Difference between tropical and mid-latitude cloud fraction (VOL)

The study by Volodin (2008) aims at the geographical distribution of clouds in climate models. ~~It was the first published emergent constraint on ECS, relying on models from CMIP3, such that both CMIP5 and CMIP6 are out-of-sample tests for this constraint. He shows that high ECS models tend to simulate a higher total cloud cover over the southern mid-latitudes and a lower total cloud cover over the tropics (relative to the multi-model mean).~~Since this early emergent constraint was originally trained on CMIP3 models, both CMIP5 and CMIP6 are out-of-sample tests for it. Volodin (2008) shows that high ECS models tend to simulate a higher total cloud cover over the southern mid-latitudes and a lower total cloud cover over the tropics (relative to the multi-model mean). This can be used to establish an emergent relationship between the ECS and the difference in tropical total cloud cover (28°S – 28°N) and the southern mid-latitude total cloud cover (56°S – 36°S). Analogous to the original study, we use the ISCCP-D2 data (Rossow and Schiffer, 1991) as observational reference. For the VOL constraint, we calculate a constrained ~~likely~~66% range of ECS of 3.74 K ± 0.64 K for CMIP5 ($R^2 = 0.38$) and 4.~~14~~21 K ± 1.~~13~~04 K for CMIP6 ($R^2 = 0.$~~16~~18), whereas the original publication gives a range of 3.6 K ± 0.4 K (standard deviation) for a climate model ensemble of CMIP3 models. The emergent constraint by Volodin (2008) ~~is one of the two emergent constraints that is highly significant for both~~shows a lower significance in the CMIP6 ensemble ($p = 0.0056$) than in the CMIP5 ensemble ($p = 0.0004$~~) and the CMIP6 ($p = 0.0057$) ensemble.~~).

3.11 Response of seasonal marine boundary layer cloud fraction to SST changes (ZHA)

Zhai et al. (2015) focus on the variations of marine boundary layer clouds (MBLC), which largely contribute to the shortwave cloud feedback and thus to the uncertainty in modeled ECS. Their central quantity is the response of the MBLC fraction to changes in the sea surface temperature (SST) in subtropical oceanic subsidence regions for both hemispheres (20° – 40°). On short (seasonal) and long (centennial under a forcing) time scales, this quantity is well correlated with ECS among an ensemble of CMIP3 and CMIP5 models. Together with observations of cloud fraction from CloudSat/CALIPSO (Mace et al., 2009), SST from AMSRE SST (AMSR-E, 2011) and vertical velocity from ERA-Interim (Dee et al., 2011), the seasonal response of MBLC fraction to changes in SST forms an emergent constraint on ECS. We assess a ~~likely~~66% ECS range of 3.35 K ± 0.~~72~~74 K for CMIP5 ($R^2 = 0.05$) and 3.~~81~~79 K ± 0.~~60~~67 K for CMIP6 ($R^2 = 0.$~~61~~62). In their original publication, Zhai et al. (2015) found an ECS range of 3.90 K ± 0.45 K (standard deviation) for a combination of CMIP3 and CMIP5 models. In terms of statistical significance, the results of the ZHA constraints are somewhat surprising: although CMIP5 data (in combination with CMIP3 data) were successfully used in their original publication, our approach finds that the statistical significance of the emergent relationship ~~for the~~ is much higher in the unseen CMIP6 ensemble ($p < 0.0001$) than in the previously available CMIP5 ~~models is far from statistically significant.~~ensemble ($p = 0.2567$). The ZHA constraint is the only emergent constraint analyzed here that shows this extreme behavior (only one other constraint, SHS, shows a slightly higher significance in CMIP6; all other constraints show lower significances in CMIP6). The reason for ~~this disagreement~~the erratic skill in CMIP5 is the set of climate models used. For our analysis, we use 11 additional CMIP5 models that were not used in the original publication (i.e. ACCESS1-0, ACCESS1-3, bcc-csm1-1, bcc-csm1-1-m, CCSM4, GFDL-ESM2G, GFDL-ESM2M, IPSL-CM5A-MR, IPSL-CM5B-LR, MPI-ESM-MR and MPI-ESM-P). Due to a lack of publicly available data, the model CESM1-CAM5 that is used in the original publication is not included in our analysis. The effect of choosing different subsets of CMIP5 models on the emergent relationship is illustrated in Figure 6. Using the original CMIP5 models from the original publication gives a considerably higher correlation ($R^2 = 0.38$) than using all available CMIP5 models ($R^2 = 0.05$). This result shows a strong dependency of this emergent constraint on the subset of climate models used. ~~Moreover, strangely and uniquely among the metrics examined here, the ZHA constraint is highly significant for the CMIP6 ensemble ($p = 0.0022$) but far from significant in the updated CMIP5 ensemble ($p = 0.1195$).~~Nonetheless, the performance on CMIP6 models is, surprisingly, the best of all the constraints, and much better than on either subset of CMIP5 models.

3.12 ~~Best estimates for~~Constrained ECS ranges and statistical significance of the 11 emergent constraints

In most cases, the emergent relationships (left columns of Figure 2 to Figure 5) show the same sign of the slope (as expected from the theory) for CMIP5 and CMIP6, with the SU constraint being the only exception. However, the coefficient of determination (R^2) is ~~considerably~~ lower for CMIP6 compared to CMIP5 for all ~~but two constraints: SHS and~~one constraint: ZHA. The probability distributions of the constrained ECS that we obtain (~~middle~~right columns of Figure 2 to Figure 5) give similar results: except for the ZHA constraint, the constraint on the CMIP6 ensemble is weaker, i.e. the constrained PDFs derived from the CMIP6 ensemble are broader than their respective CMIP5 counterparts. As shown in Table 4, for CMIP5, the range of the best (maximum likelihood) estimates for ECS is 2.97 K to 3.88 K, while the corresponding CMIP6 best estimates are higher for ~~each of the~~almost every tested emergent constraint (TII being the only exception), resulting in a range of best estimates of 3.48 K to 4.32 K. Using the arithmetic mean of all analyzed emergent constraints, this results in a mean increase of the ECS best estimate of 12% in CMIP6 compared to CMIP5. Similarly, the size of the 66% ECS ranges (17–83% confidence) shows values of 1.16 K to 1.75 K in CMIP5 and 1.32 K to 2.70 K in CMIP6, resulting in an ~~ECS range of 3.41 K to 4.36 K.~~

~~The~~increase of 51% averaged over all emergent constraints. In summary, the R^2 of the emergent relationships and the constrained range of ECS each ~~show a strong dependency~~depend strongly on the climate model ensemble used, even though a physical explanation is given for each emergent constraint that is thought to be valid for every climate model ensemble ~~describing the real world. In order to assess changes in the skill of the emergent constraints when moving from CMIP5 to CMIP6 we use the degree of statistical significance relative to a standard $p = 0.05$ threshold (“highly”, “barely”, “almost”, “far from”) of the different emergent constraints shown in the right columns of Figure 2 to Figure 5 as a proxy. We consider reductions in the skill of the constraint as significant if the interquartile ranges of the bootstrapped correlation coefficient r for CMIP5 and CMIP6 data do not overlap (colored boxes in Figure 7). This failure to overlap is seen for all emergent constraints~~

~~but SHS, confirming that most of the constraints have lost more skill than could be explained, by sampling uncertainty alone, if the models were independent. Only two constraints (SHL and VOL) show skill with high significance for both the CMIP5 and CMIP6 ensembles. Two more emergent constraints (BRI and TIH) are highly significant for CMIP5 but barely significant for CMIP6. Three other constraints are far from significant on CMIP6 (LIP, SHD and SU), but only one fails a significance test on CMIP5 (ZHA). Figure 7 gives an overview showing boxplots of the distributions for each of the 11 emergent constraints. Following the definition given in section 2.3, an emergent relationship is either highly or barely significant if the whiskers spanning the one-sided 95% confidence interval do not cross the horizontal line indicating $r = 0$. Emergent relationships with whiskers crossing that line are either almost significant or far from significant. For CMIP5, eight constraints are either highly or barely significant (BRI, COX, LIP, SHD, SHL, TIH, TII and VOL), while for CMIP6, only five constraints are either highly or barely significant (BRI, SHL, TIH, VOL and ZHA). Note that for better comparability, the signs of the correlation coefficients r from the emergent relationships with expected negative slopes have been changed in Figure 7. Thus, positive values of r in the figure indicate that the correlation coefficient matches the sign of the expected (published) relationship, while negative values indicate that the correlation coefficient does not match the expected sign. The latter is only the case for the SU constraint evaluated on the CMIP6 ensemble~~. The same behavior is found for the statistical significance of the emergent relationships using the null hypothesis that there is no correlation between the predictor and ECS (see section 2.3). Except for the ZHA and SHS constraints, every emergent relationship investigated shows a lower statistical significance (i.e. higher p-value) in the CMIP6 ensemble than in the CMIP5 ensemble. If a conventional significance test of $p < 0.05$ was applied, eight of the 11 constraints would pass this test on CMIP5 model data but only five (BRI, SHL, SHS, VOL and ZHA) would pass this test on CMIP6. This is still much better than would be expected purely by chance. Hence, there is still skill in at least a few of the constraints, but it is much less than suggested in nearly all of the initial studies.

4 Discussion

As shown in the previous sections, most emergent relationships show smaller coefficients of determination when evaluated on the new CMIP6 ensemble compared to the CMIP5 ensemble. In this section, we discuss possible reasons for, and implications of, these differences. As reported by Caldwell et al. (2014), the large amount of data provided by modern ESMs can generate spurious correlations of variables between past climate and ECS just by chance, especially when only a small number of climate models is considered. This would cause the performance of the emergent constraint to be reduced on out-of-sample data (like the new CMIP6 ensemble), since the emergent relationship appeared just by chance and not because of a physically based mechanism.

A further reason for the weaker emergent relationships in CMIP6 may be the increased complexity of the participating ESMs. Each emergent constraint approach is based on the assumption that a single observable process or physical aspect in the current climate dominates the uncertainty in ECS. Some emergent constraints such as ZHA and BRI relate changes in cloud properties (~~here~~ZHA: low-level cloud fraction~~ and~~; BRI: cloud reflectivity) on seasonal or interannual time-scales ~~(here:~~ driven by changes in SST~~)~~ to ECS. This means that it has to be implicitly assumed that the observable changes in these properties on seasonal or interannual time-scales are ~~basically driven by the same mechanisms as the changes in cloud properties~~related to those occurring as a result of climate forcing~~.~~ in a way successfully captured by the ESMs. While this assumption seems to make sense, we do not know whether ~~it actually holds in a significantly different climate or if other or additional mechanisms also might become important. For this reason it also remains unclear if~~the ESMs cover all relevant processes of the ~~regions or cloud regimes that have been selected based on present-day climate and that are used to calculate~~real Earth system. For example, it may be possible that there exist processes that are unimportant in the ESMs (and hence are not captured by the emergent constraints~~ will be equally important under significant climate change. For example Lauer et al. (2010) showed with a regional climate model that~~) but are actually important in reality. This lack of relevant processes may lead to an overconfident constraint. Thus, the ~~relationship between cloud amount and lower tropospheric stability in~~ more complex ESMs of the ~~stratocumulus deck over~~CMIP6 ensemble are more likely to capture relevant processes of the ~~Southeast Pacific derived from present-day data will~~true climate which leads to weaker emergent relationships. On the other hand, emergent constraints on the less complex CMIP3 and CMIP5 ensemble may be ~~altered under global warming~~overconfident.

For CMIP6 models, Zelinka et al. (2020) showed that cloud feedbacks and thus ECS in high-sensitivity models are ~~dominated by~~to some extent associated with changes in clouds over the Southern Ocean, while in CMIP3 and CMIP5 the uncertainty in cloud feedbacks is dominated by clouds in the subtropical subsidence regions. One might speculate that a possible reason for this might be an improved simulation of clouds over the Southern Ocean in some models (Bodas-Salcedo et al., 2019; Gettelman et al., 2019a), as shown for some pre-CMIP6 model versions evaluated by Lauer et al. (2018). The findings of Zelinka et al. (2020) ~~could~~can also at least partly explain the larger inter-model spread in climate sensitivity due to ~~more and different regions / clouds types dominating the~~a greater diversity of cloud feedbacks~~ resulting~~, which also results in a weaker emergent constraint compared with CMIP5 models, as most of them constrained low-cloud feedbacks. They found that on average, the shortwave low cloud feedback is larger in CMIP6 than in CMIP5, which they primarily relate to changes in the representation of clouds. As a possible explanation, Zelinka et al. (2020) give an increase in mean-state supercooled liquid water (i.e. increase in the cloud water liquid fraction) in mixed-phase clouds resulting in less pronounced increases in low-level cloud cover and water content with warmer SSTs particularly in mid-latitudes.

~~We~~Our findings suggest that the process-oriented emergent constraints (i.e. all of the emergent constraints investigated here except COX) are only successful in constraining ECS as long as the uncertainty in ECS is dominated by the same process or feedback. In the CMIP5 ensemble, cloud feedback is the main contributor to the spread in ECS with low-level clouds in tropical subsidence regions dominating the spread in cloud feedback (e.g. Ceppi et al. (2017)). If any other process or feedback is biased (or missing) in the ensemble as a whole, then these process-oriented emergent constraints will be biased in their estimates of ECS. The appearance of diverse new feedback processes in CMIP6 could explain the reduced skill when applied to CMIP6 data, and a tendency for these to be positive would explain the upward shift in the model ECS distribution that is not captures by the CMIP5-trained constraints. Process-oriented emergent constraints are therefore perhaps best thought of as constraints on the processes that they target, rather than constraints on ECS.

Emergent constraints that use global temperature change as a way of constraining ECS could in principle not suffer from the same problem. If one feedback is biased in an ensemble the constraint might still work as both, global temperature change and ECS, might similarly reflect the sum of all feedbacks. Emergent constraints of this kind include e.g. the tropical temperature during the Last Glacial Maximum (Hargreaves et al., 2012), tropical temperature anomalies during the mid-Pliocene Warm Period (Hargreaves and Annan, 2016), and post-1970s warming (Jimenez-de-la-Cuesta and Mauritsen, 2019). This seems to be supported by the findings of Tokarska et al. (2020), who tested an emergent constraint for the transient climate response based on recent global warming trends on the CMIP5 and CMIP6 ensembles with similar results for both model ensembles. However, these temperature-based estimates are sensitive to assumptions about forcings and unforced decadal temperature variations, which could also be incorrect, as could model-predicted relationships between feedbacks on short and long time scales that are implicit in most such measures. Indeed, the significance of the COX constraint heavily dropped from CMIP5 to CMIP6 ($p = 0.0032$ to $p = 0.5415$), similar to other constraints in this study.

~~note that also observational uncertainties can play a role as using different observational datasets for a given variable as a proxy for observational uncertainty might lead to different emergent constraints. As this study uses only one combination of observational dataset(s) to calculate the emergent constraints as in the original published emergent constraint studies, the error estimations given by our analysis are expected to underestimate the true error. This could be investigated by systematic tests using different observational datasets and/or combinations of thereof as a proxy for observational uncertainty. Where available, additionally observational uncertainty estimates could be used to give better estimates of the likely constrained range of ECS. A major challenge associated with this is, however, to determine how observational uncertainties propagate to the space-time scales represented by the models because of the typically not well known correlation of observational errors in space and time (e.g. Bellprat et al. (2017)).~~

5 Summary

This paper assesses 11 different emergent constraints on ECS, of which most are directly or indirectly related to cloud feedbacks, by applying them to results from ESMs contributing to CMIP5 and CMIP6. Of particular interest are the results from CMIP6, since all analyzed emergent constraints were published prior to the availability of CMIP6 data. In summary, ~~we assess a range of 2.97 K to 3.88 K for the best estimates of ECS for CMIP5 and a range of 3.41 K to 4.36 K for CMIP6. This increase in~~ the best estimate of ECS ~~can be found for~~averaged over all emergent constraints increases by 12% when moving from relationships trained on CMIP5 to those trained on CMIP6. Some increase is predicted by every constraint ~~that~~ we analyzed, and can be at least partly explained by the increased multi-model mean ECS of CMIP6 which was not accompanied by systematic changes in the constraint variables that could explain this increase – leading to regression fits with higher intercept values at observed constraint values. This is also illustrated by the CMIP5 and CMIP6 multi-model means in the left columns of Figure 2 to Figure 5 (colored dots), in which the connecting line between the CMIP5 and CMIP6 multi-model mean is not parallel to the CMIP5 emergent relationships for all emergent constraints. ~~However, these results need to be treated~~

~~with great care as the analysis showed that all considered emergent relationships are sensitive to outliers and the subset of the climate model ensemble used to fit the emergent relationship. Moreover, our results also show that except for ZHA and SHS~~ Our results also show that except for ZHA, all emergent relationships are weaker (in terms of the coefficient of determination R^2) in CMIP6 compared to CMIP5, which means that the corresponding emergent relationships are able to explain less of the ECS variation simulated by the newer CMIP6 models than by those of CMIP5. This is also demonstrated by the statistical significance, which is lower in CMIP6 than in CMIP5 for all emergent constraint except for ZHA and SHS. As described in section 2.3, our test for statistical significance uses the null hypothesis that there is no correlation between the predictor variable of the emergent constraint and ECS. Further evidence of the decreased performance of the emergent constraints in CMIP6 is given by the size of the constrained ECS ranges, which increases by 51% in CMIP6 compared to CMIP5 on average (using the arithmetic mean of all emergent constraints). Moreover, our more detailed analysis of the ZHA constraint (see Figure 6) showed that this emergent constraint is very sensitive to outliers and the subset of the climate model ensemble used to fit the emergent relationship. Such a behavior might not be unique to the ZHA constraint but could apply to other emergent constraints as well. This in turn suggests that the number of climate models commonly used for emergent constraints might be too low leading to non-robust relationships.

~~Of the 11 emergent constraints analyzed, four are found to be "working" in the sense that they show statistically significant skill on both the CMIP5 and CMIP6 ensembles: BRI, SHL, TIH and VOL. In contrast, the three emergent constraints LIP, SHD and SU are found to be "not working anymore" as their *p*-values are well above 0.1 in CMIP6 (far from significant). COX, TII and SHS are somewhat in-between and could be grouped as "indeterminate" as their *p*-values in CMIP6 dropped from highly or barely significant to almost significant. It is noteworthy that among the group "working" three out of the four emergent constraints point to rather high ECS values of above 4 K in CMIP6, while among the group "not working anymore" two out of three emergent constraints point to rather small ECS values of 3.3 K and less. This might be evidence that emergent constraints on ECS might point to rather higher than lower values in CMIP6.~~

~~Typically, studies proposing a single emergent constraint on ECS do not explicitly take into account model interdependency and all approaches discussed above apply a linear regression of some kind to the model data. This means that it is implicitly assumed that the individual data points (i.e. climate models) are independent. As some modeling groups provide output from multiple ESMs and some ESMs from different modeling groups share components and code, this is clearly not the case. Duplicated code in multiple models is expected to lead to an overestimation of the sample size of a model ensemble and may result in spurious correlations (Sanderson et al., 2015). This limitation also applies to this study as the tests for significance assume that all models are independent. Possible approaches could be to stop treating all models equally by either applying a model weighting based on a model's interdependence with the other models or by simply reducing the ensemble size taking into account models only that are above a given (yet to be defined) interdependence score. Promising approaches to quantify the model interdependency that could be followed include, for example, the studies of Sanderson et al. (2015); Sanderson et al. (2017) and Knutti et al. (2017b).~~

~~A further limitation of this study involves the calculation of significance for the different climate model ensembles: our non-parametric bootstrapping approach calculates the spread of sample values one might obtain if the truth looked like the sample, not the spread of true values consistent with the sample. This is particularly relevant for datasets with a large R^2 and a small sample size, in which case the bootstrap spread will be overconfident. However, since most emergent constraints show small R^2, this effect is expected to be small in our study. Also the calculation of the ECS itself is a source of uncertainty: even though widely used in literature, the Gregory regression method (Gregory et al., 2004) is known to be only an approximation of the true climate sensitivity. A recent paper (Rugenstein et al., 2020) shows that the true equilibrium warming obtained from integrating the climate models until a new equilibrium is reached is 17% (median) higher than the one estimated from the first 150 years of the simulation as done in the Gregory method. However, only a few ESMs provide simulations long enough to assess the true climate sensitivity. The CMIP endorsed LongRunMIP (Rugenstein et al., 2019) could be a promising way to estimate the true climate sensitivity that can then be used to reevaluate emergent constraints and their proposed underlying physical mechanisms.~~

Our analysis makes a number of simplifying assumptions common to other studies, such as model independence, discussed in sections 2.1 and 2.2. These assumptions affect the significance of emergent relationships and the PDFs of ECS based on a constraint. However, they do not affect our main conclusions here, which concern the change in performance on CMIP6 relative to CMIP5 and the implications for robustness and future use of emergent constraints.

