HORIZON 2020
THEME SC5-2017

EUCP
European Climate Prediction system

(GRANT AGREEMENT 776613)

European Climate Prediction system (EUCP)

Deliverable D2.2

Evaluation report on different methods to produce UQs/PDFs

Deliverable Title	*Deliverable 2.2: Quantifying uncertainty in projections of future European climate: a multi-model multi-method approach.*	
Brief Description	*Evaluation of the different methods to produce uncertainty quantifications/PDFs*	
WP number	2	
Lead Beneficiary	*Ben Booth, Met Office*	
Contributors	*Carol McSweeney, Met Office* *Lukas Brunner, ETH* *Ben Booth, Met Office* *Aurelien Ribes, CNRM* *Said Qasmi, CNRM* *Marianna Benassi, CMCC* *Silvio Gualdi, CMCC* *Antonio Navarra, CMCC* *Gabi Hegerl, University of Edinburgh* *Sabine Undorf, University of Edinburgh* *Andrew Ballinger, University of Edinburgh* *Reto Knutti, ETH* *Rita Noghero, ICTP* *Erika Coppola, ICTP* *Hylke de Vries, KNMI* *Glen Harris, Met Office* *Jason Lowe, Met Office* *Chris O'Reilly, University of Oxford* *Daniel Befort, University of Oxford*	
Creation Date	*14/10/2019*	
Version Number	*0.3*	
Version Date	*23/10/2019*	
Deliverable Due Date	*30/11/2019*	
Actual Delivery Date	*22/11/2020 (post publication date of associated peer reviewed publication)*	
Nature of the Deliverable	x	*R – Report*
		P - Prototype
		D - Demonstrator
		O – Other
Dissemination Level/ Audience	x	*PU - Public*
		PP - Restricted to other programme participants, including the Commission services
		RE - Restricted to a group specified by the consortium, including the Commission services
		CO - Confidential, only for members of the consortium, including the Commission services

Contents

Supplement A: Published paper: '**Brunner et al. (2020) Comparing Methods to Constrain Future European Climate Projections Using a Consistent Framework'.**
Available at: https://doi.org/10.1175/JCLI-D-19-0953.1

Supplement B (attached): Detailed descriptions of methods to produce UQs/PDFs

Version	Date	Modified by	Comments
2.0	26/11/2019	Carol McSweeney	Minor updates in response to internal reviews
3.0	03/11/2020	Carol McSweeney	Updated to include published paper, updated paper citations.

1. Executive Summary

The project partners in Tasks 2.1 and 2.2. of EUCP WP2 have contributed towards an evaluation of use of observations to reduce uncertainty in future climate projections and an intercomparison of several approaches to quantifying uncertainty in future projections out to 40 years ahead. This forms the basis of a scientific paper (Supplement A) which constitutes the core of the extended deliverable 2.2. This deliverable document comprises: (i) this short report providing some context for the paper in terms of the wider project and it objectives, (ii) the draft paper (supplement A) and (ii) a more detailed description of each method (Supplement B).

This piece of work represents the contributions of both tasks 2.1 and tasks 2.2, which, in practice, are intrinsically linked and not suited to assessing separately. For this reason, the partners in tasks 2.1 and 2.2 have submitted this extended deliverable 2.2.

The paper is the first example of a direct multi-method comparison of a wide range of methods based on a common set of regions, variables, time periods and seasons. A number of the results that emerge from this analysis we expect to inform the upcoming IPCC sixth assessment report. The intercomparison indicates that the multi-model methods consistently project a median warming of 2-3 degrees by the 2050s for the combined European region, irrespective of the observational constraints and the projection methodology employed. This is an encouraging result, and points towards a level of increased confidence that the various methodologies are downweighting poor models due to common underlying biases.
A second headline message is that the different methods do vary much more significantly in the uncertainty ranges around those median temperature projections. This result has the most important implications for users of climate projections who are more 'risk averse' and need to know what the upper end of the range of plausible projections might be because some methods will generate more conservative estimates than others. For precipitation, we see some consensus on decreases in mean rainfall over Central and Mediterranean Europe. However, there remains a large range in magnitudes of that change across the methods, such that the spread of median estimates remains large.

This is an important piece of work because it allows us to (a) establish how different the uncertainty estimates might be from the different methods, and thus, assess how robust the messages to users about future climate uncertainty might be, and (b) explain the differences between the methods; such as the different observational constraints applied, the inherent assumptions made within each method, and the extent to which they account for a wide range of contributing uncertainties.

This information about the source of the uncertainty estimates will help us to make informed decisions about which methods might be most appropriate in different contexts, and to provide relevant guidance to users. This underpinning information will, importantly, help to shape the outputs of Task 2.2 and deliverable 2.3 ('Production of UQ/PDFs'). For example, the workpackage will have to consider whether to recommend one or two 'most suitable' methods or a synthesis of methods. Further planned work will include exploring whether applying a common verification method could provide some objective measure of method 'skill' to help to inform these decisions. Future work being planned under EUCP will also align with the decision maker community to consider how the information in multi-method projections might be used and interpreted in some real world case studies.

2. Project Objectives

WITH THIS DELIVERABLE, EUCP HAS CONTRIBUTED TO THE ACHIEVEMENT OF THE FOLLOWING OBJECTIVES (DESCRIPTION OF ACTION, SECTION 1.1):

No.	Objective	Yes	No
1	Develop an ensembles climate prediction system based on high-resolution climate models for the European region for the near-term (~1-40 years)		x
2	Use the climate prediction system to produce consistent, authoritative and actionable climate information	x	
3	Demonstrate the value of this climate prediction system through high impact extreme weather events in the near past and near future		x
4	Develop, and publish, methodologies, good practice and guidance for producing and using EUCP's authoritative climate predictions for 1-40 year timescales	x	

3. Detailed Report

3.1 Quantifying uncertainty in projections of future European climate: a multi-model multi-method approach.

The project partners in tasks 2.1 and 2.2. have contributed towards a joint comparison of multiple approaches to quantifying uncertainty in future projections out to 40 years ahead. This comparison forms the basis of a scientific paper which, in turn, forms the core of Deliverables 2.1 and 2.2.

The paper is the first example of a direct multi-method comparison of a wide range of methods based on a common set of regions, variables, time periods and seasons. This is an important piece of work because it allows us to (a) establish how different the uncertainty estimates might be from the different methods, and thus, assess how robust the messages to users about future climate uncertainty might be and (b) explain and discuss fundamental differences between the methods; such as the different observational constraints applied, the inherent assumptions made within each method, and the extent to which they account for a wide range of contributing uncertainties. This allows us to make informed decisions about which methods might be most appropriate in different contexts, and to provide relevant guidance to users.

The core results are presented within a scientific journal paper that we intend to submit to *Environmental Research Letters* (or a journal of equivalent impact level) in ahead of the IPCC deadline (December 2019). Delivering these results as part of a scientific publication aids the visibility of our results within the wider community working on European projections, and demonstrates European leadership in developing and deploying multiple approaches for those working in other parts of the world. A number of the results that emerge from this analysis we expect to inform the upcoming IPCC assessment report.

The paper explores several diverse methods for quantifying uncertainties in future climate. Six methods explore methods for quantifying uncertainties in future climate, of which five methods use the range of multi-model projections and one is based on calibration of a single model ensemble. A further two use large single model ensembles to quantify the spread due to internal variability alone, providing contextual information about the proportion of uncertainty in future changes that relates to internal varaibility rather than uncertainty in climate signal. All methods are applied (where scientifically feasible) to summer (June-July-August) average temperature and precipitation, under RCP8.5, for eight common domains at different spatial scales – the three large IPCC SREX domains covering Northern Europe, Central Europe and the Mediterannean, a large European region combining those three regions, and four smaller 'grid-point' scale domains.

The intercomparison indicates that the multi-model methods consistently indicate a median projection of 2-3 degrees by the 2050s for the combined European region, irrespective of the observational constraints and the projection methodology employed. This is an encouraging result, and points towards a level of increased confidence that the various methodologies are downweighting poor models due to common underlying biases. A second headline message is that the different methods imply larger spread in the inferred ranges around those median temperature projections. The widest uncertainty ranges stem from the UKCP bayesian approach due to its additional sampling of carbon cycle uncertainty drawn from perturbed parameter experiments over other methods. This result has the most important implications for users of climate projections who are more 'risk averse' and need to know what the upper end of the range of plausible projections might be because some methods will generate more conservative estimates than others.

For precipitation, we see some consensus on decreases in mean rainfall over Central and Mediterannean Europe. However, there remains a large range in magnitudes of that change across the methods, such that the spread of median estimates remains large. Further work is required to understand how the fundamental differences between these methods drive such different projection ranges. It is less clear at this stage which of the precipitation projection products would be suitable for release as projection products.

This multi-method comparison provides an important evaluation of methods that will help to inform the later tasks in Work Package 2. The next steps for WP2 partners are to extend their analysis to include additional timeperiods, season and variables, and look at case studies which illustrate greater role of internal variability (e.g. nearer time frames, smaller spatial scales). With respect to the additional variables we will look towards quantifying uncertainty in more user relevant metrics, such as common indices of extremes. The work package partners also plan to explore whether a common verification method could provide some objective measure of method 'skill'. This could possibly use CMIP6 models to provide an 'out of sample' cross validation. Further, the work package partners will consider how these methods might be usefully applied to outermost regions of the EU.

Planned future work under EUCP will also align with the decision maker community to consider how the information in multi-method projections might be used and interpreted in some real world case studies. It is clear from the multi-method comparison that that the methods differ in a number of ways beyond the quantitative comparison of the resulting projections, such as their different underpinning assumptions, levels of sophistication and different characteristics of the outputs such as their spatial and physical coherency. As we look towards the next deliverable under Task 2.2 (Deliverable 2.3) we plan to confront users with these methods, in order to improve our understanding of the implications and relative importance of these differences for those working with the projections.

3.2 Detailed descriptions of methods to produce UQs/PDFs

The methods applied by the WP2 partners in the intercomparison activity are described in more detail than the scientific paper allows in the additional appendices to this document. See Supplement B for these detailed descriptions of the methods.

3.3 Published peer reviewed papers

Brunner, L, Ruth Lorenz, Marius Zumwald, and Reto Knutti (2019). ''Quantifying uncertainty in European climate projections using combined performance-independence weighting'' *Environ. Res. Lett.* DOI: 10.1088/1748-9326/ab492f

Brunner, Lukas., Carol McSweeney; Andrew P. Ballinger; Daniel J. Befort; Marianna Benassi; Ben Booth; Erika Coppola; Hylke de Vries; Glen Harris; Gabriele C. Hegerl Reto Knutti; Geert Lenderink; Jason Lowe; Rita Nogherotto; Chris O'Reilly; Saïd Qasmi; Aurélien Ribes; Paolo Stocchi; Sabine Undorf (2020) Comparing Methods to Constrain Future European Climate Projections Using a Consistent Framework. *Journal of Climat*e 33(20):8671-8672. https://doi.org/10.1175/JCLI-D-19-0953.1

3.4 Papers in preparation related to this work

O'Reilly et al. (*in prep.*) Calibrating large ensemble projections using observational data.

Ribes A., S. Qasmi; N. Gillett (*In press*) Making climate projections conditional on historical observations. Science Advances.

Harris, G.R., J.M.Murphy, D.M.H.Sexton. B.B.B.Booth (*in preparation.)* Probabilistic projections for regional climate change accounting for Earth System modelling uncertainty.

De Vries, H. and Lenderink, G. *(in prep*.) Methods for estimating internal variability at different time scales in climate model ensembles.

3.5 Summary of progress towards objectives

This work represents significant progress towards two of the projects high-level objectives.

2. Use the climate prediction system to produce consistent, authoritative and actionable climate information

This comparison of methods addresses some key issues surrounding the question of how a climate prediction system could best provide climate information that is consistent, authoritative and actionable. The diversity of uncertainty quantification methods currently available means that different products provided could offer inconsistent information to users on the upper and lower ranges of plausible climate change in Europe under a given emissions scenario. It therefore represents a significant challenge to the work package to reconcile those different approaches and the differing results in order to ensure that an EUCP product captures the strengths of the different methods without compromising the consistency of the information. The results from this intercomparison work pave the way for exploring two important avenues. The first is to explore whether it is feasible to synthesize more consistent information from these multiple methods, using the information about strengths and weaknesses and the different results from each method. A 'consensus'

product might, for example, use the spread from multiple results, or might consist of a new method which draws on the strongest components of multiple methods.