ECS is the product of the complex interactions of the many components~~.~~ and feedbacks. Thus, constraining ECS with a single physical process might overly simplify this problem. Such single process-oriented emergent constraints therefore do not seem to be helpful in constraining ECS but should probably rather be thought of as constraints for the process or feedback they are actually targeting (if that can be clearly identified). With increasing computational resources available to climate science, more and more detailed ~~of these~~process interactions can be taken into account in a modern ESM. In contrast, the predecessor versions CMIP3 and CMIP5 were less complex with ~~fewer components~~simpler atmospheric process representations, so constraining uncertainties of a single dominant process may have allowed for ~~a~~an apparently more successful ~~constraining~~constraint of ECS than would be achieved in more complex models. As a conclusion, we argue that to constrain ECS ~~for the latest generation of climate models~~in a more robust way, it might be beneficial to apply multivariate approaches that are able to consider multiple (different) relevant physical processes ~~at once and thus are able to get a broader picture of the complex reality. New~~and feedbacks at once and thus are able to get a broader picture of the complex reality. A possible approach for this is given by Bretherton and Caldwell (2020), who combine the information from multiple emergent constraints on ECS using a multivariate Gaussian and multiple linear regression. For the CMIP3 and CMIP5 ensembles, they find an increased best estimate relative to the unweighted ensemble mean similar to the participating individual emergent constraints, but with lower uncertainty range. Moreover, new machine learning techniques are a promising avenue forward for such multivariate approaches and for constraining uncertainties in multi-model projections ~~(Schlund et al., in review)~~(Schlund et al., accepted) with the aim of further improving climate modelling and analysis (Reichstein et al., 2019).

6 Tables

Label	Reference	Short description of x-axis	Variables	Observations
BRI	(Brient and Schneider, 2016)	Response of shortwave cloud reflectivity to changes in sea surface temperature [% K^{-1}]	• Surface temperature (*ts*) • Relative humidity (*hur*) • Top-of-atmosphere (TOA) outgoing shortwave radiation (*rsut*) • TOA outgoing shortwave radiation assuming clear sky (*rsutcs*) • TOA incoming shortwave radiation (*rsdt*)	HadISST (*tos*) (Rayner et al., 2003), ERA-Interim (*hur*) (Dee et al., 2011), CERES-EBAF (*rsut*, *rsutcs*, *rsdt*) (Loeb et al., 2018) [2001-2005]
COX	(Cox et al., 2018)	Temperature variability metric [K]	• Surface air temperature (*tas*)	HadCRUT4 (Morice et al., 2012) [1880-2014]
LIP	(Lipat et al., 2017)	Southern hemisphere Hadley cell extent [°]	• Northward wind (*va*)	ERA-Interim (Dee et al., 2011) [1980-2005]
SHD	(Sherwood et al., 2014)	*D* index (large-scale lower-tropospheric mixing) [1]	• Vertical velocity (*wap*)	ERA-Interim (Dee et al., 2011) [1989-1998]
SHL	(Sherwood et al., 2014)	Lower tropospheric mixing index (LTMI) [1]	• Relative humidity (*hur*) • Air temperature (*ta*) • Vertical velocity (*wap*)	ERA-Interim (Dee et al., 2011) [1989-1998]
SHS	(Sherwood et al., 2014)	*S* index (small-scale lower-tropospheric mixing) [1]	• Relative humidity (*hur*) • Air temperature (*ta*) • Vertical velocity (*wap*)	ERA-Interim (Dee et al., 2011) [1989-1998]
SU	(Su et al., 2014)	Error in vertical profile of relative humidity [1]	• Relative humidity (*hur*)	AIRS (below 300hPa) (Aumann et al., 2003), MLS-Aura (above

Label	Reference	Short description of x-axis	Variables	Observations
				300hPa) (Beer, 2006) [2005-2010]
TIH	(Tian, 2015)	Tropical mid-tropospheric humidity index [%]	• Specific humidity (*hus*)	AIRS (Aumann et al., 2003) [2003-2005]
TII	(Tian, 2015)	Southern ITCZ index [mm day^{-1}]	• Precipitation (*pr*)	GPCP (Adler et al., 2003) [1986-2005]
VOL	(Volodin, 2008)	Difference in total cloud fraction between tropics (28°S – 28°N) and southern mid-latitudes (56°S – 36°S) [%]	• Total cloud area fraction (*clt*)	ISCCP D-2 (Rossow and Schiffer, 1991) [1980-2000]*
ZHA	(Zhai et al., 2015)	Response of seasonal marine boundary layer cloud fraction to change in sea surface temperature [% K^{-1}]	• Cloud area fraction (*cl*) • Sea surface temperature (*tos*) • Vertical velocity (*wap*)	CloudSat/CALIPSO (Mace et al., 2009), AMSRE SST (AMSR-E, 2011), ERA-Interim (Dee et al., 2011) [1980-2004]*

Table 1: Overview of the 11 emergent constraints on the ECS used in this study. Observations marked with an asterisk (*) are identical with the ones used in original publication.

Index used in plots	Model	Reference
1	ACCESS1-0	(Dix et al., 2013)
2	ACCESS1-3	(Dix et al., 2013)
3	BNU-ESM	(Ji et al., 2014)
4	CCSM4	(Gent et al., 2011; Meehl et al., 2012)
5	CNRM-CM5	(Voldoire et al., 2013)
6	CNRM-CM5-2	(Voldoire et al., 2013)
7	CSIRO-Mk3-6-0	(Rotstayn et al., 2012)
8	CanESM2	(Arora et al., 2011)
9	FGOALS-g2	(Li et al., 2013)
10	GFDL-CM3	(Donner et al., 2011)
11	GFDL-ESM2G	(Dunne et al., 2012)
12	GFDL-ESM2M	(Dunne et al., 2012)
13	GISS-E2-H	(Schmidt et al., 2006)
14	GISS-E2-R	(Schmidt et al., 2006)
15	HadGEM2-ES	(Collins et al., 2011)
16	IPSL-CM5A-LR	(Dufresne et al., 2013)
17	IPSL-CM5A-MR	(Dufresne et al., 2013)
18	IPSL-CM5B-LR	(Dufresne et al., 2013)
19	MIROC-ESM	(Watanabe et al., 2011)
20	MIROC5	(Watanabe et al., 2010)
21	MPI-ESM-LR	(Giorgetta et al., 2013)
22	MPI-ESM-MR	(Giorgetta et al., 2013)
23	MPI-ESM-P	(Giorgetta et al., 2013)
24	MRI-CGCM3	(Yukimoto et al., 2012)
25	NorESM1-M	(Bentsen et al., 2013; Iversen et al., 2013)
26	bcc-csm1-1	(Wu et al., 2014)
27	bcc-csm1-1-m	(Wu et al., 2014)
28	inmcm4	(Volodin et al., 2010)

Table 2: List of CMIP5 models alongside the index used in the figures of this study and a reference.

Index used in plots	Model	Reference
29	ACCESS-CM2	(Bi et al., 2013)
30	ACCESS-ESM1-5	(Law et al., 2017; Ziehn et al., 2017)
~~29~~31	AWI-CM-1-1-MR	(Rackow et al., 2018; Sidorenko et al., 2015)
~~30~~32	BCC-CSM2-MR	(Wu et al., 2019)
~~31~~33	BCC-ESM1	(Wu et al., 2019)
~~32~~34	CAMS-CSM1-0	(Rong et al., 2018)
35	CAS-ESM2-0	(Wang et al., 2020)
~~33~~36	CESM2	(~~Gettelman~~Danabasoglu et al., ~~2019b~~2020)
37	CESM2-FV2	(Danabasoglu et al., 2020)
~~34~~38	CESM2-WACCM	(Danabasoglu et al., 2020; Gettelman et al., 2019b)
39	CESM2-WACCM-FV2	(Danabasoglu et al., 2020; Gettelman et al., 2019b)
40	CMCC-CM2-SR5	(Cherchi et al., 2019)
~~35~~41	CNRM-CM6-1	(Voldoire et al., 2019)
~~36~~42	CNRM-CM6-1-HR	(Voldoire et al., 2019)
~~37~~43	CNRM-ESM2-1	(Séférian et al., 2019)
~~38~~44	CanESM5	(Swart et al., 2019)
~~39~~45	E3SM-1-0	(Golaz et al., 2019)
~~40~~46	EC-Earth3-Veg	~~(Wyser et al., 2019)~~(Wyser et al., 2020)
~~41~~47	FGOALS-f3-L	(Guo et al., 2020; He et al., 2019; He et al., 2020)
~~42~~48	~~GFDL-CM4~~FGOALS-g3	~~(Held et al., 2019)~~(Li et al., 2020)
~~43~~	~~GFDL-ESM4~~	~~(Held et al., 2019)~~
~~44~~49	GISS-E2-1-G	(Rind et al., 2020)
~~45~~50	GISS-E2-1-H	(Rind et al., 2020)
~~46~~51	HadGEM3-GC31-LL	(Kuhlbrodt et al., 2018)
52	HadGEM3-GC31-MM	(Williams et al., 2018)
~~47~~53	INM-CM4-8	(Volodin et al., 2017a; Volodin et al., 2017b)
~~48~~54	INM-CM5-0	(Volodin et al., 2017a; Volodin et al., 2017b)
~~49~~55	IPSL-CM6A-LR	(Boucher et al~~.)~~., 2020)
56	KACE-1-0-G	(Lee et al., 2020a)
~~50~~57	MCM-UA-1-0	(Delworth et al., 2002)

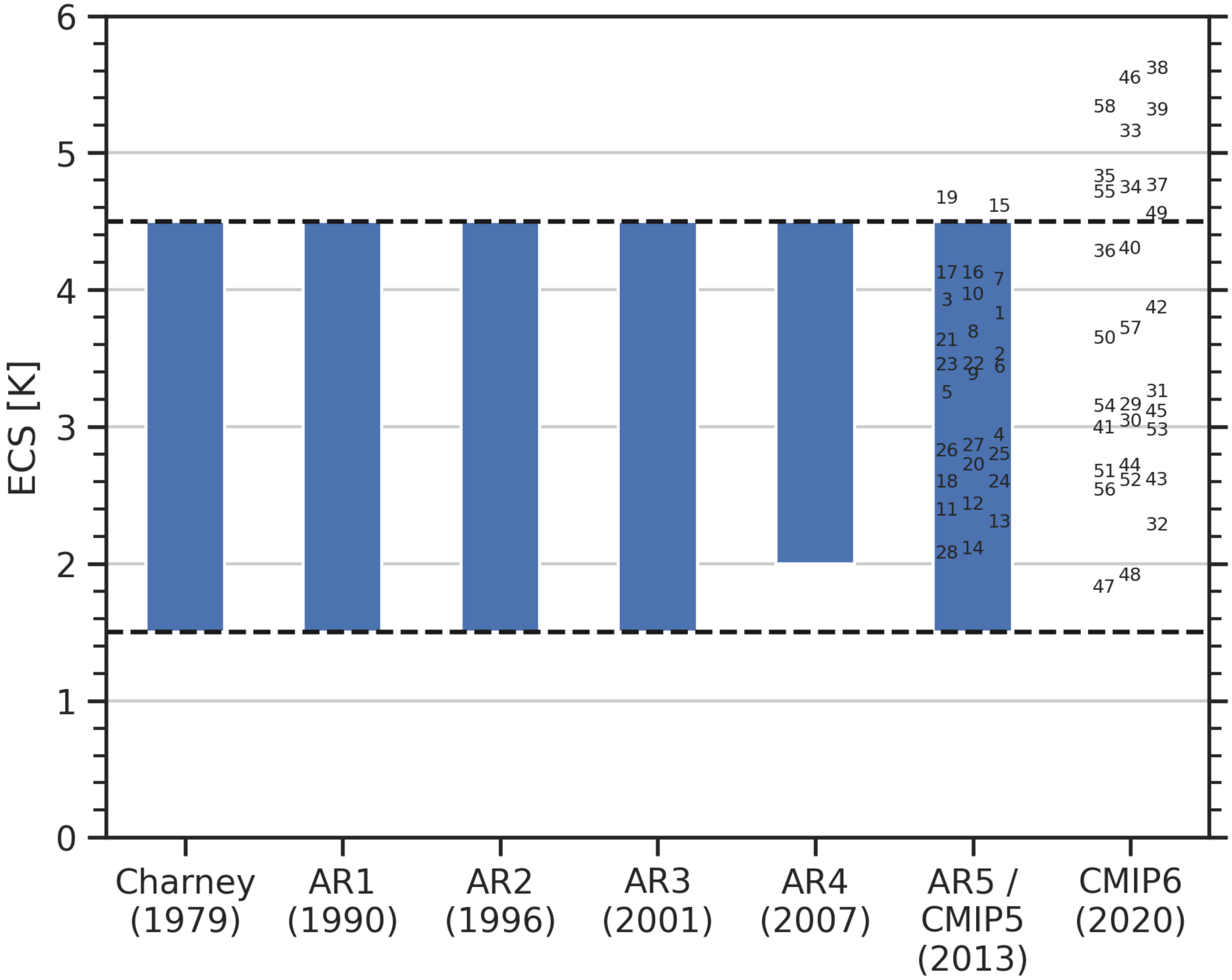
Index used in plots	**Model**	**Reference**
~~51~~58	MIROC-ES2L	(Hajima et al., 2020)
~~52~~59	MIROC6	(Tatebe et al., 2019)
60	MPI-ESM-1-2-HAM	(Mauritsen et al., 2019)
~~53~~61	MPI-ESM1-2-HR	(Muller et al., 2018)
62	MPI-ESM1-2-LR	(Mauritsen et al., 2019)
~~54~~63	MRI-ESM2-0	(Yukimoto et al., 2019)
~~55~~64	NESM3	(Cao et al., 2018)
65	NorCPM1	(Counillon et al., 2016)
~~56~~66	NorESM2-LM	(Seland et al., 2020)
67	NorESM2-MM	(Seland et al., 2020)
~~57~~68	SAM0-UNICON	(Park et al., 2019)
69	TaiESM1	(Lee et al., 2020b)
~~58~~70	UKESM1-0-LL	(Sellar et al., 2019)

Table 3: As Table 2 but for CMIP6 models included in this study.

Label	ECS (original publication)	ECS (CMIP5) [K]	ECS (CMIP6) [K]	*p* (CMIP5)	*p* (CMIP6)
BRI	most likely 4.0 K, < 2.30 K very unlikely (90% confidence)	3.72 ± 0.~~56~~59	4.~~36~~32 ± 1.~~16~~07	0.~~0002~~0005	0.~~0219~~0355
COX	2.8 K ± 0.6 K	3.03 ± 0.~~71~~73	3.~~44~~71 ± 1.~~15~~09	0.~~0010~~0032	0.~~0545~~5415
LIP	no best estimate given	2.97 ± 0.~~76~~75	3.~~66~~75 ± 1.~~27~~11	0.~~0043~~0228	0.~~2039~~6791
SHD	none – see SHL	3.65 ± 0.~~63~~64	3.~~74~~77 ± 1.~~11~~06	0.~~0006~~0037	0.~~1120~~2805
SHL	most likely 4 K with lower limit 3 K	3.42 ± 0.~~63~~65	3.~~65~~67 ± 1.~~05~~06	0.~~0001~~0002	0.~~0114~~0138
SHS	none – see SHL	3.07 ± 0.~~68~~73	3.~~41~~48 ± 1.~~13~~07	0.~~0581~~0647	0.~~0638~~0396
SU	most likely 4 K with lower limit 3 K	3.30 ± 0.~~90~~88	3.~~69~~77 ± 1.~~59~~35	0.~~0919~~1676	0.~~8573~~1935
TIH	most likely 4.0 K	3.88 ± 0.~~78~~75	4.~~07~~15 ± 1.~~21~~10	0.~~0002~~0089	0.~~0454~~1348
TII	most likely 4.0 K	3.87 ± 0.~~70~~67	~~4.11~~3.84 ± 1.~~17~~09	0.~~0004~~0013	0.~~0634~~8236
VOL	3.6 K ± 0.4 K (standard deviation)	3.74 ± 0.64	4.~~14~~21 ± 1.~~13~~04	0.0004	0.~~0057~~0056
ZHA	3.90 K ± 0.45 K (standard deviation)	3.35 ± 0.~~72~~74	3.~~81~~79 ± 0.~~60~~67	0.~~1195~~2567	< 0.~~0022~~0001

Table 4: Overview of the constrained ECS ranges and p-values for all 11 analyzed emergent constraints. If not further specified, the uncertainty ranges correspond to the ~~IPCC's "likely range" (~~66% confidence~~).~~ intervals (17–83%). For CMIP5 and CMIP6, these are evaluated from the probability distribution given by equation (6) (see also ~~middle~~right columns of Figure 2 to Figure 5). Note that even though CMIP5 models were used for some constraints in the original publications, the constrained ranges in column 2 and column 3 might differ due to the use of different models (in this paper, we use output from all CMIP models that is publicly available, see data availability section for details). The *p*-values describing the significance of the emergent relationships are defined as the probability ~~that *r* (given by the bootstrapped distribution) exhibits the opposite sign as originally expected from the emergent relationship (see shaded areas in the right columns of Figure 2 to Figure 5).~~to obtain an absolute correlation coefficient |*r*| or higher under the null hypothesis that the true underlying correlation coefficient between the predictor and ECS is zero. Smaller p-values point to higher significance and vice versa (for details see section 2.3).

7 Figures



ECS [K]
0
1
2
3
4
5
6
Charney (1979)
AR1 (1990)
AR2 (1996)
AR3 (2001)
AR4 (2007)
AR5 / CMIP5 (2013)
CMIP6 (2020)

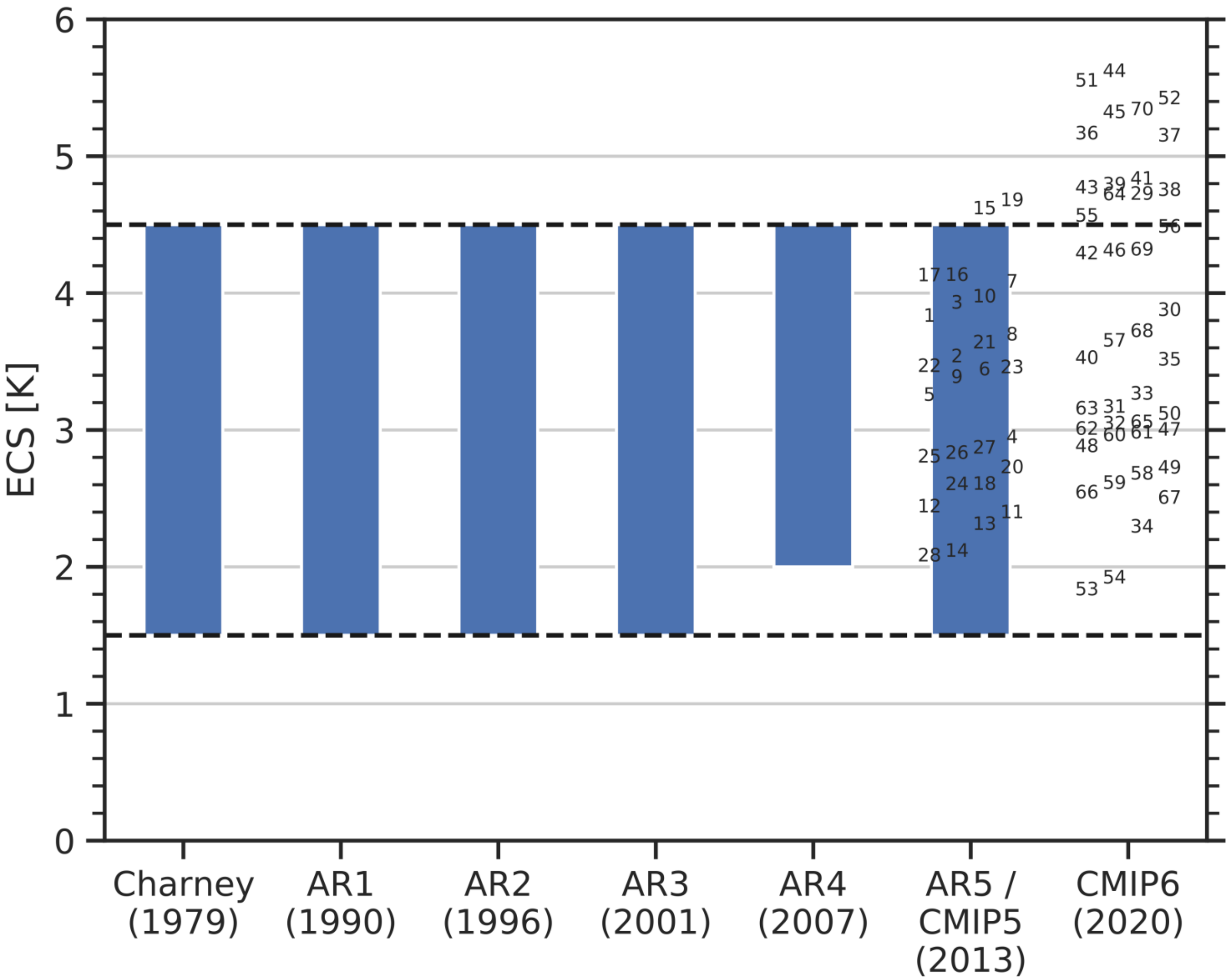
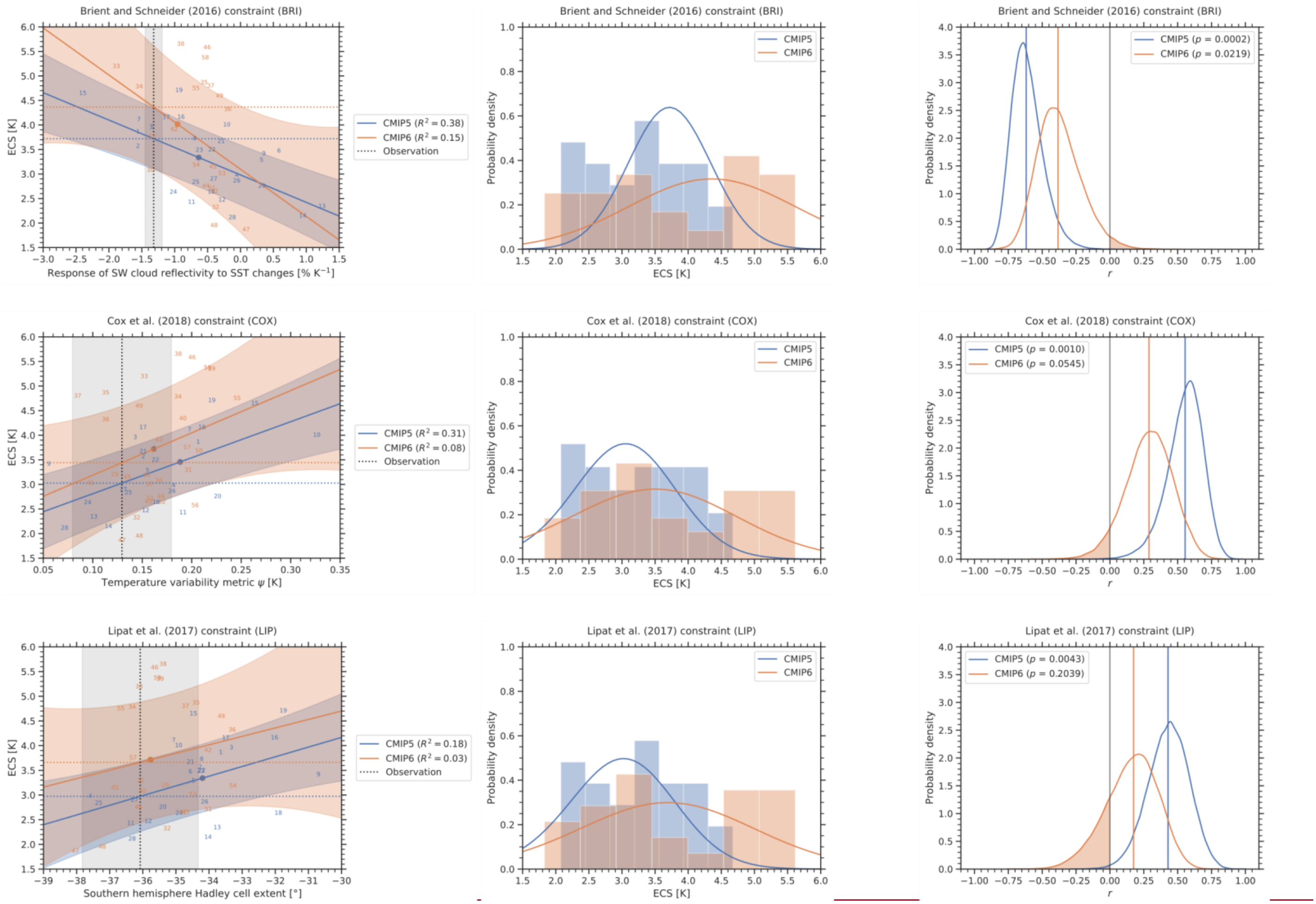


Figure 1: Assessed ECS ranges (blue bars) from the Charney report (Charney et al., 1979) and the different Assessment Reports (ARs) of the Intergovernmental Panel on Climate Change (IPCC). The numbers correspond to individual CMIP5 and CMIP6 models; see Table A1 and Table A2 for details. Adapted and updated from Meehl et al. (2020).


Brient and Schneider (2016) constraint (BRI)
ECS [K]
Response of SW cloud reflectivity to SST changes [% K^{-1}]
CMIP5 ($R^2 = 0.38$)
CMIP6 ($R^2 = 0.15$)
Observation
Probability density
ECS [K]
CMIP5
CMIP6
CMIP5 ($p = 0.0002$)
CMIP6 ($p = 0.0219$)
r
Cox et al. (2018) constraint (COX)
Temperature variability metric ψ [K]
CMIP5 ($R^2 = 0.31$)
CMIP6 ($R^2 = 0.08$)
Observation
CMIP5 ($p = 0.0010$)
CMIP6 ($p = 0.0545$)
Lipat et al. (2017) constraint (LIP)
Southern hemisphere Hadley cell extent [°]
CMIP5 ($R^2 = 0.18$)
CMIP6 ($R^2 = 0.03$)
Observation
CMIP5 ($p = 0.0043$)
CMIP6 ($p = 0.2039$)

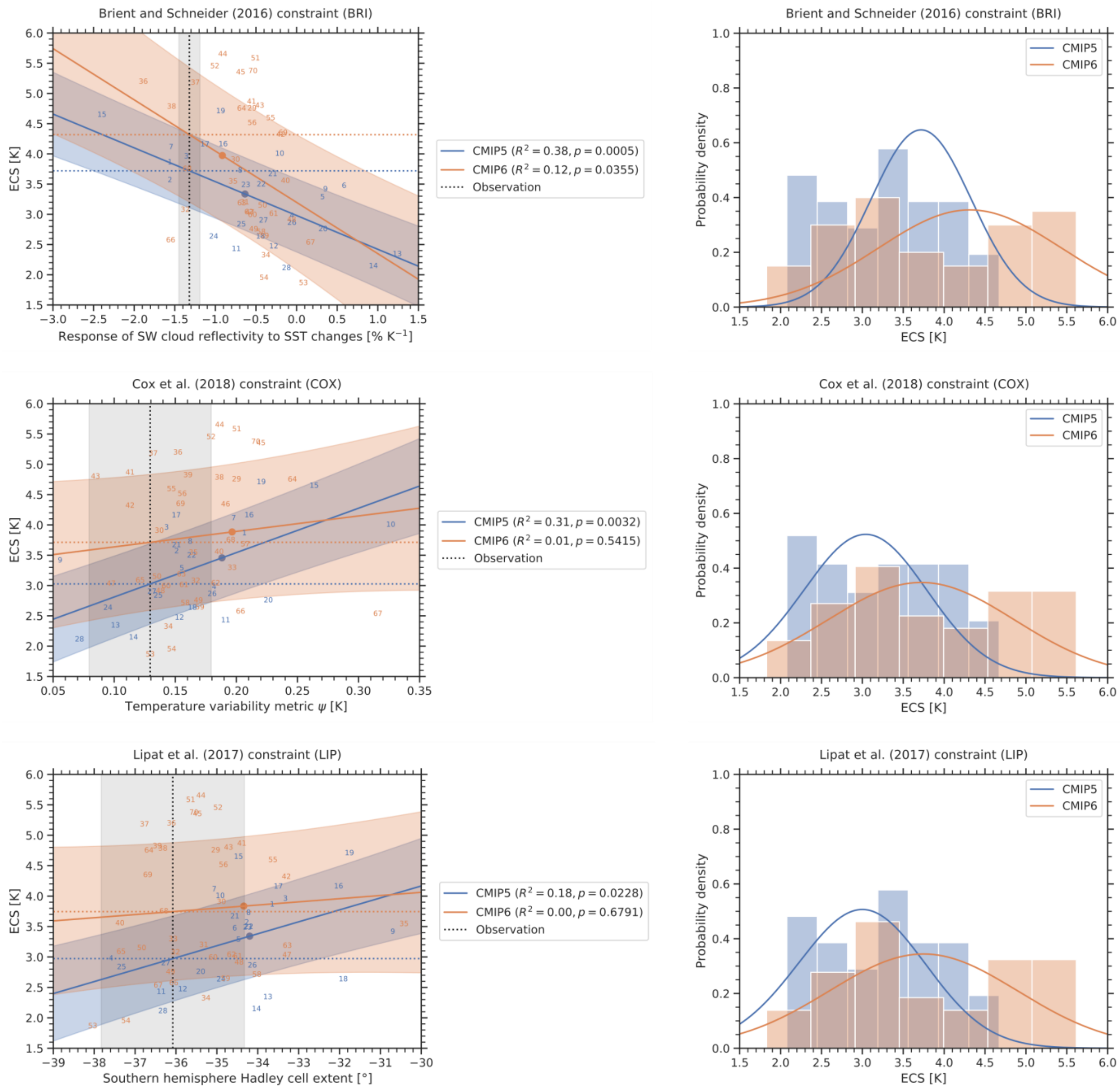
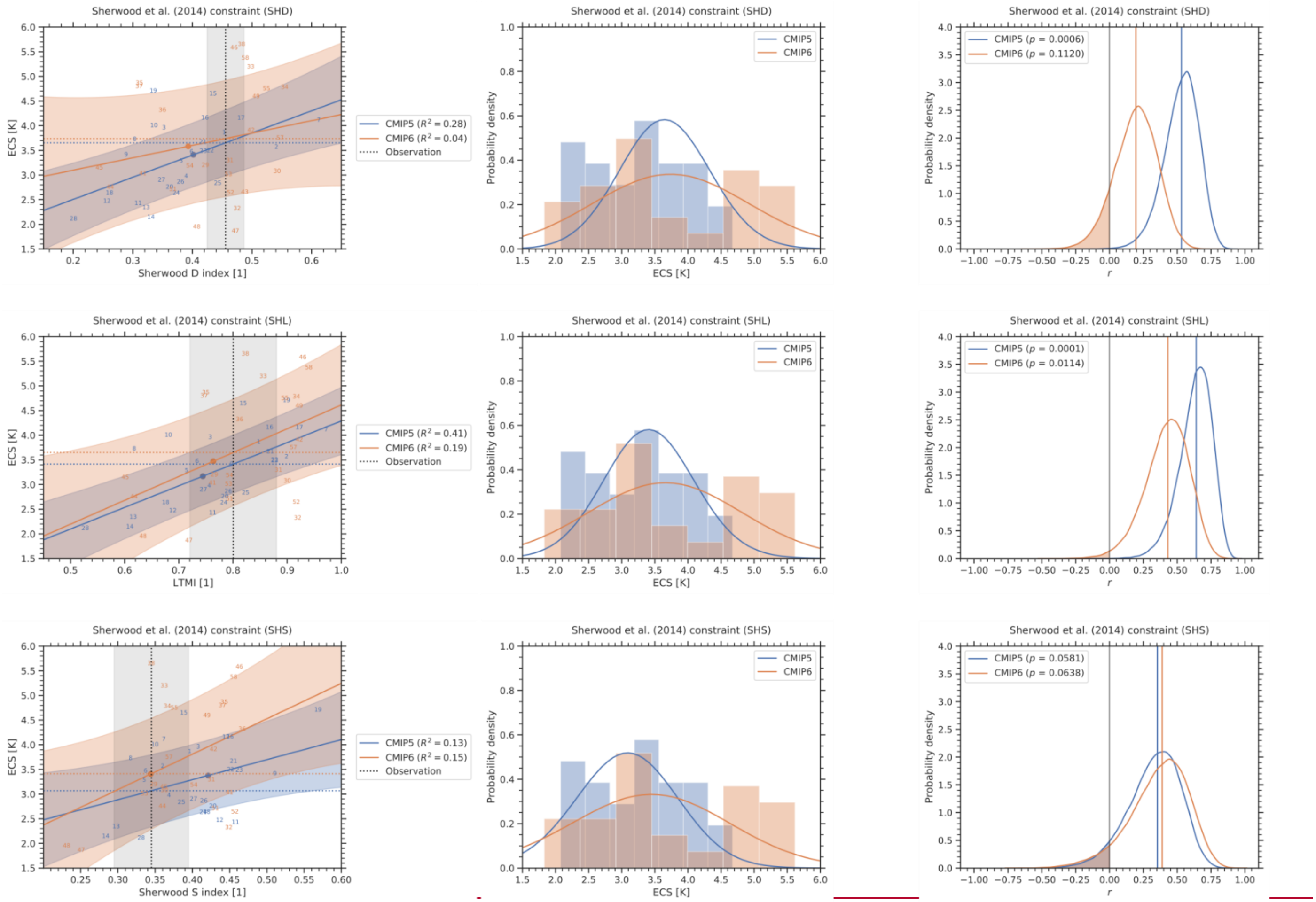


Figure 2: Emergent constraints BRI, COX and LIP applied to the CMIP5 ensemble (blue) and CMIP6 ensemble (orange). Left column: emergent relationships (solid blue and orange lines) for the CMIP models (numbers for individual models are specified in Table A1 and Table A2). The shaded areas around the regression lines correspond to the standard prediction errors (equation (3)), which defines the error in the regression model itself. The vertical black dashed line corresponds to the observational reference (see Table 1 for details on the individual observational datasets used) with its uncertainty range given as standard error (gray shaded area). The horizontal dashed lines show the best estimates of the constrained ECS for CMIP5 (blue) and CMIP6 (orange). The colored dots mark the CMIP5 (blue) and CMIP6 (orange) multi-model means. ~~Middle~~The p-value in the legend corresponds to the hypothesis test introduced in section 2.3 and describes the probability to obtain an absolute correlation coefficient $|r|$ or higher under the null hypothesis that the true underlying correlation coefficient between the predictor and ECS is zero. Right column: probability densities for the constrained ECS following equation (6) (solid lines) and the unconstrained model ensembles (histograms). Note that for each individual emergent constraint, a different subset of climate models is used due to the availability of data (see Table A1 and Table A2 for details). Thus, these histograms may differ for the different constraints. ~~Right column: probability distributions of the linear (Pearson) correlation coefficient r for 100,000 bootstrap samples (solid lines). The vertical solid lines correspond to the value of r when the complete original sample is used for its calculation. The vertical gray line illustrates $r = 0$, i.e. no correlation between the emergent constraint x-axis and ECS. The shaded areas correspond to the p-values describing~~

~~the significance of the emergent relationships. These are defined as the probability that r (given by the bootstrapped distribution) exhibits the opposite sign as originally expected from the emergent relationship (see section 2.3). Smaller p-values point to higher significance and vice versa.~~


Sherwood et al. (2014) constraint (SHD)
Sherwood D index [1]
ECS [K]
CMIP5 ($R^2 = 0.28$)
CMIP6 ($R^2 = 0.04$)
Observation
Probability density
CMIP5
CMIP6
CMIP5 ($p = 0.0006$)
CMIP6 ($p = 0.1120$)
r
Sherwood et al. (2014) constraint (SHL)
LTMI [1]
CMIP5 ($R^2 = 0.41$)
CMIP6 ($R^2 = 0.19$)
CMIP5 ($p = 0.0001$)
CMIP6 ($p = 0.0114$)
Sherwood et al. (2014) constraint (SHS)
Sherwood S index [1]
CMIP5 ($R^2 = 0.13$)
CMIP6 ($R^2 = 0.15$)
CMIP5 ($p = 0.0581$)
CMIP6 ($p = 0.0638$)