The second is to improve the transparency of these methods by exposing differences between methods and the differences in result in order to improve the guidance on which datasets might be most appropriate in different user cases.

4. Develop, and publish, methodologies, good practice and guidance for producing and using EUCP's authoritative climate predictions for 1-40 year timescales

This piece of work represents an important basis for providing advice on good practice for producing and using probabilistic climate projections. The intercomparison will provide the evidence base for (a) the selection of methods, or synthesis products disseminated in EUCP in WP2 Task 2.3, as well as (b) forming the basis for good guidance materials by exposing the differences between different methods and their implications for different users.

4. Lessons Learnt and links Built

WP2 have encountered challenges as well and positive experiences in these early stages of EUCP.

- The intercomparison study in task 2.2 has proved to be a useful focus for the workpackage and has drawn together the collective understanding of these diverse methods and facilitated really strong engagement in the work package. This is a really positive 'lesson' from the work undertaken so far.
- CMIP6 data have not been available as early as was originally anticipated. While a small number modelling centres had already populated CMIP6 by early 2019, the UQ methods required a projections simulations from a significant number of models to be available much sooner than they have been. In the absence of this a substantive CMIP6 ensemble, the work package has been able to press forward with the intercomparison study to evaluate the different methods using CMIP5. Some groups have been able to test some methodology on an early subset of CMIP6, and several groups plan to extend their work to include CMIP6 within Task 2.3. Given that WP3 simulations are based on CMIP5-generation modelling, working with CMIP5 data in this early stage does have the benefit of bridging more directly to WP3.
- An important 'lesson learned' early in this workpackage is that the treatment of observational constraints and uncertainty quantification are inherently linked to one another. The original proposal had the expectation that the activities focused on identifying useful observational constraints could be a distinct and self-contained activity (Task 2.1). In practice, two main factors meant that assessment of observational constraints were intrinsically linked to their implementation within the climate projection methodologies. Firstly, any quantitative assessment of the impact of observations needed the implementation to compare projections with and without the selected observations. Secondly, many of the insights into the value of selected observations proved to be method specific, due to philosophical and practical choices made by individual projection methodologies. The valuable outcome, therefore, has not been two distinct assessments (firstly on the most useful observational constraints and secondly on

their implementation) but rather the comparison of methods and their observational constraints in terms of their combined impact on future projected changes. This is reflected in the merged nature of this deliverable.

There are number of areas where WP2 has begun to build strong links with other work packages.

- WP2 has engaged with WP3 on their selection of boundary conditions for their convective permitting regional climate modelling. In practice, pressures from short timescales to get simulations going, limited availability of global climate model boundary conditions in CMIP5, and workflows closely tied to institutional modelling capability meant that many of these decisions were made on pragmatic grounds rather than designs to optimally sample the range of projected future changes. WP2 was represented at the initial WP3 workshop where we provided context on which global climate models were likely to be screened out as less plausible. Longer term, WP2 will look to demonstrate methods to sub-select global driving data that could inform selection of boundary data for future regional modelling activities.

- A number of areas of discussion have developed around the role of WP2 in providing the broader uncertainty context for small downscaled ensemble subsets in WP3. This also has relevance for Task 5.5 in WP5, where various activities will explore cases where projection information might be used from different sources and the user may have to make choices, or use a combination of pieces of data. WP5 also deals with the exploration of uncertainty quantification methodologies in the context of merging intialised and un-initialised runs. The Callibration method (CALL) developed by partners at the University of Oxford jointly under WP2 and WP5 is designed to address some of the bias correction issues that need to be addressed in order to offer a merged product. A joint Work Package 2 and 5 workshop will be held on this topic in June 2020.

- Now that we have produced our initial WP2 climate projections initial conversations have started with uses of climate projection data within WP4. Peter Greve (IIASA) joined us for our WP2 workshop in SMHI in Sweden (Oct. 2019) to discuss how we can link this information through to their impact modelling, both at river and water modelling at IIASA and more widely in WP4. This exposed some interesting challenges. For example, the water modelling requires spatially and temporally coherent realizations to drive their impact modelling. This is not necessarily consistent with the projection probabilities described in this deliverable. We are starting to think of ways to bridge this gap, whether this is using probabilistic projections to provide context to WP4 impact modelling; employing weather generators to translate the probabilities into realizations; or identifying the subset of methodologies that can provide weights on individual realizations.

SUPPLEMENT B

Contents

B.1 Climate model weighting by independence and performance (ClimWIP) - ETHZ

Author: Lukas Brunner and Ruth Lorenz

Short summary of method

A weighting method based on model performance and independence is used to better quantity the uncertainty and increase the skill of climate projections in Europe. The weights for each model are based on a pool of diagnostics which represent the its distance several observational data sets on the one hand and the distances to all other models on the other hand. This approach can be easily be applied to different of target variables, time periods, and geographical regions, depending on the availability of observations to inform the weighting.

Key References

Brunner *et al.*, 2019: *This paper includes a detailed method description and presents results based on the common WP2 settings [EUCP WP2 publication]*

Lorenz *et al.*, 2018: *This paper investigates the effect of the weighting method applied to maximum temperature in North America and describes the selection of diagnostics to inform the weighting.*

Knutti *et al.*, 2017: *This paper looks into the effect of the weighting method using the case of Arctic sea ice extent.*

Data

The weighting is applied to all available CMIP5 (Taylor, Stouffer and Meehl, 2012) models and initial-condition members. We use monthly data regridded to a regular 2.5°×2.5° grid and combine historical runs and the representative concentration pathway 8.5 (RCP8.5) (van Vuuren *et al.*, 2011) in the period 1995-2014 to initialize our weighting approach and RCP8.5 from 2041-2060 as target period.

As reference we use reanalyses data from the European Centre for Medium-Range Weather Forecasts' (ECMWF) ERA-Interim (Dee *et al.*, 2011) and from the National Aeronautics and Space Administrations (NASA) MERRA2 (Molod *et al.*, 2015) as well as a data set based purely on observations, which is combined from multiple sources for different variables (E-OBS, Cornes *et al.* (2018); CERES, (Wielicki B. A. *et al.*, 1996))

Details of method

The method is based on a combination of historical model performance and model independence. To inform the performance weighting we use six diagnostics, which can be based on any CMIP5 output variable for which observations are available. The calculation of a diagnostic follows a straight-forward approach: (i) variable, region, time period, and season are selected, (ii) the climatology (CLIM) or standard deviation (STD) is calculated, (iii) the point-to-point difference

between model and the observational spread is calculated, and finally (iv) the area weighted root mean squared error is calculated over the selected region.

The independence weighting is informed by diagnostics which can be based on any CMIP5 output variable which is available for all models. In practice, we use the same diagnostics as for the performance weighting, with the only difference that the point-to-point difference (step (iii) above) is calculated between each model pair.
The weights are calculated following an approach used by Lorenz *et al.* (2018),which is in turn based on the work by Knutti *et al.* (2017), Sanderson, Knutti and Caldwell (2015a), and Sanderson, Knutti and Caldwell (2015b). Each weight w_i is a combination of the observational distance parameter D_i (informing the performance weighting) and the model distance parameter S_{ij} (informing the independence weighting):

$$w_i = \frac{e^{\frac{D_i}{\sigma_D}}}{1 + \sum_{j \neq i}^{M} e^{\frac{S_{ij}}{\sigma_S}}}$$

with the total number of model runs M and the shape parameters σ_D and σ_S. For details on the estimation of these shape parameters see Lorenz *et al.* (2018) and Brunner *et al.* (2019).

Observational constraints applied

When discussed in the context of observational constraints model weighting schemes take a somewhat special role. They use multiple observation-based variables, sometimes termed *diagnostics* to inform the weights. Ideally, informing diagnostisc should show a strong relationship between the simulated historical values and the target which one tried to predict. In a classical emergent constraints setting adding additional constraints usually leads to a strong sub-selection of models and even none of the models being consistent with the applied constraints any more. In contrast, weighting approaches tend to have an ideal number of diagnostics to inform the weighting. Using less diagnostics leads to very strong weighting and is therefore prone to overconfidence while using many diagnostics converges towards equal weighting. (Lorenz *et al.*, 2018). A discussion of the differences between emergent constraints and model weighting can be found, for example, in Alex Hall *et al.*, (2019).
Sanderson, Wehner and Knutti, (2017) use the seasonal climatologies of 12 variables to inform their weights. These variables include basic ones such as temperature, precipitation, radiation fluxes, and pressure as well as more specific ones, such as coldest and warmest days and nights (see Sanderson, Wehner and Knutti (2017), Table 1 for a full overview). The weights are then applied to projections of temperature and precipitation in North America. Lorenz *et al.* (2018) use a very similar approach and apply model weighting to maximum temperature in North America as well as a sub-region. From a pool of over 20 possible diagnostics (including climatologies, variances, and trends) they select the most informative nine based on correlations between historical values and change in the target variable. Among them are maximum temperature (climatology, variance, and trend), precipitation (climatology and variance), and surface humidity (variance) (see Lorenz *et al.* (2018), Table 1 for a full overview).

More specifically, Knutti *et al.* (2017) focus on September sea ice extend in the Arctic. Four diagnostics are selected based on expert knowledge: the September sea ice extend climatology and trend as well as the climatology and variability in the surface air temperature.

Here, ETHZ applies weights to constrain projections of summer and winter temperature and precipitation in different European regions as detailed in Brunner *et al.* (2019). Six diagnostics based are selected: temperature (climatology), precipitation (climatology), shortwave downward radiation (climatology), shortwave upward radiation (climatology and variance), longwave downward radiation (variance).

Key assumptions

- Future model performance can be inferred from historical model performance (based on a range of diagnostics) (e.g., Knutti (2008)).
- The dependence of models between each other can be estimated by their output (e.g., Knutti, Masson and Gettelman (2013)).
- Selection of the ideal diagnostics to inform the weighting as well as a reasonable number of diagnostics to use (e.g., Lorenz *et al.* (2018)).
- How to translate observation-model distance into model performance (performance shape parameter, how strongly do we weight for model performance) and how to translate the model-model distance into model independence (i.e., when are two model independent; independence shape parameter) (Knutti *et al.*, 2017).

Limitations

- Availability of gridded observations to produce diagnostics relevant for the target variable.
- Sufficient domain to average over and avoid the effect of natural variability on the weights.
- The observational spread (i.e., from different sources) needs to be smaller than the model spread.
- No extrapolation beyond original model range possible.

References

Alex Hall *et al.* (2019) 'Progressing emergent constraints on future climate change', *Nature Climate Change*. Springer US. doi: 10.1038/s41558-019-0436-6.

Brunner, L. *et al.* (2019) 'Quantifying uncertainty in European climate projections using combined performance-independence weighting', *Environmental Research Letters*. doi: 10.1088/1748-9326/ab492f.

Cornes, R. C. *et al.* (2018) 'An Ensemble Version of the E-OBS Temperature and Precipitation Datasets', *Journal of Geophysical Research: Atmospheres*, pp. 1–19. doi: 10.1029/2017JD028200.

Dee, D. P. *et al.* (2011) 'The ERA-Interim reanalysis: Configuration and performance of the data assimilation system', *Quarterly Journal of the Royal Meteorological Society*, 137(656), pp. 553–597. doi: 10.1002/qj.828.

Knutti, R. (2008) 'Should we believe model predictions of future climate change?', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 366(1885), pp. 4647–4664. doi: 10.1098/rsta.2008.0169.

Knutti, R. *et al.* (2017) 'A climate model projection weighting scheme accounting for performance and interdependence', *Geophysical Research Letters*, 44(4), pp. 1909–1918. doi: 10.1002/2016GL072012.

Knutti, R., Masson, D. and Gettelman, A. (2013) 'Climate model genealogy: Generation CMIP5 and how we got there', *Geophysical Research Letters*, 40(6), pp. 1194–1199. doi: 10.1002/grl.50256.

Lorenz, R. *et al.* (2018) 'Prospects and caveats of weighting climate models for summer maximum temperature projections over North America', *Journal of Geophysical Research: Atmospheres*. doi: 10.1029/2017JD027992.