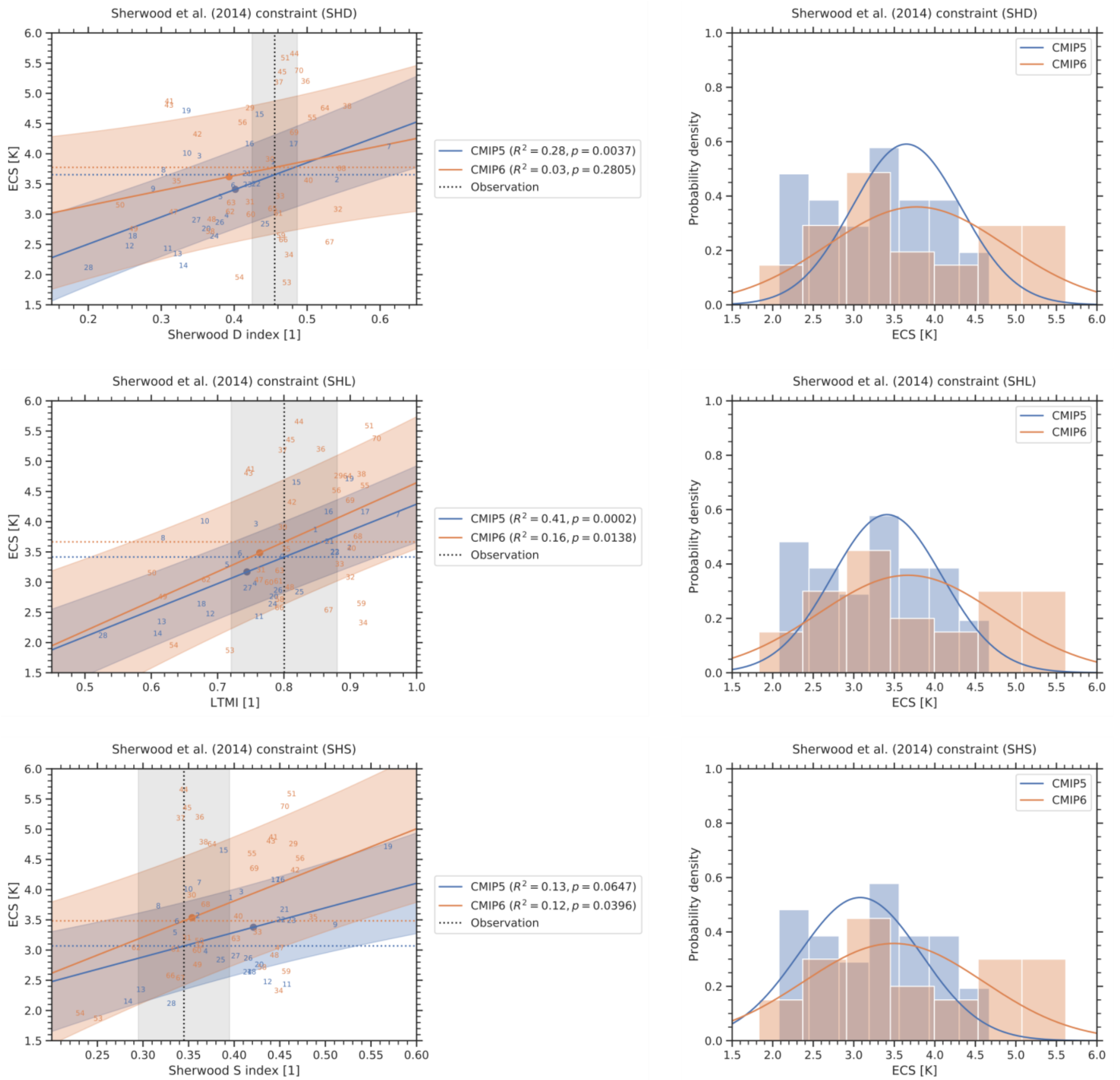
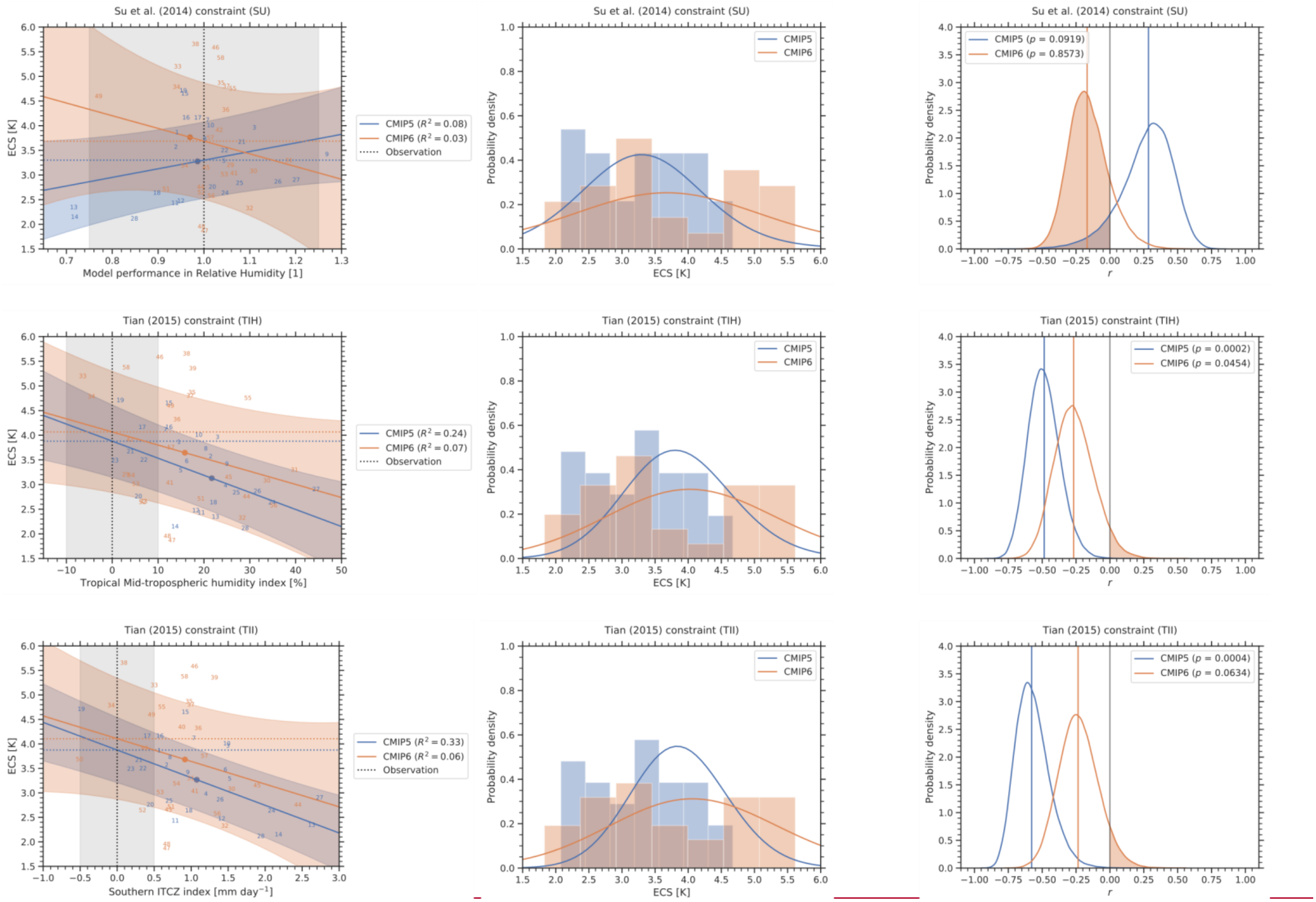


Figure 3: As Figure 2, but for the constraints SHD, SHL and SHS.


Su et al. (2014) constraint (SU)
ECS [K]
Model performance in Relative Humidity [1]
CMIP5 ($R^2 = 0.08$)
CMIP6 ($R^2 = 0.03$)
Observation
Probability density
ECS [K]
CMIP5
CMIP6
CMIP5 ($p = 0.0919$)
CMIP6 ($p = 0.8573$)
r
Tian (2015) constraint (TIH)
Tropical Mid-tropospheric humidity index [%]
CMIP5 ($R^2 = 0.24$)
CMIP6 ($R^2 = 0.07$)
Observation
CMIP5 ($p = 0.0002$)
CMIP6 ($p = 0.0454$)
Tian (2015) constraint (TII)
Southern ITCZ index [mm day^{-1}]
CMIP5 ($R^2 = 0.33$)
CMIP6 ($R^2 = 0.06$)
Observation
CMIP5 ($p = 0.0004$)
CMIP6 ($p = 0.0634$)

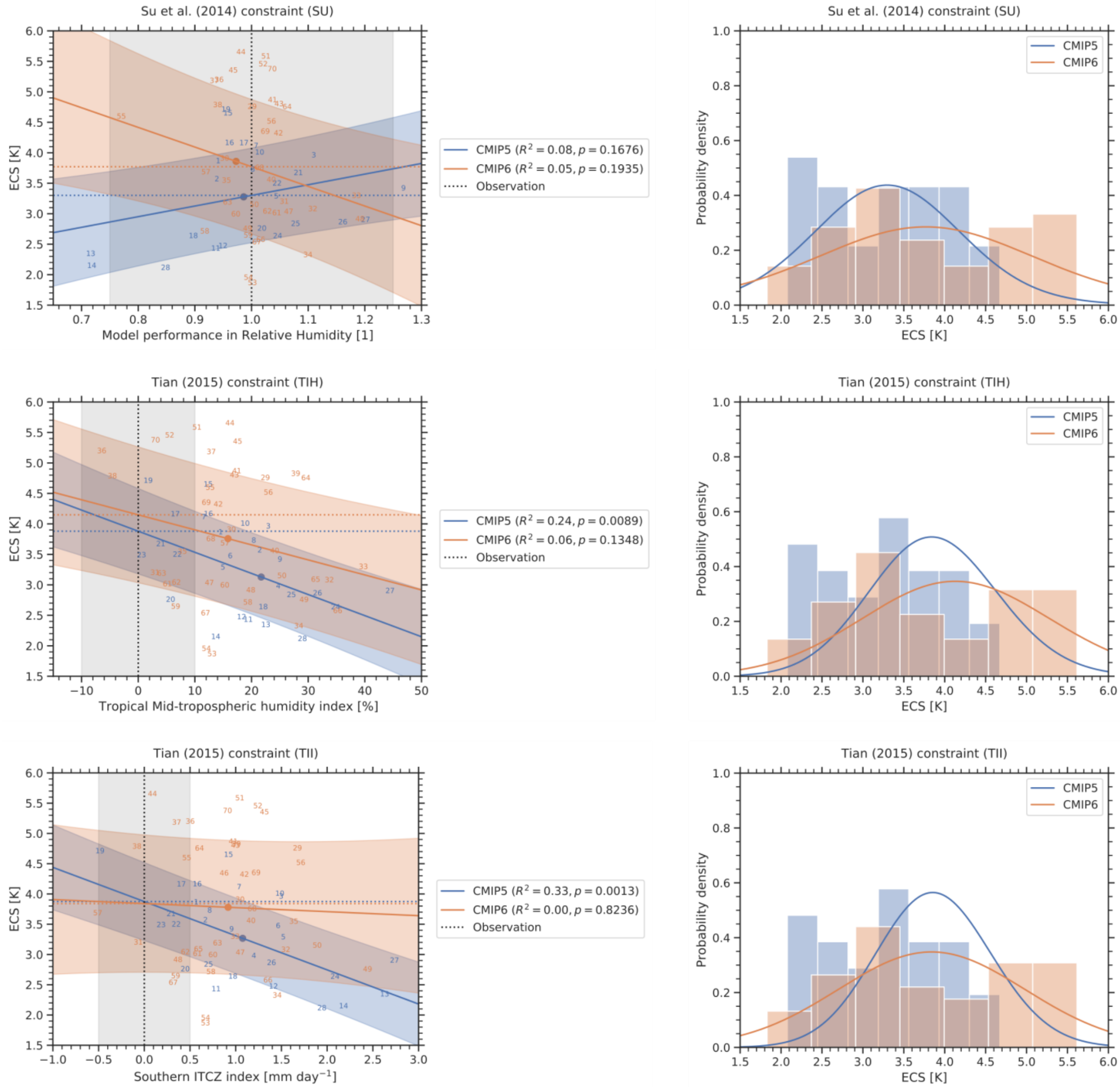
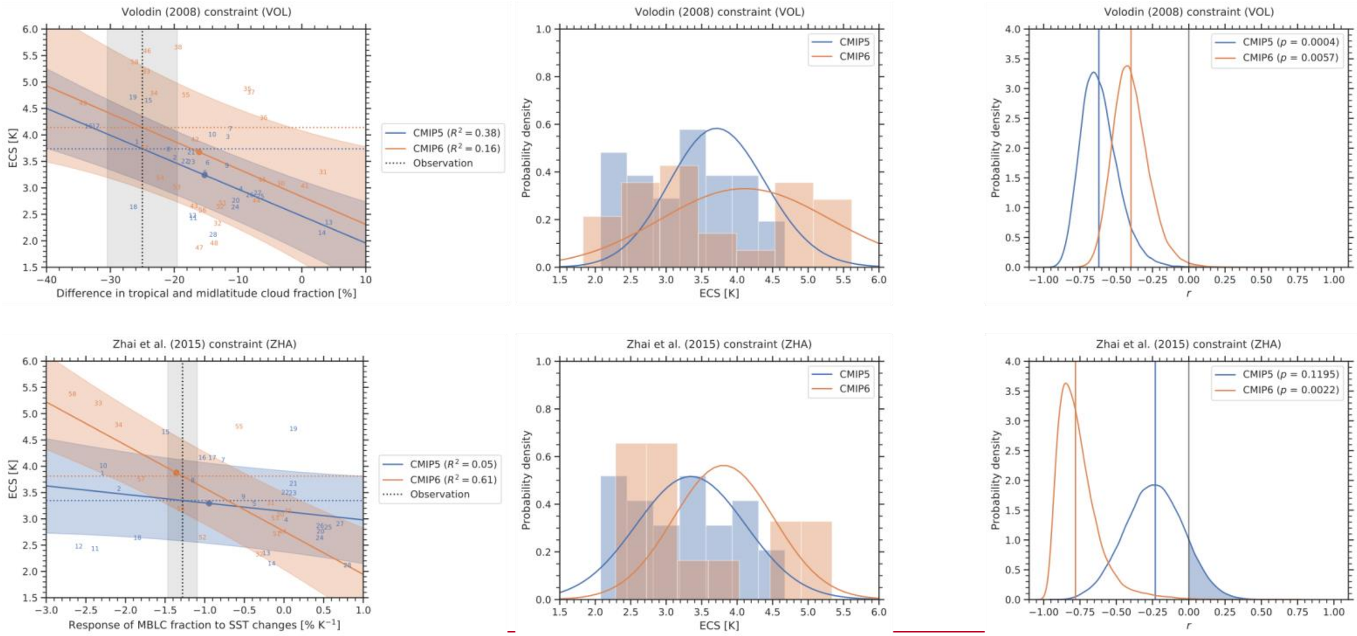


Figure 4: As Figure 2, but for the constraints SU, TIH and TII.


Volodin (2008) constraint (VOL)
Difference in tropical and midlatitude cloud fraction [%]
ECS [K]
CMIP5 ($R^2 = 0.38$)
CMIP6 ($R^2 = 0.16$)
Observation
Probability density
CMIP5
CMIP6
CMIP5 ($p = 0.0004$)
CMIP6 ($p = 0.0057$)
r
Zhai et al. (2015) constraint (ZHA)
Response of MBLC fraction to SST changes [% K^{-1}]
CMIP5 ($R^2 = 0.05$)
CMIP6 ($R^2 = 0.61$)
CMIP5 ($p = 0.1195$)
CMIP6 ($p = 0.0022$)

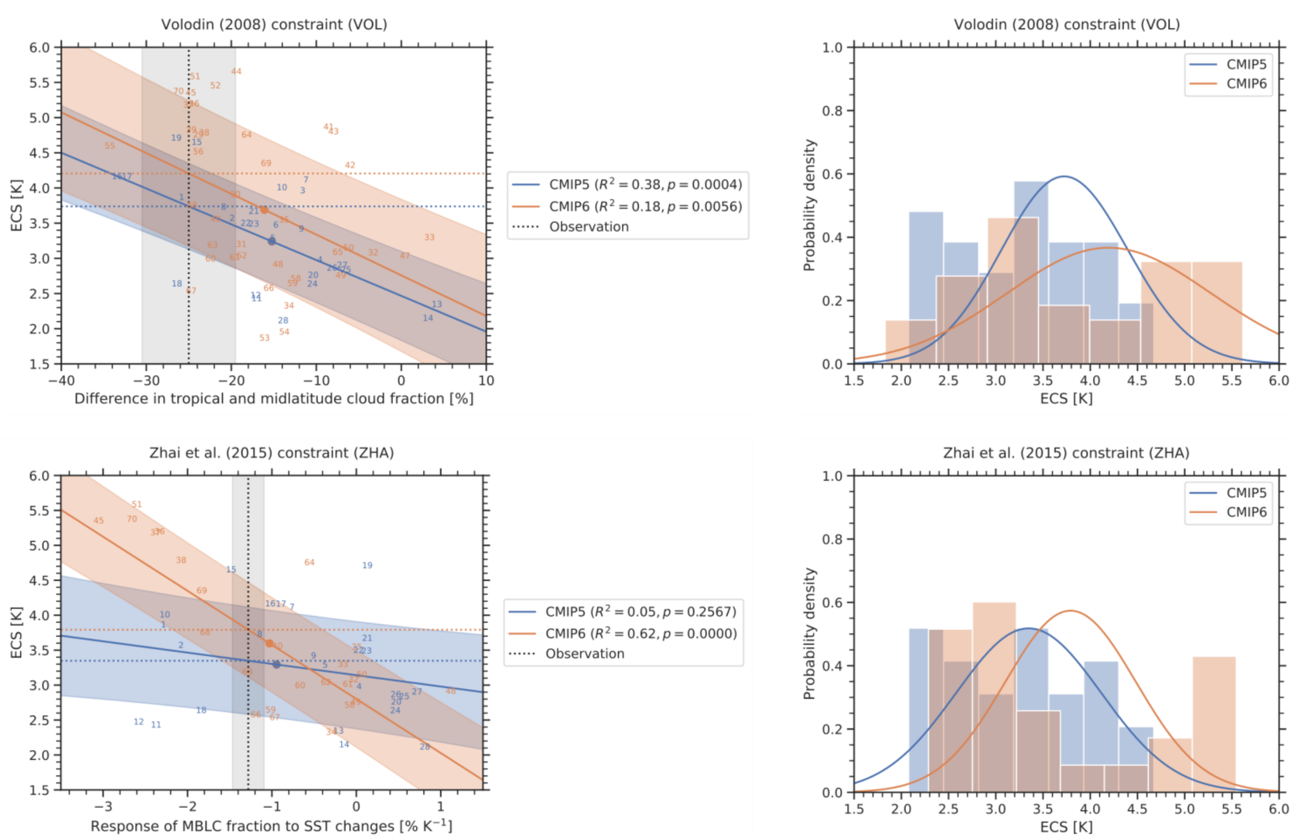
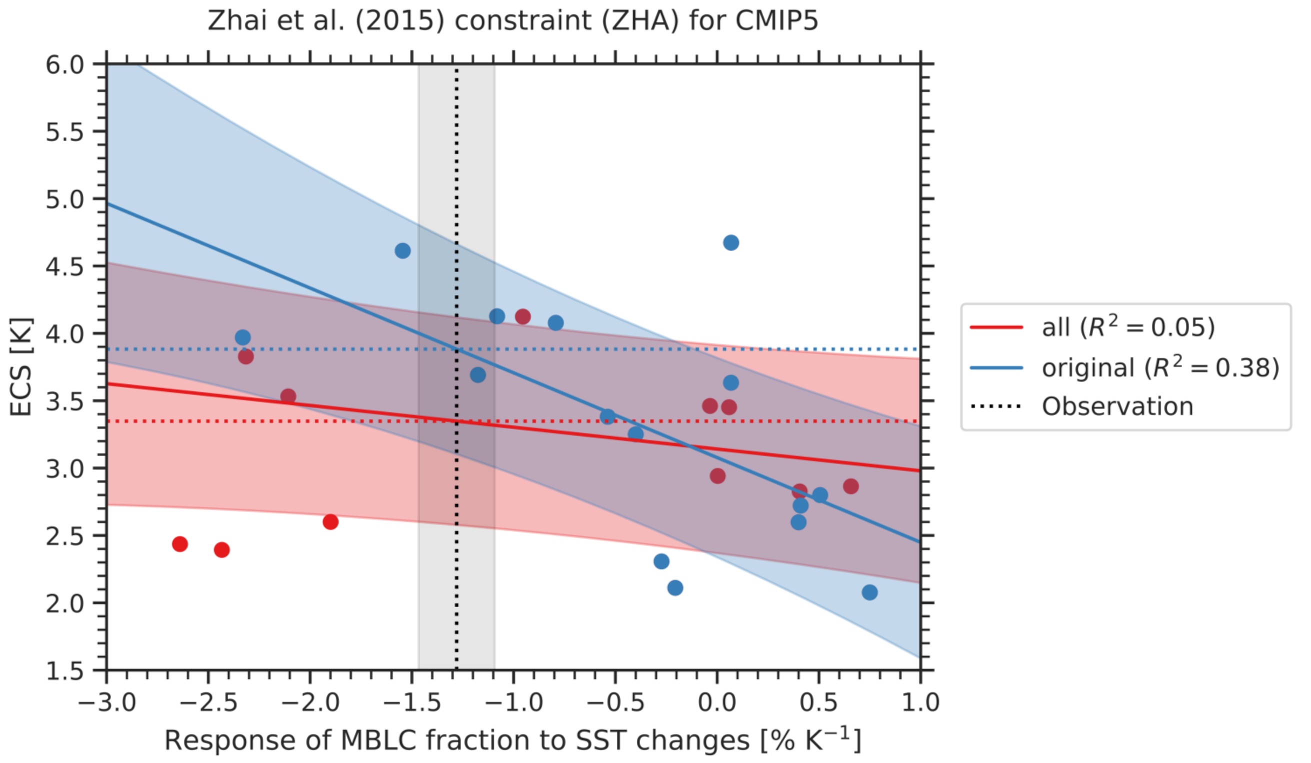

Volodin (2008) constraint (VOL)
Difference in tropical and midlatitude cloud fraction [%]
ECS [K]
CMIP5 ($R^2 = 0.38, p = 0.0004$)
CMIP6 ($R^2 = 0.18, p = 0.0056$)
Observation
Probability density
CMIP5
CMIP6
Zhai et al. (2015) constraint (ZHA)
Response of MBLC fraction to SST changes [% K^{-1}]
CMIP5 ($R^2 = 0.05, p = 0.2567$)
CMIP6 ($R^2 = 0.62, p = 0.0000$)

Figure 5: As Figure 2, but for the constraints VOL and ZHA.