Molod, A. *et al.* (2015) 'Development of the GEOS-5 atmospheric general circulation model: Evolution from MERRA to MERRA2', *Geoscientific Model Development*, 8(5), pp. 1339–1356. doi: 10.5194/gmd-8-1339-2015.

Sanderson, B. M., Knutti, R. and Caldwell, P. (2015a) 'A representative democracy to reduce interdependency in a multimodel ensemble', *Journal of Climate*, 28(13), pp. 5171–5194. doi: 10.1175/JCLI-D-14-00362.1.

Sanderson, B. M., Knutti, R. and Caldwell, P. (2015b) 'Addressing interdependency in a multimodel ensemble by interpolation of model properties', *Journal of Climate*, 28(13), pp. 5150–5170. doi: 10.1175/JCLI-D-14-00361.1.

Sanderson, B. M., Wehner, M. and Knutti, R. (2017) 'Skill and independence weighting for multi-model assessments', *Geoscientific Model Development*, 10(6), pp. 2379–2395. doi: 10.5194/gmd-10-2379-2017.

Taylor, K. E., Stouffer, R. J. and Meehl, G. A. (2012) 'An overview of CMIP5 and the experiment design', *Bulletin of the American Meteorological Society*, 93(4), pp. 485–498. doi: 10.1175/BAMS-D-11-00094.1.

van Vuuren, D. P. *et al.* (2011) 'The representative concentration pathways: An overview', *Climatic Change*, 109(1), pp. 5–31. doi: 10.1007/s10584-011-0148-z.

Wielicki B. A. *et al.* (1996) 'Clouds and the Earth's Radiant Energy System (CERES): An Earth Observing System Experiment', *Bulletin of the American Meteorological Society*, 77(5), pp. 853–868.

B.2 The Reliability Ensemble Averaging method (REA) - ICTP

Authors: Sticchi Paolo, Nogherotto Rita, Erika Coppola, and Filippo Giorgi

Short summary of the method

The weighting method is based on two "reliability" criteria: the model performance criterion in reproducing present day-climate and the model convergence criterion that consider the spread of the climate change signal across the models. The philosophy of the REA approach is to minimize the contribution of simulations that either perform poorly in the representation of present-day climate or are outlier in the ensemble for the future projections, in order to reduce the uncertainty and increase the skill of the climate model ensemble in Europe.

Key references

Giorgi, F., and L.O. Mearns, 2002: Calculation of average, uncertainty range and reliability of regional climate changes from AOGCM simulations via the "Reliability Ensemble Averaging (REA)" method. Journal of Climate, 15, 1141-1158.

Giorgi, F., and L.O. Mearns, 2003: Probability of regional climate change calculated using the Reliability Ensemble Averaging (REA) method. Geophysical Research Letters, 30, 1629.

Data

The weighting is applied to the available CMIP5 (Taylor, Stouffer and Meehl, 2012) models that we have in house. The monthly model data are regridded into a regular 2.5°×2.5° grid and we use the historical period 1995-2014 has a reference period to calculate the first reliability criteria and the mid of century future time slice period (2041-2060) for the scenario concentration pathway 8.5 (RCP8.5) (van Vuuren *et al.*, 2011). As observed reference data we used the E-OBS gridded dataset.

Details of method

Two general "reliability criteria" are used to assess the reliability of regional climate change ensemble simulations. The first is based on the ability of a climate model to reproduce different aspects of present-day climate: the better a model performance is, the higher the reliability of that model. We refer to this as the "model performance" criterion. The second criterion is based on the convergence of the different model simulations for a given forcing scenario. The greater it is the convergence the higher it is the reliability of the signal that is little sensitive to the model differences. We refer to this as the "model convergence" criterion. This method is called "reliability ensemble averaging" (REA).

In our REA method, the average change, $(\Delta\tilde{T})$ is given by a weighted average of the ensemble members, that is:

$$\tilde{A}(\Delta T) = (\Delta\tilde{T}) = \frac{\sum_i R_i \Delta T_i}{\sum R_i} \qquad (1)$$

where the operator $\tilde{A}$ denotes the REA averaging and R_i is a model reliability factor defined as

$$R_i = \left[\left(R_{B,i}\right)^m x \left(R_{D,i}\right)^n\right]^{[1/(mxn)]} = \left\{\left[\frac{T}{\text{abs}(B_{T,i})}\right]^m \left[\frac{T}{\text{abs}(D_{T,i})}\right]^n\right\}^{[1/(mxn)]} \quad (2)$$

$\boldsymbol{R_{B,i}}$ is a factor that measures the model reliability as a function of the model bias ($\boldsymbol{B_{T,i}}$) in simulating present-day value of a given variable. The higher the bias is the lower is the model reliability. Here the bias is defined as the difference between simulated and observed mean value of the variable for the present-day period. $\boldsymbol{R_{D,i}}$ is a factor that measures the model reliability in terms of the distance ($\boldsymbol{D_{T,i}}$) of the change calculated by a given model from the REA ensemble average change. The higher is the distance the lower is the model reliability. Therefore, the distance is a measure of the degree of convergence of a given model toward the average of the others. To summarize $\boldsymbol{R_{B,i}}$ is a measure of the model performance criterion while $\boldsymbol{R_{D,i}}$ is a measure of the model convergence criterion.

The distance $D_{T,i}$ is calculated using an iterative procedure. A first guess of $D_{T,i}$ is the distance of each ΔT_i from the ensemble average change that is, $[D_{T,i}]_1 = [\Delta T_i - \bar{\Delta T}]$ with

$$\bar{\Delta T} = \frac{1}{N}\sum \Delta T_i$$

The first guess values are then used in eq. (1) and eq. (2) to obtain a first-order REA average change $[\bar{\Delta T}]_1$, which is then used to recalculate the distance of each individual model as $\left[D_{T,i}\right]_2 = [\Delta T_i - [\bar{\Delta T}]_1]$ and repeat the iteration. Typically, this procedure converges quickly after several iterations.

The parameters *m* and *n* in eq. (2) can be used to give different weigh to each criterion. For most calculations *m* and *n* are assumed to be equal to 1, which gives equal weight to both criteria. Also, R_B and R_D are set to 1 when *B* and *D* are respectively smaller than ϵ. To summarize eq.(2) states that a model projection is "reliable" when both its bias and distance from the ensemble average are within the natural variability, so that $R_B = R_D = R = 1$. As the bias and/or distance grow, the reliability of a given model simulation decreases.

The parameter ϵ in eq. (2) is a measure of natural variability in 30-yr average regional temperature and precipitation.
In order to calculate ϵ, we consider the time series of observed regionally temperature and precipitation for the twentieth century from the E-OBS dataset. We then compute yearly averages of the time series after linearly detrending the data (to remove century-scale trends) and estimated ϵ as the variance of these 30-yearly averages.
The concept of reliability factor can be used to estimate the probability of future climate change from the model ensemble. Before doing that, we note that the likelihood that a given model-simulated change will actually happen is generally not known, since future conditions are not known and therefore the model cannot be validated in its ability to predict climate change. Also not known is the probability distribution of the simulated changes, since this would require a very large sample of model simulations. As a result, some assumptions need to be made concerning the likelihood of a model outcome.

In our method, the likelihood associated with a model-simulated change (Pm_i) is proportional to the reliability parameter defined in eq. (2). The normalization of this likelihood yields the definition

$$P_{m_i} = \frac{R_i}{\sum_{j=1}^{N} Rj} \qquad (3)$$

In other words we assume that the change simulated by a more reliable model is more likely to occur. From equation (3) it follows that, for a given emission scenario, the probability of a climate change exceeding a certain threshold ΔT_{th} is given by :

$$\Delta T_i >= \Delta T_{th} \qquad (4)$$

$$Pm_{DT_{th}} = \sum_i Pm_i = \sum_i \left[\frac{R_i}{\sum_{j=1}^{N} R_j} \right]$$

where $Pm_{\Delta Tth}$ is the probability of the temperature change being greater than the threshold ΔT_{th} in the scenario considered.

From the probability of exceedance we computed the Probability Density Function (PDF) as the difference between $Pm_{\Delta Tth} - Pm_{\Delta Tth-1}$ for the 2.5°x2.5° boxes centered on Svealand, Dusseldorf, Madrid and Transylvania (Fig. 1-2) and for the three SREX regions (NEU, CEU, MED) + the entire European domain (NEU-CEU-MED) (Fig. 3-6).
The changes are calculated for the period 2041-2060 relative to the 1995-2014 period.

Limitations

- The applicability of the method is limited by the availability of observations
- The method at the moment doesn't consider multiple observations and the spread of the observational dataset.

Results

We show here some preliminary results applying the method using the CMIP5 models reported in Table 1.

Table 1: CMIP5 models used in the REA method calculations.

BCC-CSM1.1	CSIRO-ACCESS-3	MIROC-ESM
BCC-CSM1.1M	CSIRO-MK36	MPI-ESM-LR
BNU-ESM	EC-EARTH	MPI-ESM-MR
CanESM2	FIO-ESM	NCAR-CCSM4
CMCC-CM	GFDL-ESM2M	NCAR-CESM1-BGCNCAR-CESM1-CAM5
CMCC-CMS	HadGEM2-ES	NorESM1-M
CSIRO-ACCESS-1	IPSL-CM5A-MR	