Zhai et al. (2015) constraint (ZHA) for CMIP5
ECS [K]
Response of MBLC fraction to SST changes [% K^{-1}]
all ($R^2 = 0.05$)
original ($R^2 = 0.38$)
Observation

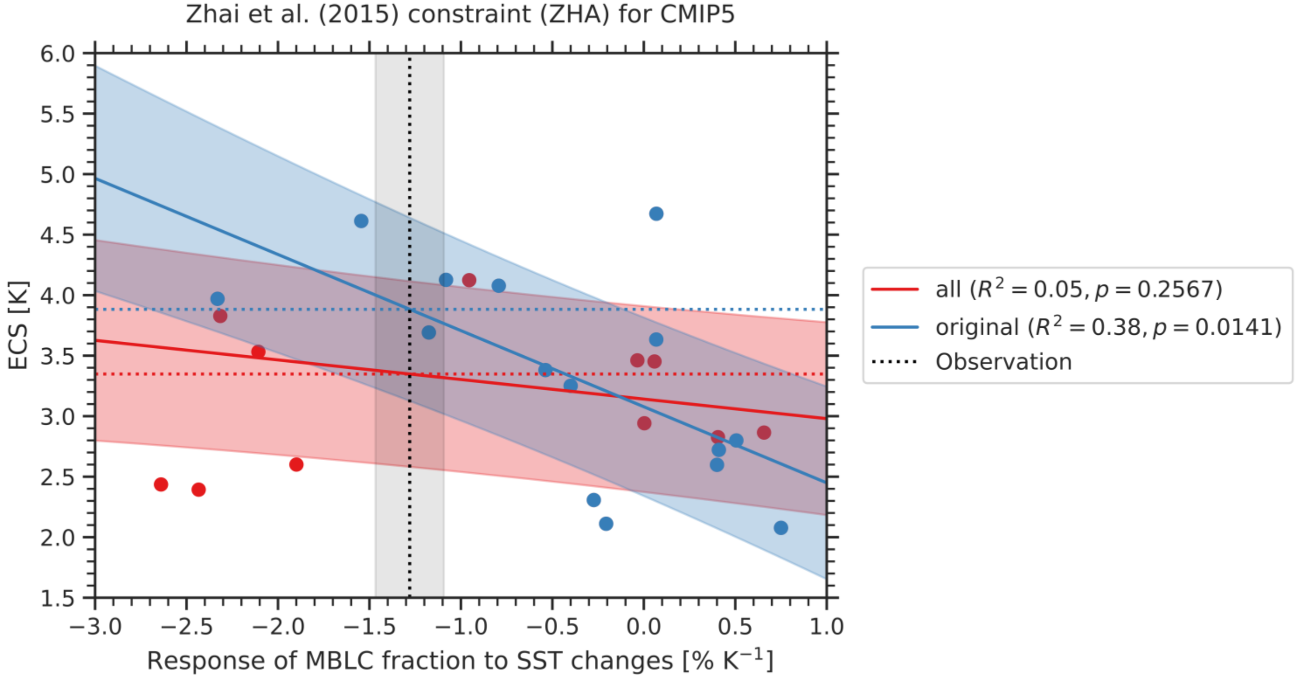
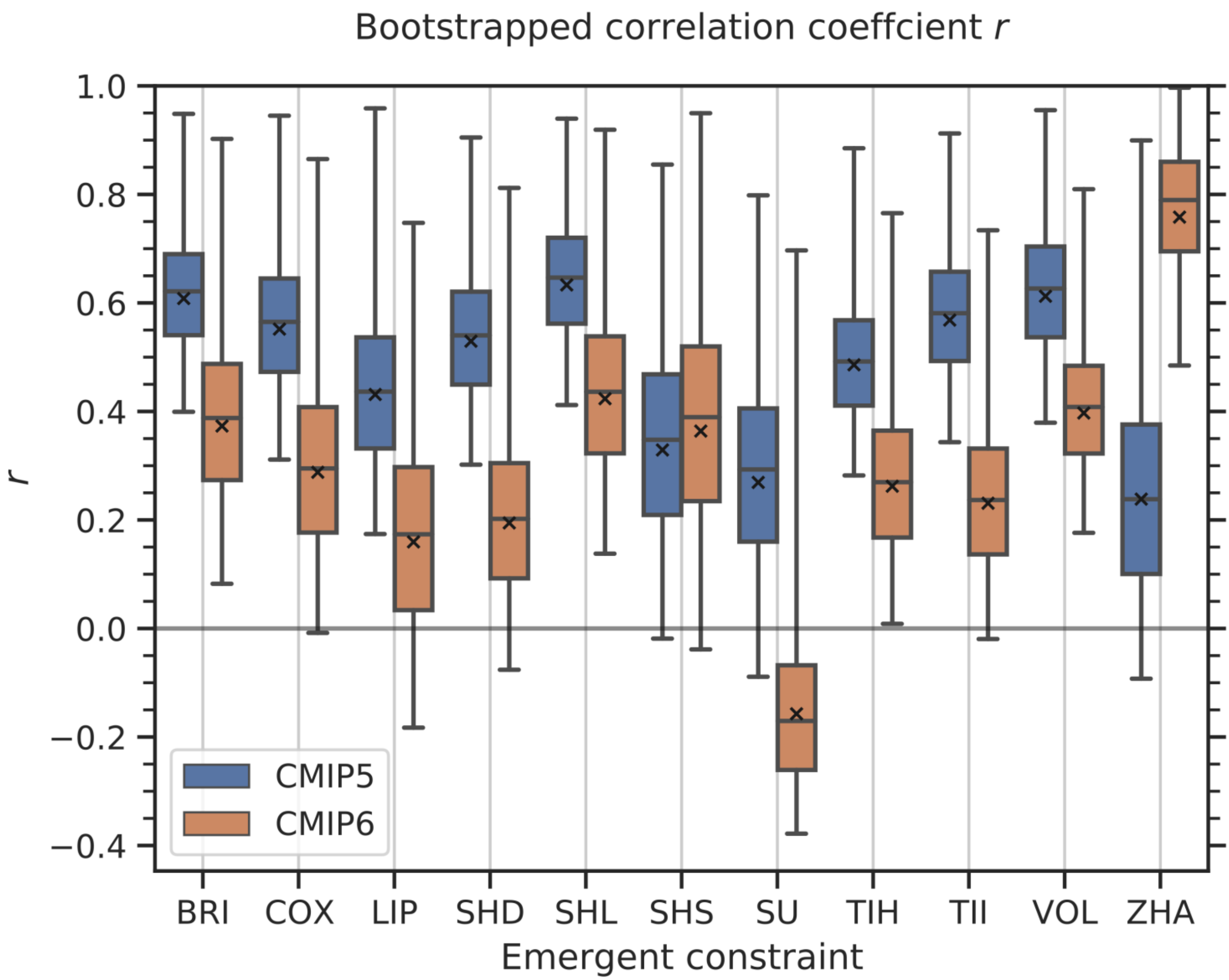

Zhai et al. (2015) constraint (ZHA) for CMIP5
ECS [K]
Response of MBLC fraction to SST changes [% K^{-1}]
all ($R^2 = 0.05, p = 0.2567$)
original ($R^2 = 0.38, p = 0.0141$)
Observation

Figure 6: Emergent relationship ZHA (Zhai et al., 2015) for different subsets of CMIP5 models. Blue circles show the 15 CMIP5 models used in the original publication (except for CESM1-CAM5): the solid blue line and blue shaded area show the emergent relationships evaluated on these models including the uncertainty range. In our study, we added 11 more CMIP5 models (red circles). The corresponding emergent relationship that considers all available CMIP5 models is shown in red colors. This relationship shows a ~~considerable~~considerably lower coefficient of determination (R^2) and higher *p*-value than the relationship using the original subset of CMIP5 models. The vertical dashed line and shaded area correspond to the observational reference and the horizontal dashed lines to the corresponding ECS constraints using this observation.



~~**Figure 7: Boxplots of the distribution of the 100,000 bootstrap samples of the linear (Pearson) correlation coefficient *r* for the different emergent constraints (x-axis) evaluated on different climate model ensembles (different colors). For better transparency, the sign of correlation coefficients from emergent relationships with expected negative slopes have been changed. Thus, positive values of *r* indicate that *r* matches the sign of the expected relationship, while negative values indicate that *r* does not match the expected sign. The "x" shows the mean of the distribution, the horizontal central line the median, the edges of the box the 25% and 75% percentile and the whiskers the 5% and 100% percentiles which form the one-sided 95% confidence interval of *r*. An emergent relationship is highly or barely significant on a climate model ensemble if the one-sided 95% confidence interval does not include *r* = 0 (horizontal gray line). Otherwise it is called "almost significant" or "far from significant" (see section 2.3 for details). Only four emergent constraints are either highly or barely significant on both climate model ensembles: BRI, SHL, TIH and VOL.**~~

8 Appendix A

	Model	ECS	BRI	COX	LIP	SHD	SHL	SHS	SU	TIH	TII	VOL	ZHA
1	ACCESS1-0	3.83	-1.59	0.20	-33.70	0.45	0.84	0.39	0.94	14.14	0.54	-26.16	-2.32
2	ACCESS1-3	3.53	-1.59	0.15	-34.32	0.54	0.90	0.36	0.93	21.05	0.64	-20.21	-2.11
3	BNU-ESM	3.92	-1.39	0.14	-33.38	0.35	0.75	0.41	1.11	22.56	1.47	-11.91	
4	CCSM4	2.94	-0.09	0.18	-37.64	0.39	0.75	0.37		24.32	1.17	-9.88	0.00
5	CNRM-CM5	3.25	0.29	0.15	-34.53	0.38	0.71	0.33	1.04	14.54	1.50	-15.43	-0.40
6	CNRM-CM5-2	3.44	0.56		-34.62	0.40	0.73	0.33		15.83	1.44	-15.08	
7	CSIRO-Mk3-6-0	4.08	-1.57	0.20	-35.12	0.61	0.97	0.36	1.00	11.19	1.01	-11.53	-0.79
8	CanESM2	3.69	-0.72	0.16	-34.28	0.30	0.61	0.31	1.00	20.00	0.69	-21.21	-1.18
9	FGOALS-g2	3.38	0.33	0.05	-30.75	0.29	0.79	0.51	1.26	24.61	0.93	-12.06	-0.54
10	GFDL-CM3	3.97	-0.26	0.32	-35.03	0.33	0.67	0.34	1.01	18.05	1.43	-14.67	-2.33
11	GFDL-ESM2G	2.39	-0.80	0.19	-36.47	0.30	0.76	0.45	0.93	18.60	0.74	-17.63	-2.43
12	GFDL-ESM2M	2.44	-0.34	0.15	-35.95	0.25	0.68	0.43	0.94	17.43	1.36	-17.75	-2.64
13	GISS-E2-H	2.31	1.18	0.10	-33.87	0.32	0.61	0.29	0.71	21.70	2.58	3.57	-0.27
14	GISS-E2-R	2.11	0.89	0.11	-34.14	0.32	0.60	0.28	0.71	12.88	2.13	2.52	-0.21
15	HadGEM2-ES	4.61	-2.45	0.26	-34.58	0.43	0.81	0.38	0.95	11.54	0.87	-24.69	-1.54
16	IPSL-CM5A-LR	4.13	-0.96	0.21	-32.13	0.41	0.86	0.45	0.95	11.58	0.53	-34.04	-1.08
17	IPSL-CM5A-MR	4.12	-1.18	0.15	-33.61	0.48	0.92	0.44	0.98	5.73	0.36	-32.90	-0.95
18	IPSL-CM5B-LR	2.60	-0.50	0.16	-32.02	0.25	0.67	0.41	0.89	21.25	0.92	-27.02	-1.90
19	MIROC-ESM	4.67	-0.99	0.22	-31.87	0.33	0.89	0.56	0.95	0.94	-0.54	-27.09	0.07
20	MIROC5	2.72	0.27	0.22	-35.51	0.36	0.78	0.42	1.01	4.90	0.40	-10.99	0.41
21	MPI-ESM-LR	3.63	-0.35	0.15	-34.67	0.41	0.86	0.45	1.07	3.16	0.24	-17.96	0.07
22	MPI-ESM-MR	3.46	-0.49	0.16	-34.33	0.42	0.87	0.45	1.04	6.10	0.30	-18.92	-0.04
23	MPI-ESM-P	3.45	-0.68		-34.36	0.41	0.87	0.46		-0.23	0.13	-17.97	0.06
24	MRI-CGCM3	2.60	-1.08	0.09	-35.01	0.37	0.78	0.41	1.04	34.02	2.04	-11.08	0.40
25	NorESM1-M	2.80	-0.74	0.13	-37.44	0.44	0.82	0.38	1.07	26.23	0.65	-7.13	0.51
26	bcc-csm1-1	2.83	-0.11	0.18	-34.25	0.37	0.78	0.41	1.15	30.84	1.34	-8.77	0.41
27	bcc-csm1-1-m	2.86	-0.46	0.13	-36.36	0.34	0.74	0.40	1.19	43.64	2.69	-7.56	0.66
28	inmcm4	2.08	-0.18	0.07	-36.43	0.19	0.52	0.33	0.84	28.16	1.89	-14.52	0.75

Table A1: All participating CMIP5 models including their ECS values (in K) and x-axis values for the different emergent constraints. Details on all constraints (including the units) are given in Table 1. The leftmost column corresponds to the index used in all plots.

	Model	ECS	BRI	COX	LIP	SHD	SHL	SHS	SU	TIH	TII	VOL	ZHA
29	ACCESS-CM2	4.72	-0.60	0.20	-35.14	0.42	0.88	0.46	0.99	21.63	1.62	-24.48	
30	ACCESS-ESM1-5	3.87	-0.81	0.13	-35.00	0.44	0.79	0.35	0.94	15.62	1.00	-20.15	
~~29~~31	AWI-CM-1-1-MR	3.16	-0.70	~~0.12~~	-35.43	0.42	0.76	0.34	1.05	2.15	-0.12	-19.44	
~~30~~32	BCC-CSM2-MR	3.04	-1.43	0.16	-36.12	0.54	0.89	0.36	1.10	32.92	1.50	-3.93	-0.10
~~31~~33	BCC-ESM1	3.26		0.19	-36.18	0.46	0.88	0.42	1.18	38.93	0.94	2.69	-0.22
~~32~~34	CAMS-CSM1-0	2.29	-0.44	0.14	-35.38	0.47	0.91	0.44	1.09	27.52	1.40	-13.84	-0.36
35	CAS-ESM2-0	3.51	-0.84	0.16	-30.53	0.31	0.80	0.48	0.95	6.98	1.58	-14.42	-0.06
~~33~~36	CESM2	5.16	-1.94	0.15	-36.22	0.49	0.85	0.36	0.93	-7.25	0.45	-24.99	-2.39
37	CESM2-FV2	5.14	-1.30	0.13	-36.88	0.45	0.79	0.34	0.93	12.09	0.30	-25.68	-2.44
~~34~~38	CESM2-WACCM	4.75	-1.60	0.18	-36.43	0.55	0.91	0.36	0.93	-5.37	-0.13	-23.81	-2.14
39	CESM2-WACCM-FV2	4.79		0.16	-36.57					26.98	0.96	-25.34	
40	CMCC-CM2-SR5	3.52	-0.19	0.18	-37.48	0.50	0.90	0.40	1.03	23.28	1.12	-22.46	-1.00
~~35~~41	CNRM-CM6-1	4.83	-0.61	0.11	-34.50	0.30	0.74	0.44	1.03	16.59	0.92	-9.16	
~~36~~42	CNRM-CM6-1-HR	4.28	-0.25	0.11	-33.41	0.34	0.81	0.46	1.04	13.31	1.04	-6.63	
~~37~~43	CNRM-ESM2-1	4.76	-0.51	0.08	-34.83	0.30	0.74	0.44	1.04	16.17	0.95	-8.60	
~~38~~44	CanESM5	5.62	-0.97	0.18	-35.50	0.48	0.82	0.34	0.97	15.39	0.04	-20.02	
~~39~~45	E3SM-1-0	5.32	-0.74	0.22	-35.59	0.46	0.80	0.34	0.96	16.73	1.26	-25.38	-3.11
~~40~~46	EC-Earth3-Veg	4.31		0.19							0.82		
~~41~~47	FGOALS-f3-L	3.00	-0.63	0.09	-33.40	0.31	0.76	0.44	1.06	11.75	1.00	-0.19	
~~42~~48	~~GFDL-CM4~~FGOALS-g3	~~3.87~~2.88	-~~1.06~~0.12	0.~~16~~13	-34.~~14~~57	0.~~49~~36	0.~~92~~80	0.~~42~~44	1.~~03~~18	~~3.08~~19.06	0.32	-~~17.36~~15.13	1.06
43	~~GFDL-ESM4~~	~~2.62~~		~~0.15~~	~~-34.82~~	~~0.48~~				~~6.07~~	~~0.64~~	~~-17.53~~	

	Model	ECS	BRI	COX	LIP	SHD	SHL	SHS	SU	TIH	TII	VOL	ZHA
~~44~~49	GISS-E2-1-G	2.72	-0.58	0.17	-36.24	0.26	0.61	0.35	0.98	28.46	2.39	-7.75	-0.07
~~45~~50	GISS-E2-1-H	3.11	-0.48	0.13	-36.95	0.24	0.59	0.36	1.00	24.55	1.84	-6.82	0.00
~~46~~51	HadGEM3-GC31-LL	5.55	-0.56	0.20	-35.75	0.46	0.92	0.46	1.02	9.56	0.99	-24.87	-2.66
52	HadGEM3-GC31-MM	5.42	-1.06	0.18	-35.09				1.01	4.73	1.19	-22.48	
~~47~~53	INM-CM4-8	1.83	0.03	0.13	-38.15	0.47	0.71	0.25	0.99	12.22	0.62	-16.71	
~~48~~54	INM-CM5-0	1.92	-0.46	0.14	-37.34	0.40	0.63	0.23	0.99	11.22	0.62	-14.38	
~~49~~55	IPSL-CM6A-LR	4.56	-0.38	0.14	-33.74	0.50	0.92	0.41	0.76	11.90	0.41	-34.90	
56	KACE-1-0-G	4.48	-0.60	0.15	-34.95	0.41	0.87	0.47	1.03	22.16	1.66	-24.51	
~~50~~57	MCM-UA-1-0	3.65		0.20					0.91	14.50	-0.56		
~~51~~58	MIROC-ES2L	2.68	-0.49	0.15	-34.13	0.36	0.79	0.43	0.91	18.53	0.68	-13.03	-0.14
~~52~~59	MIROC6	2.61	-0.45	0.17	-34.90	0.46	0.91	0.45	0.99	5.79	0.29	-13.43	-1.08
60	MPI-ESM-1-2-HAM	2.96	-0.60	0.14	-35.21	0.42	0.77	0.35	0.96	14.47	0.70	-23.06	-0.73
~~53~~61	MPI-ESM1-2-HR	2.98	-0.34	0.15	-34.60	0.45	0.78	0.33	1.04	4.33	0.53	-20.21	-0.16
62	MPI-ESM1-2-LR	3.00	-0.64	0.18	-34.76	0.39	0.68	0.29	1.02	5.92	0.40	-19.40	-0.42
~~54~~63	MRI-ESM2-0	3.15	-0.73	0.15	-33.39	0.39	0.79	0.40	0.95	3.33	0.75	-22.83	-1.35
~~55~~64	NESM3	4.72	-0.73	0.24	-36.77	0.52	0.89	0.37	1.05	28.78	0.55	-18.81	-0.62
65	NorCPM1	3.05		0.12	-37.45	0.45				30.50	0.54	-8.09	
~~56~~66	NorESM2-LM	2.54	-1.61	0.20	-36.17	0.46	0.79	0.33	1.01	34.40	1.30	-16.22	-1.25
67	NorESM2-MM	2.50	0.12	0.31	-36.55	0.52	0.86	0.34	1.00	11.07	0.27	-25.33	-1.03
~~57~~68	SAM0-UNICON	3.72	-1.40	0.19	-36.41	0.54	0.90	0.36	1.01	11.99	1.13	-25.25	-1.85

	Model	ECS	BRI	COX	LIP	SHD	SHL	SHS	SU	TIH	TII	VOL	ZHA
69	TaiESM1	4.31	-0.22	0.15	-36.80	0.48	0.89	0.42	1.02	11.20	1.17	-16.53	-1.89
~~58~~70	UKESM1-0-LL	5.34	-0.59	0.21	-35.67	0.48	0.93	0.45	1.03	2.22	0.86	-26.83	-2.72

Table A2: As Table A1, but for the CMIP6 models.

9 Code availability

The corresponding ESMValTool recipe that can be used to reproduce the figures of this paper will be included in ESMValTool v2~~.0~~ (Eyring et al., 2020; Lauer et al., 2020; Righi et al., 2020) at the time of publication of this paper. ESMValTool v2~~.0~~ is released under the Apache License, VERSION 2.0. The latest release of ESMValTool v2~~.0~~ is publicly available on Zenodo at https://doi.org/10.5281/zenodo.3401363. The source code of the ESMValCore package, which is installed as a dependency of the ESMValTool v2~~.0~~, is also publicly available on Zenodo at https://doi.org/10.5281/zenodo.3387139. ESMValTool and ESMValCore are developed on the GitHub repositories available at https://github.com/ESMValGroup.

10 Data availability

CMIP5 and CMIP6 model output (see Table 2 and Table 3) is available through the Earth System Grid Foundation (ESGF) and can be directly used within the ESMValTool (e.g. https://esgf-data.dkrz.de/projects/esgf-dkrz/). Downloading instructions and preprocessing scripts for the observational datasets detailed in Table 1 are included in the ESMValTool distribution.

11 Author contribution

MS led the writing and analysis of the paper. MS and AL coded the emergent constraints in the ESMValTool. VE, PG and SCS contributed to the concept of the study and the interpretation of the results. All authors contributed to the writing of the manuscript.

12 Competing interests

The authors declare that they have no conflict of interest.

13 Acknowledgements

This work has been supported by the European Union's Horizon 2020 Framework Programme for Research and Innovation "Coordinated Research in Earth Systems and Climate: Experiments, kNowledge, Dissemination and Outreach (CRESCENDO)" project under Grant Agreement No. 641816, the European Union's Horizon 2020 project "Climate-Carbon Interactions in the Coming Century" (4C) under Grant Agreement No. 821003 and the "Advanced Earth System Model Evaluation for CMIP (EVal4CMIP)" project funded by the Helmholtz Society. We acknowledge the World Climate Research Programme (WCRP), which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP5 and CMIP6. We thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP and

ESGF. The computational resources of the Deutsches Klimarechenzentrum (DKRZ, Hamburg, Germany) that allowed the analysis of this study are kindly acknowledged. We also thank Lisa Bock for helpful comments on the manuscript.

14 References

Adler, R. F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P. P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S., Bolvin, D., Gruber, A., Susskind, J., Arkin, P., and Nelkin, E.: The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979-present), J Hydrometeorol, 4, 1147-1167, 2003.

Allen, M. R. and Ingram, W. J.: Constraints on future changes in climate and the hydrologic cycle, Nature, 419, 224-+, 2002.

AMSR-E: AMSR-E Level 3 Sea Surface Temperature for Climate Model Comparison. Ver. 1. PO.DAAC, CA, USA, Remote Sensing Systems, doi: https://doi.org/10.5067/SST00-1D1M1, 2011. 2011.

Andrews, T., Gregory, J. M., Webb, M. J., and Taylor, K. E.: Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models, Geophys Res Lett, 39, 2012.

Annan, J. D., Hargreaves, J. C., Mauritsen, T., and Stevens, B.: What could we learn about climate sensitivity from variability in the surface temperature record?, Earth Syst Dynam, 11, 709-719, 2020.

Arora, V. K., Scinocca, J. F., Boer, G. J., Christian, J. R., Denman, K. L., Flato, G. M., Kharin, V. V., Lee, W. G., and Merryfield, W. J.: Carbon emission limits required to satisfy future representative concentration pathways of greenhouse gases, Geophys Res Lett, 38, 2011.

Aumann, H. H., Chahine, M. T., Gautier, C., Goldberg, M. D., Kalnay, E., McMillin, L. M., Revercomb, H., Rosenkranz, P. W., Smith, W. L., Staelin, D. H., Strow, L. L., and Susskind, J.: AIRS/AMSU/HSB on the aqua mission: Design, science objectives, data products, and processing systems, Ieee T Geosci Remote, 41, 253-264, 2003.

Beer, R.: TES on the Aura mission: Scientific objectives, measurements, and analysis overview, Ieee T Geosci Remote, 44, 1102-1105, 2006.

Bellprat, O., Massonnet, F., Siegert, S., Prodhomme, C., Macias-Gomez, D., Guemas, V., and Doblas-Reyes, F.: Uncertainty propagation in observational references to climate model scales, Remote Sens Environ, 203, 101-108, 2017.

Bellucci, A., Gualdi, S., and Navarra, A.: The double-ITCZ syndrome in coupled general circulation models: the role of large-scale vertical circulation regimes, Journal of Climate, 23, 1127-1145, 2010.

Bentsen, M., Bethke, I., Debernard, J. B., Iversen, T., Kirkevag, A., Seland, O., Drange, H., Roelandt, C., Seierstad, I. A., Hoose, C., and Kristjansson, J. E.: The Norwegian Earth System Model, NorESM1-M - Part 1: Description and basic evaluation of the physical climate, Geosci Model Dev, 6, 687-720, 2013.

Bi, D. H., Dix, M., Marsland, S. J., O'Farrell, S., Rashid, H. A., Uotila, P., Hirst, A. C., Kowalczyk, E., Golebiewski, M., Sullivan, A., Yan, H. L., Hannah, N., Franklin, C., Sun, Z. A., Vohralik, P., Watterson, I., Zhou, X. B., Fiedler, R., Collier, M., Ma, Y. M., Noonan, J., Stevens, L., Uhe, P., Zhu, H. Y., Griffies, S. M., Hill, R., Harris, C., and Puri, K.: The ACCESS coupled model: description, control climate and evaluation, Aust Meteorol Ocean, 63, 41-64, 2013.

Bodas-Salcedo, A., Mulcahy, J. P., Andrews, T., Williams, K. D., Ringer, M. A., Field, P. R., and Elsaesser, G. S.: Strong Dependence of Atmospheric Feedbacks on Mixed-Phase Microphysics and Aerosol-Cloud Interactions in HadGEM3, J Adv Model Earth Sy, 11, 1735-1758, 2019.

Boucher, O., Randall, D., Artaxo, P., Bretherton, C., Feingold, G., Forster, P., Kerminen, V.-M., Kondo, Y., Liao, H., and Lohmann, U.: Clouds and aerosols. In: Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, 2013.

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., Bekki, S., Bonnet, R., Bony, S., Bopp, L., Braconnot, P., Brockmann, P., Cadule, P., Caubel, A., Cheruy, F., Codron, F., Cozic, A., Cugnet, D., D'Andrea, F., Davini, P., de Lavergne, C., Denvil, S., Deshayes, J., Devilliers, M., Ducharne, A., Dufresne, J.~~-~~. L., Dupont, E., ~~Éthé~~Ethe, C., Fairhead, L., Falletti, L., Flavoni, S., Foujols, M.~~-~~. A., Gardoll, S., Gastineau, G., Ghattas, J., Grandpeix, J.~~-~~. Y., Guenet, B., Guez, L. E., Guilyardi, ~~É~~E., Guimberteau, M., Hauglustaine, D., Hourdin, F., Idelkadi, A., Joussaume, S., Kageyama, M., Khodri, M., Krinner, G., Lebas, N., Levavasseur, G., ~~Lévy~~Levy, C., Li, L., Lott, F., Lurton, T., Luyssaert, S., Madec, G., Madeleine, J.~~-~~. B., Maignan, F., Marchand, M., Marti, O., Mellul, L., Meurdesoif, Y., Mignot, J., Musat, I., ~~Ottlé~~Ottle, C., Peylin, P., Planton, Y., Polcher, J., Rio, C., Rochetin, N., Rousset, C., Sepulchre, P., Sima, A., Swingedouw, D., ~~Thiéblemont~~Thieblemont, R., Traore, A. K., Vancoppenolle, M., Vial, J., Vialard, J., Viovy, N., and Vuichard, N.: Presentation and ~~evaluation~~Evaluation of the IPSL-CM6A-LR ~~climate model, n/a, e2019MS002010~~Climate Model, J Adv Model Earth Sy, 12, 2020.

Bretherton, C. S. and Caldwell, P. M.: Combining Emergent Constraints for Climate Sensitivity, J Climate, 33, 7413-7430, 2020.

Brient, F.: Reducing Uncertainties in Climate Projections with Emergent Constraints: Concepts, Examples and Prospects, Adv Atmos Sci, 37, 1-15, 2020.

Brient, F. and Schneider, T.: Constraints on Climate Sensitivity from Space-Based Measurements of Low-Cloud Reflection, J Climate, 29, 5821-5835, 2016.

Brown, P. T., Stolpe, M. B., and Caldeira, K.: Assumptions for emergent constraints, Nature, 563, E1-E3, 2018.

Caldwell, P. M., Bretherton, C. S., Zelinka, M. D., Klein, S. A., Santer, B. D., and Sanderson, B. M.: Statistical significance of climate sensitivity predictors obtained by data mining, Geophys Res Lett, 41, 1803-1808, 2014.

Caldwell, P. M., Zelinka, M. D., and Klein, S. A.: Evaluating Emergent Constraints on Equilibrium Climate Sensitivity, J Climate, 31, 3921-3942, 2018.

Cao, J., Wang, B., Yang, Y. M., Ma, L. B., Li, J., Sun, B., Bao, Y., He, J., Zhou, X., and Wu, L. G.: The NUIST Earth System Model (NESM) version 3: description and preliminary evaluation, Geosci Model Dev, 11, 2975-2993, 2018.

Ceppi, P., Brient, F., Zelinka, M. D., and Hartmann, D. L.: Cloud feedback mechanisms and their representation in global climate models, Wires Clim Change, 8, 2017.

Charney, J. G., Arakawa, A., Baker, D. J., Bolin, B., Dickinson, R. E., Goody, R. M., Leith, C. E., Stommel, H. M., and Wunsch, C. I.: Carbon dioxide and climate: a scientific assessment, National Academy of Sciences, Washington, DC, 1979.

Cherchi, A., Fogli, P. G., Lovato, T., Peano, D., Iovino, D., Gualdi, S., Masina, S., Scoccimarro, E., Materia, S., Bellucci, A., and Navarra, A.: Global Mean Climate and Main Patterns of Variability in the CMCC-CM2 Coupled Model, J Adv Model Earth Sy, 11, 185-209, 2019.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., Gao, X., Gutowski, W. J., Johns, T., and Krinner, G.: Long-term climate change: projections, commitments and irreversibility. In: Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, 2013.

Collins, W. J., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Halloran, P., Hinton, T., Hughes, J., Jones, C. D., Joshi, M., Liddicoat, S., Martin, G., O'Connor, F., Rae, J., Senior, C., Sitch, S., Totterdell, I., Wiltshire, A., and Woodward, S.: Development and evaluation of an Earth-System model-HadGEM2, Geosci Model Dev, 4, 1051-1075, 2011.

Counillon, F., Keenlyside, N., Bethke, I., Wang, Y. G., Billeau, S., Shen, M. L., and Bentsen, M.: Flow-dependent assimilation of sea surface temperature in isopycnal coordinates with the Norwegian Climate Prediction Model, Tellus A, 68, 2016.

Cox, P. M., Huntingford, C., and Williamson, M. S.: Emergent constraint on equilibrium climate sensitivity from global temperature variability, Nature, 553, 319-+, 2018.

Danabasoglu, G., Lamarque, J. F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., Emmons, L. K., Fasullo, J., Garcia, R., Gettelman, A., Hannay, C., Holland, M. M., Large, W. G., Lauritzen, P. H., Lawrence, D. M., Lenaerts, J. T. M., Lindsay, K., Lipscomb, W. H., Mills, M. J., Neale, R., Oleson, K. W., Otto-Bliesner, B., Phillips, A. S., Sacks, W., Tilmes, S., van Kampenhout, L., Vertenstein, M., Bertini, A., Dennis, J., Deser, C., Fischer, C., Fox-Kemper, B., Kay, J. E., Kinnison, D., Kushner, P. J., Larson, V. E., Long, M. C., Mickelson, S., Moore, J. K., Nienhouse, E., Polvani, L., Rasch, P. J., and Strand, W. G.: The Community Earth System Model Version 2 (CESM2), J Adv Model Earth Sy, 12, 2020.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Holm, E. V., Isaksen, L., Kallberg, P., Kohler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thepaut, J. N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q J Roy Meteor Soc, 137, 553-597, 2011.

Delworth, T. L., Stouffer, R. J., Dixon, K. W., Spelman, M. J., Knutson, T. R., Broccoli, A. J., Kushner, P. J., and Wetherald, R. T.: Review of simulations of climate variability and change with the GFDL R30 coupled climate model, Clim Dynam, 19, 555-574, 2002.

Dix, M., Vohralik, P., Bi, D. H., Rashid, H., Marsland, S., O'Farrell, S., Uotila, P., Hirst, T., Kowalczyk, E., Sullivan, A., Yan, H. L., Franklin, C., Sun, Z. A., Watterson, I., Collier, M., Noonan, J., Rotstayn, L., Stevens, L., Uhe, P., and Puri, K.: The ACCESS coupled model: documentation of core CMIP5 simulations and initial results, Aust Meteorol Ocean, 63, 83-99, 2013.

Donner, L. J., Wyman, B. L., Hemler, R. S., Horowitz, L. W., Ming, Y., Zhao, M., Golaz, J. C., Ginoux, P., Lin, S. J., Schwarzkopf, M. D., Austin, J., Alaka, G., Cooke, W. F., Delworth, T. L., Freidenreich, S. M., Gordon, C. T., Griffies, S. M., Held, I. M., Hurlin, W. J., Klein, S. A., Knutson, T. R., Langenhorst, A. R., Lee, H. C., Lin, Y. L., Magi, B. I., Malyshev, S. L., Milly, P. C. D., Naik, V., Nath, M. J., Pincus, R., Ploshay, J. J., Ramaswamy, V., Seman, C. J., Shevliakova, E., Sirutis, J. J., Stern, W. F., Stouffer, R. J., Wilson, R. J., Winton, M., Wittenberg, A. T., and Zeng, F. R.: The Dynamical Core, Physical Parameterizations, and Basic Simulation Characteristics of the Atmospheric Component AM3 of the GFDL Global Coupled Model CM3, J Climate, 24, 3484-3519, 2011.

Dufresne, J. L., Foujols, M. A., Denvil, S., Caubel, A., Marti, O., Aumont, O., Balkanski, Y., Bekki, S., Bellenger, H., Benshila, R., Bony, S., Bopp, L., Braconnot, P., Brockmann, P., Cadule, P., Cheruy, F., Codron, F., Cozic, A., Cugnet, D., de Noblet, N., Duvel, J. P., Ethe, C., Fairhead, L., Fichefet, T., Flavoni, S., Friedlingstein, P., Grandpeix, J. Y., Guez, L., Guilyardi, E., Hauglustaine, D., Hourdin, F., Idelkadi, A., Ghattas, J., Joussaume, S., Kageyama, M., Krinner, G., Labetoulle, S., Lahellec, A., Lefebvre, M. P., Lefevre, F., Levy, C., Li, Z. X., Lloyd, J., Lott, F., Madec, G., Mancip, M., Marchand, M., Masson, S., Meurdesoif, Y., Mignot, J., Musat, I., Parouty, S., Polcher, J., Rio, C., Schulz, M., Swingedouw, D., Szopa, S., Talandier, C., Terray, P., Viovy, N., and Vuichard, N.: Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5, Clim Dynam, 40, 2123-2165, 2013.

Dunne, J. P., John, J. G., Adcroft, A. J., Griffies, S. M., Hallberg, R. W., Shevliakova, E., Stouffer, R. J., Cooke, W., Dunne, K. A., Harrison, M. J., Krasting, J. P., Malyshev, S. L., Milly, P. C. D., Phillipps, P. J., Sentman, L. T., Samuels, B. L., Spelman, M. J., Winton, M., Wittenberg, A. T., and Zadeh, N.: GFDL's ESM2 Global Coupled Climate-Carbon Earth System Models. Part I: Physical Formulation and Baseline Simulation Characteristics, J Climate, 25, 6646-6665, 2012.

Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., Arnone, E., Bellprat, O., Brötz, B., Caron, L. P., Carvalhais, N., Cionni, I., Cortesi, N., Crezee, B., Davin, E., Davini, P., Debeire, K., de Mora, L., Deser, C., Docquier, D., Earnshaw, P., Ehbrecht, C., Gier, B. K., Gonzalez-Reviriego, N., Goodman, P., Hagemann, S., Hardiman, S., Hassler, B., Hunter, A., Kadow, C., Kindermann, S., Koirala, S., Koldunov, N. V., Lejeune, Q., Lembo, V., Lovato, T., Lucarini, V., Massonnet, F., Müller, B., Pandde, A., Pérez-Zanón, N., Phillips, A., Predoi, V., Russell, J., Sellar, A., Serva, F., Stacke, T., Swaminathan, R., Torralba, V., Vegas-Regidor, J., von Hardenberg, J., Weigel, K., and Zimmermann, K.: ESMValTool v2.0 – Extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP Geosci. Model Dev. Discuss., doi: 10.5194/gmd-2019-291, 2020. 1-81, 2020.

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, Geosci Model Dev, 9, 1937-1958, 2016.

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., Collins, W. D., Gier, B. K., Hall, A. D., Hoffman, F. M., Hurtt, G. C., Jahn, A., Jones, C. D., Klein, S. A., Krasting, J. P., Kwiatkowski, L., Lorenz, R., Maloney, E., Meehl, G. A., Pendergrass, A. G., Pincus, R., Ruane, A. C., Russell, J. L., Sanderson, B. M., Santer, B. D., Sherwood, S. C., Simpson, I. R., Stouffer, R. J., and Williamson, M. S.: Taking climate model evaluation to the next level, Nat Clim Change, 9, 102-110, 2019.

~~Forster, P. M., Maycock, A. C., McKenna, C. M., and Smith, C. J.: Latest climate models confirm need for urgent mitigation, Nat Clim Change, 10, 7-10, 2020.~~

Fasullo, J. T. and Trenberth, K. E.: A Less Cloudy Future: The Role of Subtropical Subsidence in Climate Sensitivity, Science, 338, 792-794, 2012.

Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., Lawrence, D. M., Neale, R. B., Rasch, P. J., Vertenstein, M., Worley, P. H., Yang, Z. L., and Zhang, M. H.: The Community Climate System Model Version 4, J Climate, 24, 4973-4991, 2011.

Gettelman, A., Hannay, C., Bacmeister, J. T., Neale, R. B., Pendergrass, A. G., Danabasoglu, G., Lamarque, J. F., Fasullo, J. T., Bailey, D. A., Lawrence, D. M., and Mills, M. J.: High Climate Sensitivity in the Community Earth System Model Version 2 (CESM2), Geophys Res Lett, 46, 8329-8337, 2019a.

Gettelman, A., Mills, M. J., Kinnison, D. E., Garcia, R. R., Smith, A. K., Marsh, D. R., Tilmes, S., Vitt, F., Bardeen, C. G., McInerny, J., Liu, H. L., Solomon, S. C., Polvani, L. M., Emmons, L. K., Lamarque, J. F., Richter, J. H., Glanville, A. S., Bacmeister, J. T., Phillips, A. S., Neale, R. B., Simpson, I. R., DuVivier, A. K., Hodzic, A., and Randel, W. J.: The Whole Atmosphere Community Climate Model Version 6 (WACCM6), J Geophys Res-Atmos, doi: 10.1029/2019jd030943, 2019b. 2019b.

Giorgetta, M. A., Jungclaus, J., Reick, C. H., Legutke, S., Bader, J., Bottinger, M., Brovkin, V., Crueger, T., Esch, M., Fieg, K., Glushak, K., Gayler, V., Haak, H., Hollweg, H. D., Ilyina, T., Kinne, S., Kornblueh, L., Matei, D., Mauritsen, T., Mikolajewicz, U., Mueller, W., Notz, D., Pithan, F., Raddatz, T., Rast, S., Redler, R., Roeckner, E., Schmidt, H., Schnur, R., Segschneider, J., Six, K. D., Stockhause, M., Timmreck, C., Wegner, J., Widmann, H., Wieners, K. H., Claussen, M., Marotzke, J., and Stevens, B.: Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5, J Adv Model Earth Sy, 5, 572-597, 2013.

Golaz, J. C., Caldwell, P. M., Van Roekel, L. P., Petersen, M. R., Tang, Q., Wolfe, J. D., Abeshu, G., Anantharaj, V., Asay-Davis, X. S., Bader, D. C., Baldwin, S. A., Bisht, G., Bogenschutz, P. A., Branstetter, M., Brunke, M. A., Brus, S. R., Burrows, S. M., Cameron-Smith, P. J., Donahue, A. S., Deakin, M., Easter, R. C., Evans, K. J., Feng, Y., Flanner, M., Foucar, J. G., Fyke, J. G., Griffin, B. M., Hannay, C., Harrop, B. E., Hoffman, M. J., Hunke, E. C., Jacob, R. L., Jacobsen, D. W., Jeffery, N., Jones, P. W., Keen, N. D., Klein, S. A., Larson, V. E., Leung, L. R., Li, H. Y., Lin, W. Y., Lipscomb, W. H., Ma, P. L., Mahajan, S., Maltrud, M. E., Mametjanov, A., McClean, J. L., McCoy, R. B., Neale, R. B., Price, S. F., Qian, Y., Rasch, P. J., Eyre, J. E. J. R., Riley, W. J., Ringler, T. D., Roberts, A. F., Roesler, E. L., Salinger, A. G., Shaheen, Z., Shi, X. Y., Singh, B., Tang, J. Y., Taylor, M. A., Thornton, P. E., Turner, A. K., Veneziani, M., Wan, H., Wang, H. L., Wang, S. L., Williams, D. N., Wolfram, P. J., Worley, P. H., Xie, S. C., Yang, Y., Yoon, J. H., Zelinka, M. D., Zender, C. S., Zeng, X. B., Zhang, C. Z., Zhang, K., Zhang, Y., Zheng, X., Zhou, T., and Zhu, Q.: The DOE E3SM Coupled Model Version 1: Overview and Evaluation at Standard Resolution, J Adv Model Earth Sy, 11, 2089-2129, 2019.

Gregory, J. M., Ingram, W. J., Palmer, M. A., Jones, G. S., Stott, P. A., Thorpe, R. B., Lowe, J. A., Johns, T. C., and Williams, K. D.: A new method for diagnosing radiative forcing and climate sensitivity, Geophys Res Lett, 31, 2004.

Guo, Y. Y., Yu, Y. Q., Lin, P. F., Liu, H. L., He, B., Bao, Q., Zhao, S. W., and Wang, X. W.: Overview of the CMIP6 Historical Experiment Datasets with the Climate System Model CAS FGOALS-f3-L, Adv Atmos Sci, doi: 10.1007/s00376-020-2004-4, 2020. 2020.

Hajima, T., Watanabe, M., Yamamoto, A., Tatebe, H., Noguchi, M. A., Abe, M., Ohgaito, R., Ito, A., Yamazaki, D., Okajima, H., Ito, A., Takata, K., Ogochi, K., Watanabe, S., and Kawamiya, M.: Development of the MIROC-ES2L Earth system model and the evaluation of biogeochemical processes and feedbacks, Geosci Model Dev, 13, 2197-2244, 2020.

Hall, A. and Qu, X.: Using the current seasonal cycle to constrain snow albedo feedback in future climate change, Geophys Res Lett, 33, 2006.

Hargreaves, J. C. and Annan, J. D.: Could the Pliocene constrain the equilibrium climate sensitivity?, Clim Past, 12, 1591-1599, 2016.

Hargreaves, J. C., Annan, J. D., Yoshimori, M., and Abe-Ouchi, A.: Can the Last Glacial Maximum constrain climate sensitivity?, Geophys Res Lett, 39, 2012.

He, B., Bao, Q., Wang, X. C., Zhou, L. J., Wu, X. F., Liu, Y. M., Wu, G. X., Chen, K. J., He, S. C., Hu, W. T., Li, J. D., Li, J. X., Nian, G. K., Wang, L., Yang, J., Zhang, M. H., and Zhang, X. Q.: CAS FGOALS-f3-L Model Datasets for CMIP6 Historical Atmospheric Model Intercomparison Project Simulation, Adv Atmos Sci, 36, 771-778, 2019.

He, B., Liu, Y. M., Wu, G. X., Bao, Q., Zhou, T. J., Wu, X. F., Wang, L., Li, J. D., Wang, X. C., Li, J. X., Hu, W. T., Zhang, X. Q., Sheng, C., and Tang, Y. Q.: CAS FGOALS-f3-L Model Datasets for CMIP6 GMMIP Tier-1 and Tier-3 Experiments, Adv Atmos Sci, 37, 18-28, 2020.

~~Held, I. M., Guo, H., Adcroft, A., Dunne, J. P., Horowitz, L. W., Krasting, J., Shevliakova, E., Winton, M., Zhao, M., Bushuk, M., Wittenberg, A. T., Wyman, B., Xiang, B., Zhang, R., Anderson, W., Balaji, V., Donner, L., Dunne, K., Durachta, J., Gauthier, P. P. G., Ginoux, P., Golaz, J. C., Griffies, S. M., Hallberg, R., Harris, L., Harrison, M., Hurlin, W., John, J., Lin, P., Lin, S. J., Malyshev, S., Menzel, R., Milly, P. C. D., Ming, Y., Naik, V., Paynter, D., Paulot, F., Rammaswamy, V., Reichl,~~

~~B., Robinson, T., Rosati, A., Seman, C., Silvers, L. G., Underwood, S., and Zadeh, N.: Structure and Performance of GFDL's CM4.0 Climate Model, J Adv Model Earth Sy, 11, 3691-3727, 2019.~~

Hirota, N., Takayabu, Y. N., Watanabe, M., and Kimoto, M. J. J. o. C.: Precipitation reproducibility over tropical oceans and its relationship to the double ITCZ problem in CMIP3 and MIROC5 climate models, 24, 4859-4873, 2011.

Iversen, T., Bentsen, M., Bethke, I., Debernard, J. B., Kirkevag, A., Seland, O., Drange, H., Kristjansson, J. E., Medhaug, I., Sand, M., and Seierstad, I. A.: The Norwegian Earth System Model, NorESM1-M - Part 2: Climate response and scenario projections, Geosci Model Dev, 6, 389-415, 2013.

Ji, D., Wang, L., Feng, J., Wu, Q., Cheng, H., Zhang, Q., Yang, J., Dong, W., Dai, Y., Gong, D., Zhang, R. H., Wang, X., Liu, J., Moore, J. C., Chen, D., and Zhou, M.: Description and basic evaluation of Beijing Normal University Earth System Model (BNU-ESM) version 1, Geosci Model Dev, 7, 2039-2064, 2014.

Jimenez-de-la-Cuesta, D. and Mauritsen, T.: Emergent constraints on Earth's transient and equilibrium response to doubled CO2 from post-1970s global warming, Nat Geosci, 12, 902-+, 2019.

Knutti, R., Rugenstein, M. A. A., and Hegerl, G. C.: Beyond equilibrium climate sensitivity, Nat Geosci, 10, 727-+, 2017a.

Knutti, R., Sedlacek, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., and Eyring, V.: A climate model projection weighting scheme accounting for performance and interdependence, Geophys Res Lett, 44, 1909-1918, 2017b.

Kuhlbrodt, T., Jones, C. G., Sellar, A., Storkey, D., Blockley, E., Stringer, M., Hill, R., Graham, T., Ridley, J., Blaker, A., Calvert, D., Copsey, D., Ellis, R., Hewitt, H., Hyder, P., Ineson, S., Mulcahy, J., Siahaan, A., and Walton, J.: The Low-Resolution Version of HadGEM3 GC3.1: Development and Evaluation for Global Climate, J Adv Model Earth Sy, 10, 2865-2888, 2018.

Lauer, A., Eyring, V., Bellprat, O., Bock, L., Gier, B. K., Hunter, A., Lorenz, R., Pérez-Zanón, N., Righi, M., Schlund, M., Senftleben, D., Weigel, K., and Zechlau, S.: Earth System Model Evaluation Tool (ESMValTool) v2.0 – diagnostics for emergent constraints and future projections from Earth system models in CMIP, Geosci. Model Dev. Discuss., 2020, 1-47, 2020.

Lauer, A., ~~Hamilton, K., Wang, Y. Q., Phillips, V. T. J., and Bennartz, R.: The Impact of Global Warming on Marine Boundary Layer Clouds over the Eastern Pacific-A Regional Model Study, J Climate, 23, 5844-5863, 2010.~~

~~Lauer, A.,~~ Jones, C., Eyring, V., Evaldsson, M., Stefan, H. A., Makela, J., Martin, G., Roehrig, R., and Wang, S. Y.: Process-level improvements in CMIP5 models and their impact on tropical variability, the Southern Ocean, and monsoons, Earth Syst Dynam, 9, 33-67, 2018.

Law, R. M., Ziehn, T., Matear, R. J., Lenton, A., Chamberlain, M. A., Stevens, L. E., Wang, Y. P., Srbinovsky, J., Bi, D. H., Yan, H. L., and Vohralik, P. F.: The carbon cycle in the Australian Community Climate and Earth System Simulator (ACCESS-ESM1) - Part 1: Model description and pre-industrial simulation, Geosci Model Dev, 10, 2567-2590, 2017.

Lee, J., Kim, J., Sun, M. A., Kim, B. H., Moon, H., Sung, H. M., Kim, J., and Byun, Y. H.: Evaluation of the Korea Meteorological Administration Advanced Community Earth-System model (K-ACE), Asia-Pac J Atmos Sci, 56, 381-395, 2020a.

Lee, W. L., Wang, Y. C., Shiu, C. J., Tsai, I. C., Tu, C. Y., Lan, Y. Y., Chen, J. P., Pan, H. L., and Hsu, H. H.: Taiwan Earth System Model Version 1: description and evaluation of mean state, Geosci Model Dev, 13, 3887-3904, 2020b.

Li, L., Yu, Y., Tang, Y., Lin, P., Xie, J., Song, M., Dong, L., Zhou, T., Liu, L., Wang, L., Pu, Y., Chen, X., Chen, L., Xie, Z., Liu, H., Zhang, L., Huang, X., Feng, T., Zheng, W., Xia, K., Liu, H., Liu, J., Wang, Y., Wang, L., Jia, B., Xie, F., Wang, B., Zhao, S., Yu, Z., Zhao, B., and Wei, J.: The Flexible Global Ocean-Atmosphere-Land System Model Grid-Point Version 3 (FGOALS-g3): Description and Evaluation, 12, e2019MS002012, 2020.

Li, L. J., Lin, P. F., Yu, Y. Q., Wang, B., Zhou, T. J., Liu, L., Liu, J. P., Bao, Q., Xu, S. M., Huang, W. Y., Xia, K., Pu, Y., Dong, L., Shen, S., Liu, Y. M., Hu, N., Liu, M. M., Sun, W. Q., Shi, X. J., Zheng, W. P., Wu, B., Song, M. R., Liu, H. L., Zhang, X. H., Wu, G. X., Xue, W., Huang, X. M., Yang, G. W., Song, Z. Y., and Qiao, F. L.: The flexible global ocean-atmosphere-land system model, Grid-point Version 2: FGOALS-g2, Adv Atmos Sci, 30, 543-560, 2013.

Lipat, B. R., Tselioudis, G., Grise, K. M., and Polvani, L. M.: CMIP5 models' shortwave cloud radiative response and climate sensitivity linked to the climatological Hadley cell extent, Geophys Res Lett, 44, 5739-5748, 2017.

Loeb, N. G., Doelling, D. R., Wang, H. L., Su, W. Y., Nguyen, C., Corbett, J. G., Liang, L. S., Mitrescu, C., Rose, F. G., and Kato, S.: Clouds and the Earth's Radiant Energy System (CERES) Energy Balanced and Filled (EBAF) Top-of-Atmosphere (TOA) Edition-4.0 Data Product, J Climate, 31, 895-918, 2018.

Mace, G. G., Zhang, Q. Q., Vaughan, M., Marchand, R., Stephens, G., Trepte, C., and Winker, D.: A description of hydrometeor layer occurrence statistics derived from the first year of merged Cloudsat and CALIPSO data, J Geophys Res-Atmos, 114, 2009.

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., Brovkin, V., Claussen, M., Crueger, T., Esch, M., Fast, I., Fiedler, S., Flaeschner, D., Gayler, V., Giorgetta, M., Goll, D. S., Haak, H., Hagemann, S., Hedemann, C., Hohenegger, C., Ilyina, T., Jahns, T., Jimenez-de-la-Cuesta, D., Jungclaus, J., Kleinen, T., Kloster, S., Kracher, D., Kinne, S., Kleberg, D., Lasslop, G., Kornblueh, L., Marotzke, J., Matei, D., Meraner, K., Mikolajewicz, U., Modali, K., Mobis, B., Muller, W. A., Nabel, J. E. M. S., Nam, C. C. W., Notz, D., Nyawira, S. S., Paulsen, H., Peters, K., Pincus, R., Pohlmann, H., Pongratz, J., Popp, M., Raddatz, T. J., Rast, S., Redler, R., Reick, C. H., Rohrschneider, T., Schemann, V., Schmidt, H., Schnur, R., Schulzweida, U., Six, K. D., Stein, L., Stemmler, I., Stevens, B., von Storch, J. S., Tian, F. X., Voigt, A., Vrese, P., Wieners, K. H., Wilkenskjeld, S., Winkler, A., and Roeckner, E.: Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM1.2) and Its Response to Increasing CO2, J Adv Model Earth Sy, 11, 998-1038, 2019.

Meehl, G. A., Senior, C. A., Eyring, V., Flato, G., Lamarque, J.-F., Stouffer, R. J., Taylor, K. E., and Schlund, M.: Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models, Sci Adv, 6, eaba1981, 2020.

Meehl, G. A., Washington, W. M., Arblaster, J. M., Hu, A. X., Teng, H. Y., Tebaldi, C., Sanderson, B. N., Lamarque, J. F., Conley, A., Strand, W. G., and White, J. B.: Climate System Response to External Forcings and Climate Change Projections in CCSM4, J Climate, 25, 3661-3683, 2012.

Morice, C. P., Kennedy, J. J., Rayner, N. A., and Jones, P. D.: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set, J Geophys Res-Atmos, 117, 2012.

Muller, W. A., Jungclaus, J. H., Mauritsen, T., Baehr, J., Bittner, M., Budich, R., Bunzel, F., Esch, M., Ghosh, R., Haak, H., Ilyina, T., Kleine, T., Kornblueh, L., Li, H., Modali, K., Notz, D., Pohlmann, H., Roeckner, E., Stemmler, I., Tian, F., and Marotzke, J.: A Higher-resolution Version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR), J Adv Model Earth Sy, 10, 1383-1413, 2018.

Park, S., Shin, J., Kim, S., Oh, E., and Kim, Y.: Global Climate Simulated by the Seoul National University Atmosphere Model Version 0 with a Unified Convection Scheme (SAM0-UNICON), J Climate, 32, 2917-2949, 2019.

Po-Chedley, S., Proistosescu, C., Armour, K. C., and Santer, B. D.: Climate constraint reflects forced signal, Nature, 563, E6-E9, 2018.

Rackow, T., Goessling, H. F., Jung, T., Sidorenko, D., Semmler, T., Barbi, D., and Handorf, D.: Towards multi-resolution global climate modeling with ECHAM6-FESOM. Part II: climate variability, Clim Dynam, 50, 2369-2394, 2018.