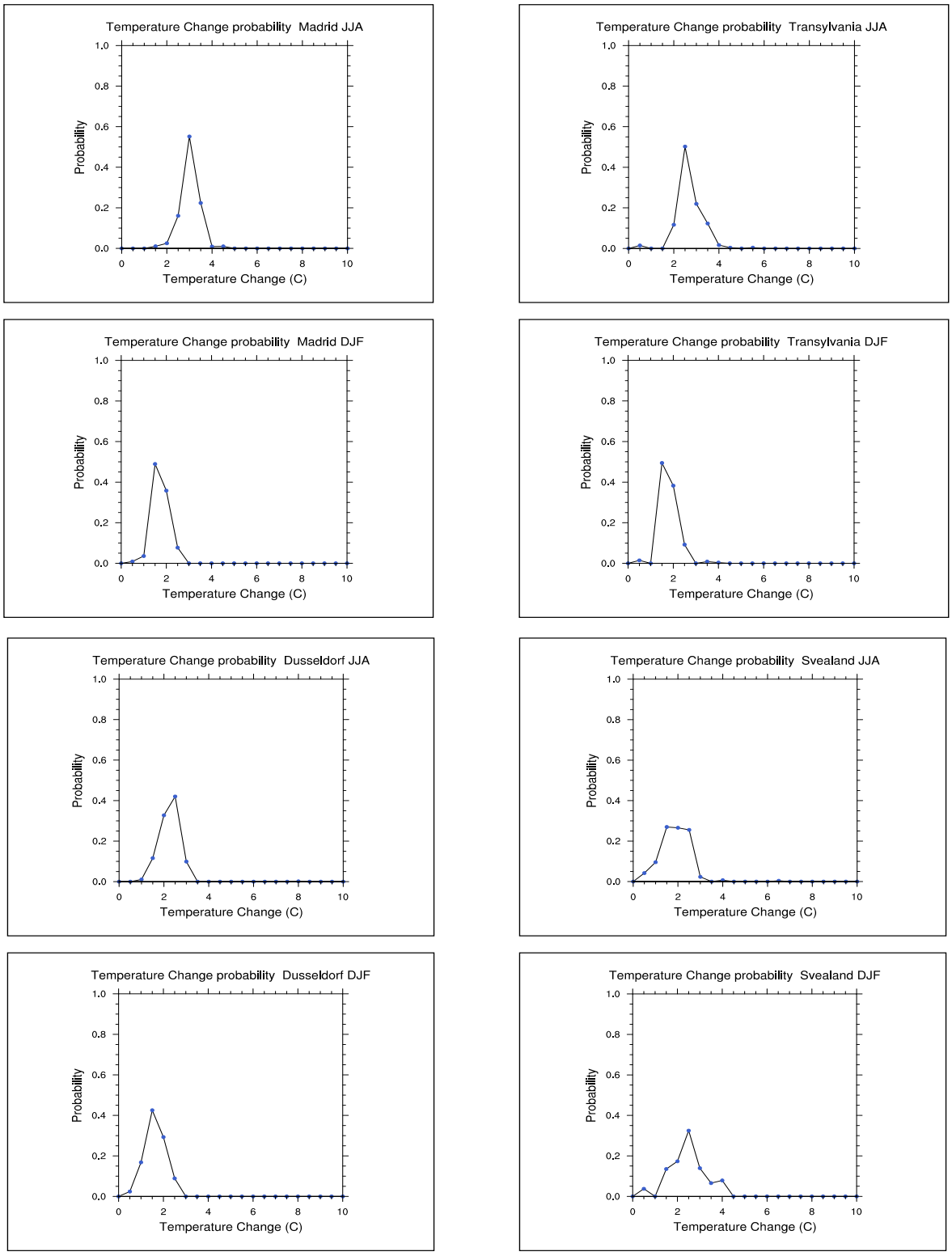


Figure 1: Temperature Change (°C) probability density functions for the four small boxes centered on Dusseldorf, Svealand, Madrid and Transylvania for the two seasons (JJA, DJF).

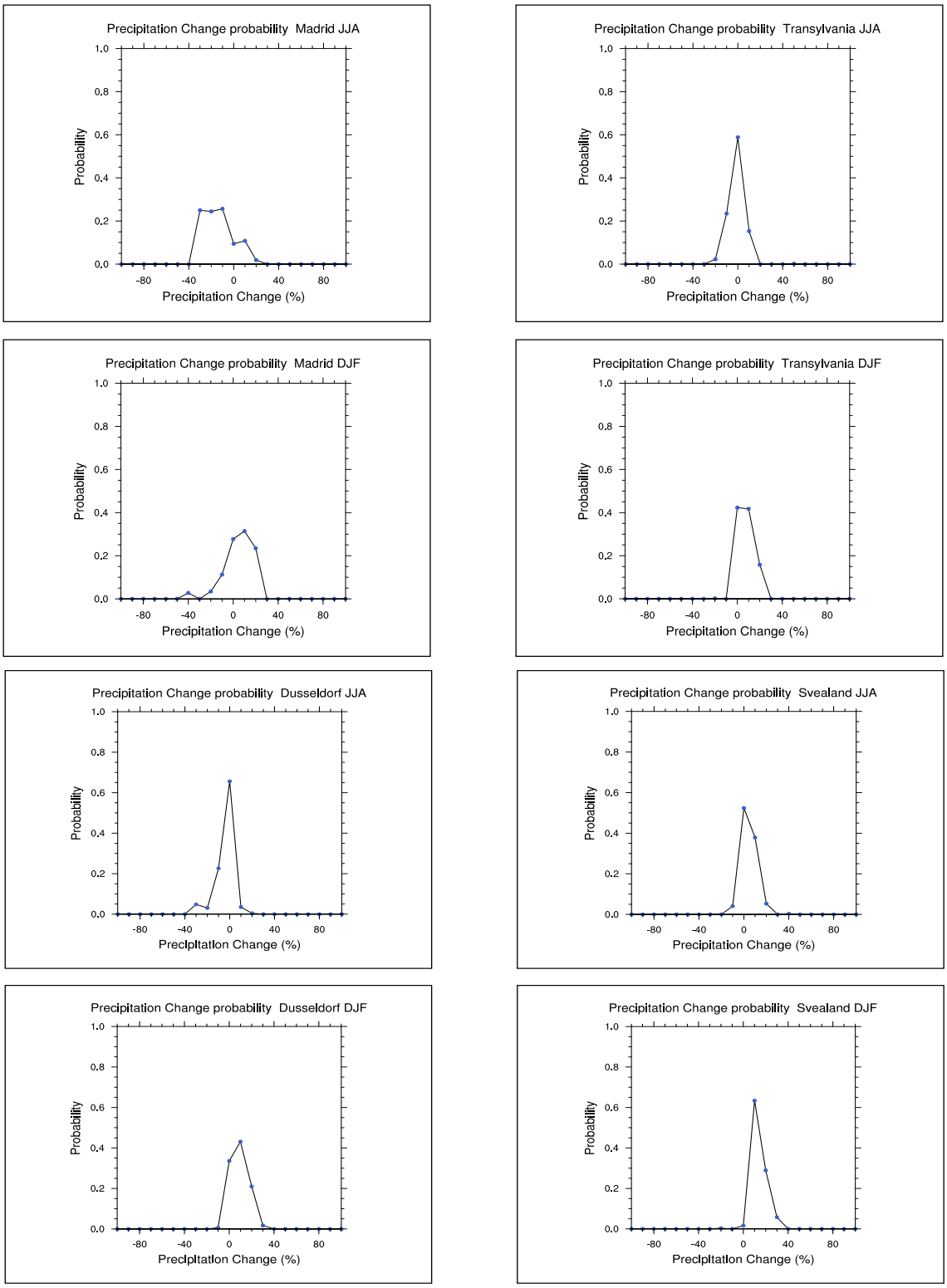


Figure 2: Precipitation Change (%) probability density functions for the four small boxes centered on Dusseldorf, Svealand, Madrid and Transylvania for the two seasons (JJA, DJF).