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., and Kaplan, A.: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, J Geophys Res-Atmos, 108, 2003.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195-204, 2019.

Renoult, M., Annan, J. D., Hargreaves, J. C., Sagoo, N., Flynn, C., Kapsch, M. L., Mikolajewicz, U., Ohgaito, R., and Mauritsen, T.: A Bayesian framework for emergent constraints: case studies of climate sensitivity with PMIP, Clim. Past Discuss., 2020, 1-29, 2020.

Riahi, K., Rao, S., Krey, V., Cho, C. H., Chirkov, V., Fischer, G., Kindermann, G., Nakicenovic, N., and Rafaj, P.: RCP 8.5-A scenario of comparatively high greenhouse gas emissions, Climatic Change, 109, 33-57, 2011.

Righi, M., Andela, B., Eyring, V., Lauer, A., Predoi, V., Schlund, M., Vegas-Regidor, J., Bock, L., Brotz, B., de Mora, L., Diblen, F., Dreyer, L., Drost, N., Earnshaw, P., Hassler, B., Koldunov, N., Little, B., Tomas, S. L., and Zimmermann, K.: Earth System Model Evaluation Tool (ESMValTool) v2.0-technical overview, Geosci Model Dev, 13, 1179-1199, 2020.

Rind, D., Orbe, C., Jonas, J., Nazarenko, L., Zhou, T., Kelley, M., Lacis, A., Shindell, D., Faluvegi, G., Romanou, A., Russell, G., Tausnev, N., Bauer, M., and Schmidt, G.: GISS Model E2.2: A Climate Model Optimized for the Middle Atmosphere—Model Structure, Climatology, Variability, and Climate Sensitivity, 125, e2019JD032204, 2020.

~~Roe, G. H. and Baker, M. B.: Why is climate sensitivity so unpredictable?, Science, 318, 629-632, 2007.~~

Rong, X. Y., Li, J., Chen, H. M., Xin, Y. F., Su, J. Z., Hua, L. J., Zhou, T. J., Qi, Y. J., Zhang, Z. Q., Zhang, G., and Li, J. D.: The CAMS Climate System Model and a Basic Evaluation of Its Climatology and Climate Variability Simulation, J Meteorol Res-Prc, 32, 839-861, 2018.

Rossow, W. B. and Schiffer, R. A.: Isccp Cloud Data Products, B Am Meteorol Soc, 72, 2-20, 1991.

Rotstayn, L. D., Jeffrey, S. J., Collier, M. A., Dravitzki, S. M., Hirst, A. C., Syktus, J. I., and Wong, K. K.: Aerosol- and greenhouse gas-induced changes in summer rainfall and circulation in the Australasian region: a study using single-forcing climate simulations, Atmos Chem Phys, 12, 6377-6404, 2012.

Rugenstein, M., Bloch-Johnson, J., Abe-Ouchi, A., Andrews, T., Beyerle, U., Cao, L., Chadha, T., Danabasoglu, G., Dufresne, J. L., Duan, L., Foujols, M. A., Frolicher, T., Geoffroy, O., Gregory, J., Knutti, R., Li, C., Marzocchi, A., Mauritsen, T., Menary, M., Moyer, E., Nazarenko, L., Paynter, D., Saint-Martin, D., Schmidt, G. A., Yamamoto, A., and Yang, S. T.: LongRunMIP: Motivation and Design for a Large Collection of Millennial-Length AOGCM Simulations, B Am Meteorol Soc, 100, 2551-2570, 2019.

Rugenstein, M., Bloch-Johnson, J., Gregory, J., Andrews, T., Mauritsen, T., Li, C., Frolicher, T. L., Paynter, D., Danabasoglu, G., Yang, S. T., Dufresne, J. L., Cao, L., Schmidt, G. A., Abe-Ouchi, A., Geoffroy, O., and Knutti, R.: Equilibrium Climate Sensitivity Estimated by Equilibrating Climate Models, Geophys Res Lett, 47, 2020.

Rypdal, M., Fredriksen, H. B., Rypdal, K., and Steene, R. J.: Emergent constraints on climate sensitivity, Nature, 563, E4-E5, 2018.

Sanderson, B. M., Knutti, R., and Caldwell, P.: A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble, J Climate, 28, 5171-5194, 2015.

Sanderson, B. M., Wehner, M., and Knutti, R.: Skill and independence weighting for multi-model assessments, Geosci Model Dev, 10, 2379-2395, 2017.

Schlund, M., Eyring, V., Camps-Valls, G., Friedlingstein, P., Gentine, P., and Reichstein, M.: Constraining uncertainty in projected gross primary production with machine learning, J Geophys Res-Biogeo, ~~in review. in review~~accepted. accepted.

Schmidt, G. A., Ruedy, R., Hansen, J. E., Aleinov, I., Bell, N., Bauer, M., Bauer, S., Cairns, B., Canuto, V., Cheng, Y., Del Genio, A., Faluvegi, G., Friend, A. D., Hall, T. M., Hu, Y. Y., Kelley, M., Kiang, N. Y., Koch, D., Lacis, A. A., Lerner, J., Lo, K. K., Miller, R. L., Nazarenko, L., Oinas, V., Perlwitz, J., Perlwitz, J., Rind, D., Romanou, A., Russell, G. L., Sato, M., Shindell, D. T., Stone, P. H., Sun, S., Tausnev, N., Thresher, D., and Yao, M. S.: Present-day atmospheric simulations using GISS ModelE: Comparison to in situ, satellite, and reanalysis data, J Climate, 19, 153-192, 2006.

Séférian, R., Nabat, P., Michou, M., Saint-Martin, D., Voldoire, A., Colin, J., Decharme, B., Delire, C., Berthet, S., Chevallier, M., Sénési, S., Franchisteguy, L., Vial, J., Mallet, M., Joetzjer, E., Geoffroy, O., Guérémy, J.-F., Moine, M.-P., Msadek, R., Ribes, A., Rocher, M., Roehrig, R., Salas-y-Mélia, D., Sanchez, E., Terray, L., Valcke, S., Waldman, R., Aumont, O., Bopp, L., Deshayes, J., Éthé, C., and Madec, G.: Evaluation of CNRM Earth System Model, CNRM-ESM2-1: Role of Earth System Processes in Present-Day and Future Climate, 11, 4182-4227, 2019.

Seland, Ø., Bentsen, M., Seland Graff, L., Olivié, D., Toniazzo, T., Gjermundsen, A., Debernard, J. B., Gupta, A. K., He, Y., Kirkevåg, A., Schwinger, J., Tjiputra, J., Schancke Aas, K., Bethke, I., Fan, Y., Griesfeller, J., Grini, A., Guo, C., Ilicak, M., Hafsahl Karset, I. H., Landgren, O., Liakka, J., Onsum Moseid, K., Nummelin, A., Spensberger, C., Tang, H., Zhang, Z., Heinze, C., Iverson, T., and Schulz, M.: The Norwegian Earth System Model, NorESM2 - Evaluation of theCMIP6 DECK and historical simulations, Geosci. Model Dev. Discuss., 2020, 1-68, 2020.

Sellar, A. A., Jones, C. G., Mulcahy, J. P., Tang, Y. M., Yool, A., Wiltshire, A., O'Connor, F. M., Stringer, M., Hill, R., Palmieri, J., Woodward, S., de Mora, L., Kuhlbrodt, T., Rumbold, S. T., Kelley, D. I., Ellis, R., Johnson, C. E., Walton, J., Abraham, N. L., Andrews, M. B., Andrews, T., Archibald, A. T., Berthou, S., Burke, E., Blockley, E., Carslaw, K., Dalvi, M., Edwards, J., Folberth, G. A., Gedney, N., Griffiths, P. T., Harper, A. B., Hendry, M. A., Hewitt, A. J., Johnson, B., Jones, A., Jones, C. D., Keeble, J., Liddicoat, S., Morgenstern, O., Parker, R. J., Predoi, V., Robertson, E., Siahaan, A., Smith, R. S., Swaminathan, R., Woodhouse, M. T., Zeng, G., and Zerroukat, M.: UKESM1: Description and Evaluation of the UK Earth System Model, J Adv Model Earth Sy, 11, 4513-4558, 2019.

Sherwood, S., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., Hegerl, G., Klein, S. A., Marvel, K. D., Rohling, E. J., Watanabe, M., Andrews, T., Braconnot, P., Bretherton, C. S., Foster, G. L., Hausfather, Z., Heydt, A. S. v. d., Knutti, R., Mauritsen, T., Norris, J. R., Proistosescu, C., Rugenstein, M., Schmidt, G. A., Tokarska, K. B., and Zelinka, M. D.: An assessment of Earth's climate sensitivity using multiple lines of evidence, Reviews of Geophysics, accepted, 2020.

Sherwood, S. C., Bony, S., and Dufresne, J. L.: Spread in model climate sensitivity traced to atmospheric convective mixing, Nature, 505, 37-+, 2014.

Sidorenko, D., Rackow, T., Jung, T., Semmler, T., Barbi, D., Danilov, S., Dethloff, K., Dorn, W., Fieg, K., Goessling, H., Handorf, D., Harig, S., Hiller, W., Juricke, S., Losch, M., Schroter, J., Sein, D. V., and Wang, Q.: Towards multi-resolution global climate modeling with ECHAM6-FESOM. Part I: model formulation and mean climate, Clim Dynam, 44, 757-780, 2015.

Smith, T. M. and Reynolds, R. W.: Extended reconstruction of global sea surface temperatures based on COADS data (1854-1997), J Climate, 16, 1495-1510, 2003.

Su, H., Jiang, J. H., Zhai, C. X., Shen, T. J., Neelin, J. D., Stephens, G. L., and Yung, Y. L.: Weakening and strengthening structures in the Hadley Circulation change under global warming and implications for cloud response and climate sensitivity, J Geophys Res-Atmos, 119, 5787-5805, 2014.

Swart, N. C., Cole, J. N. S., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., Anstey, J., Arora, V., Christian, J. R., Hanna, S., Jiao, Y. J., Lee, W. G., Majaess, F., Saenko, O. A., Seiler, C., Seinen, C., Shao, A., Sigmond, M., Solheim, L., von Salzen, K., Yang, D., and Winter, B.: The Canadian Earth System Model version 5 (CanESM5.0.3), Geosci Model Dev, 12, 4823-4873, 2019.

Tatebe, H., Ogura, T., Nitta, T., Komuro, Y., Ogochi, K., Takemura, T., Sudo, K., Sekiguchi, M., Abe, M., Saito, F., Chikira, M., Watanabe, S., Mori, M., Hirota, N., Kawatani, Y., Mochizuki, T., Yoshimura, K., Takata, K., O'ishi, R., Yamazaki, D., Suzuki, T., Kurogi, M., Kataoka, T., Watanabe, M., and Kimoto, M.: Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6, Geosci Model Dev, 12, 2727-2765, 2019.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of Cmip5 and the Experiment Design, B Am Meteorol Soc, 93, 485-498, 2012.

Tian, B. J.: Spread of model climate sensitivity linked to double-Intertropical Convergence Zone bias, Geophys Res Lett, 42, 4133-4141, 2015.

Tokarska, K. B., Stolpe, M. B., Sippel, S., Fischer, E. M., Smith, C. J., Lehner, F., and Knutti, R.: Past warming trend constrains future warming in CMIP6 models, Sci Adv, 6, 2020.

Vial, J., Dufresne, J. L., and Bony, S.: On the interpretation of inter-model spread in CMIP5 climate sensitivity estimates, Clim Dynam, 41, 3339-3362, 2013.

Voldoire, A., Saint-Martin, D., Sénési, S., Decharme, B., Alias, A., Chevallier, M., Colin, J., Guérémy, J.-F., Michou, M., Moine, M.-P., Nabat, P., Roehrig, R., Salas y Mélia, D., Séférian, R., Valcke, S., Beau, I., Belamari, S., Berthet, S., Cassou, C., Cattiaux, J., Deshayes, J., Douville, H., Ethé, C., Franchistéguy, L., Geoffroy, O., Lévy, C., Madec, G., Meurdesoif, Y., Msadek, R., Ribes, A., Sanchez-Gomez, E., Terray, L., and Waldman, R.: Evaluation of CMIP6 DECK Experiments With CNRM-CM6-1, 11, 2177-2213, 2019.

Voldoire, A., Sanchez-Gomez, E., Melia, D. S. Y., Decharme, B., Cassou, C., Senesi, S., Valcke, S., Beau, I., Alias, A., Chevallier, M., Deque, M., Deshayes, J., Douville, H., Fernandez, E., Madec, G., Maisonnave, E., Moine, M. P., Planton, S., Saint-Martin, D., Szopa, S., Tyteca, S., Alkama, R., Belamari, S., Braun, A., Coquart, L., and Chauvin, F.: The CNRM-CM5.1 global climate model: description and basic evaluation, Clim Dynam, 40, 2091-2121, 2013.

Volodin, E. M.: Relation between temperature sensitivity to doubled carbon dioxide and the distribution of clouds in current climate models, Izv Atmos Ocean Phy+, 44, 288-299, 2008.

Volodin, E. M., Dianskii, N. A., and Gusev, A. V.: Simulating present-day climate with the INMCM4.0 coupled model of the atmospheric and oceanic general circulations, Izv Atmos Ocean Phy+, 46, 414-431, 2010.

Volodin, E. M., Mortikov, E. V., Kostrykin, S. V., Galin, V. Y., Lykosov, V. N., Gritsun, A. S., Diansky, N. A., Gusev, A. V., and Yakovlev, N. G.: Simulation of modern climate with the new version of the INM RAS climate model, Izv Atmos Ocean Phy+, 53, 142-155, 2017a.

Volodin, E. M., Mortikov, E. V., Kostrykin, S. V., Galin, V. Y., Lykossov, V. N., Gritsun, A. S., Diansky, N. A., Gusev, A. V., and Iakovlev, N. G.: Simulation of the present-day climate with the climate model INMCM5, Clim Dynam, 49, 3715-3734, 2017b.

Wang, T., Jiang, J., Zhang, M., Zhang, H., He, J., Hao, H., and Chi, X.: Design and Research of CAS-CIG for Earth System Models, Earth Space Sci, 7, 2020.

Watanabe, M., Suzuki, T., O'ishi, R., Komuro, Y., Watanabe, S., Emori, S., Takemura, T., Chikira, M., Ogura, T., Sekiguchi, M., Takata, K., Yamazaki, D., Yokohata, T., Nozawa, T., Hasumi, H., Tatebe, H., and Kimoto, M.: Improved Climate Simulation by MIROC5. Mean States, Variability, and Climate Sensitivity, J Climate, 23, 6312-6335, 2010.

Watanabe, S., Hajima, T., Sudo, K., Nagashima, T., Takemura, T., Okajima, H., Nozawa, T., Kawase, H., Abe, M., Yokohata, T., Ise, T., Sato, H., Kato, E., Takata, K., Emori, S., and Kawamiya, M.: MIROC-ESM 2010: model description and basic results of CMIP5-20c3m experiments, Geosci Model Dev, 4, 845-872, 2011.

Webb, M. J., Andrews, T., Bodas-Salcedo, A., Bony, S., Bretherton, C. S., Chadwick, R., Chepfer, H., Douville, H., Good, P., Kay, J. E., Klein, S. A., Marchand, R., Medeiros, B., Siebesma, A. P., Skinner, C. B., Stevens, B., Tselioudis, G., Tsushima, Y., and Watanabe, M.: The Cloud Feedback Model Intercomparison Project (CFMIP) contribution to CMIP6, Geosci Model Dev, 10, 359-384, 2017.

Williams, K. D., Copsey, D., Blockley, E. W., Bodas-Salcedo, A., Calvert, D., Comer, R., Davis, P., Graham, T., Hewitt, H. T., Hill, R., Hyder, P., Ineson, S., Johns, T. C., Keen, A. B., Lee, R. W., Megann, A., Milton, S. F., Rae, J. G. L., Roberts, M. J., Scaife, A. A., Schiemann, R., Storkey, D., Thorpe, L., Watterson, I. G., Walters, D. N., West, A., Wood, R. A., Woollings, T., and Xavier, P. K.: The Met Office Global Coupled Model 3.0 and 3.1 (GC3.0 and GC3.1) Configurations, J Adv Model Earth Sy, 10, 357-380, 2018.

Wu, T. W., Lu, Y. X., Fang, Y. J., Xin, X. G., Li, L., Li, W. P., Jie, W. H., Zhang, J., Liu, Y. M., Zhang, L., Zhang, F., Zhang, Y. W., Wu, F. H., Li, J. L., Chu, M., Wang, Z. Z., Shi, X. L., Liu, X. W., Wei, M., Huang, A. N., Zhang, Y. C., and Liu, X. H.: The Beijing Climate Center Climate System Model (BCC-CSM): the main progress from CMIP5 to CMIP6, Geosci Model Dev, 12, 1573-1600, 2019.

Wu, T. W., Song, L. C., Li, W. P., Wang, Z. Z., Zhang, H., Xin, X. G., Zhang, Y. W., Zhang, L., Li, J. L., Wu, F. H., Liu, Y. M., Zhang, F., Shi, X. L., Chu, M., Zhang, J., Fang, Y. J., Wang, F., Lu, Y. X., Liu, X. W., Wei, M., Liu, Q. X., Zhou, W. Y., Dong, M., Zhao, Q. G., Ji, J. J., Li, L., and Zhou, M. Y.: An Overview of BCC Climate System Model Development and Application for Climate Change Studies, J Meteorol Res-Prc, 28, 34-56, 2014.

Wyser, K., van Noije, T., Yang, S. T., von Hardenberg, J., O'Donnell, D., and ~~Döscher~~Doscher, R.: On the increased climate sensitivity in the EC-Earth model from CMIP5 to CMIP6, Geosci~~.~~ Model Dev~~. Discuss., 2019, 1-~~, 13, ~~2019~~3465-3474, 2020.

Yukimoto, S., Adachi, Y., Hosaka, M., Sakami, T., Yoshimura, H., Hirabara, M., Tanaka, T. Y., Shindo, E., Tsujino, H., Deushi, M., Mizuta, R., Yabu, S., Obata, A., Nakano, H., Koshiro, T., Ose, T., and Kitoh, A.: A New Global Climate Model of the Meteorological Research Institute: MRI-CGCM3-Model Description and Basic Performance, J Meteorol Soc Jpn, 90a, 23-64, 2012.

Yukimoto, S., Kawai, H., Koshiro, T., Oshima, N., Yoshida, K., Urakawa, S., Tsujino, H., Deushi, M., Tanaka, T., Hosaka, M., Yabu, S., Yoshimura, H., Shindo, E., Mizuta, R., Obata, A., Adachi, Y., and Ishii, M.: The Meteorological Research Institute Earth System Model Version 2.0, MRI-ESM2.0: Description and Basic Evaluation of the Physical Component, J Meteorol Soc Jpn, 97, 931-965, 2019.

Zelinka, M. D., Myers, T. A., Mccoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., Klein, S. A., and Taylor, K. E.: Causes of Higher Climate Sensitivity in CMIP6 Models, Geophys Res Lett, 47, 2020.

Zhai, C. X., Jiang, J. H., and Su, H.: Long-term cloud change imprinted in seasonal cloud variation: More evidence of high climate sensitivity, Geophys Res Lett, 42, 8729-8737, 2015.

Ziehn, T., Lenton, A., Law, R. M., Matear, R. J., and Chamberlain, M. A.: The carbon cycle in the Australian Community Climate and Earth System Simulator (ACCESS-ESM1) - Part 2: Historical simulations, Geosci Model Dev, 10, 2591-2614, 2017.