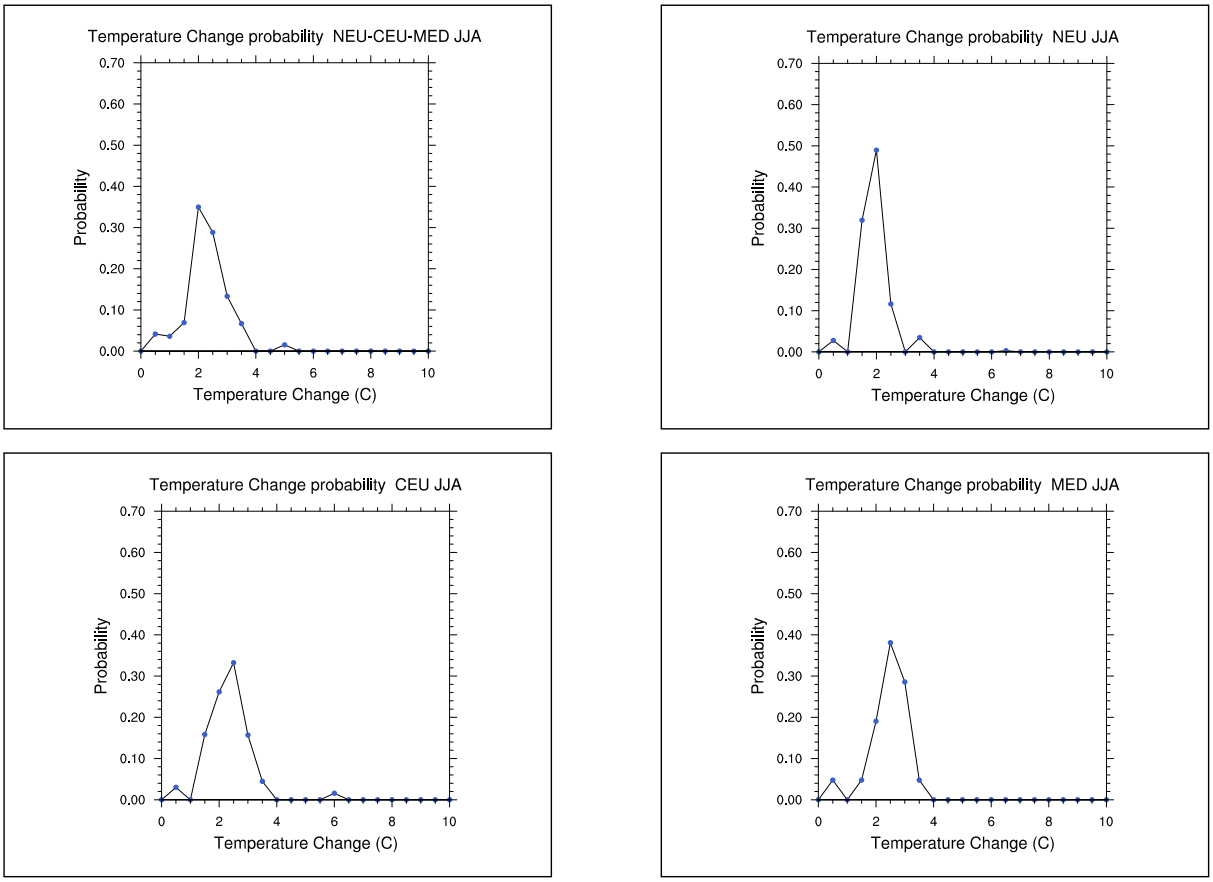


Figure 3: Temperature Change(°C) probability density functions for the whole European area (NEU-CEU-MED), and for the NEU, MED, CEU regions in the JJA season.

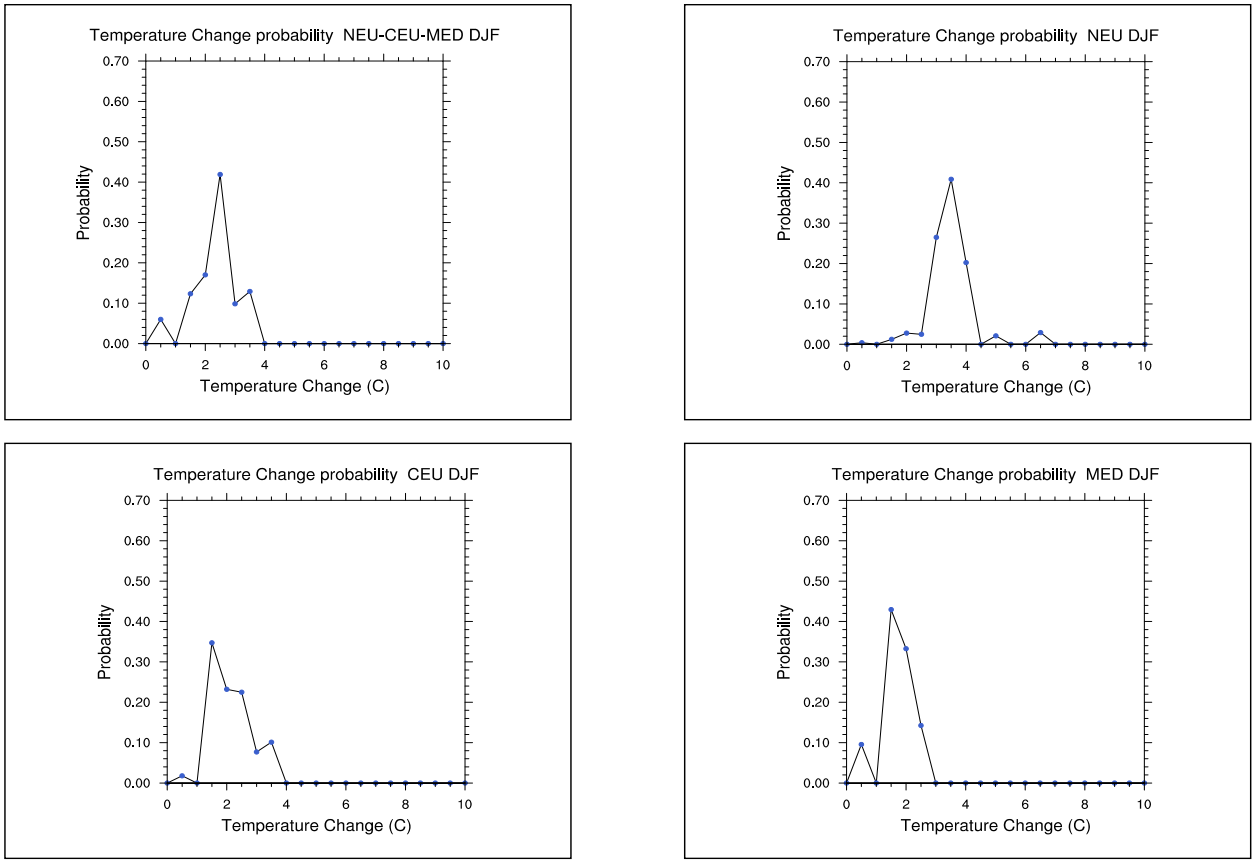


Figure 4: Temperature Change (°C) probability density functions for the the whole European area (NEU-CEU-MED), and for the NEU, MED, CEU regions in the DJF season.

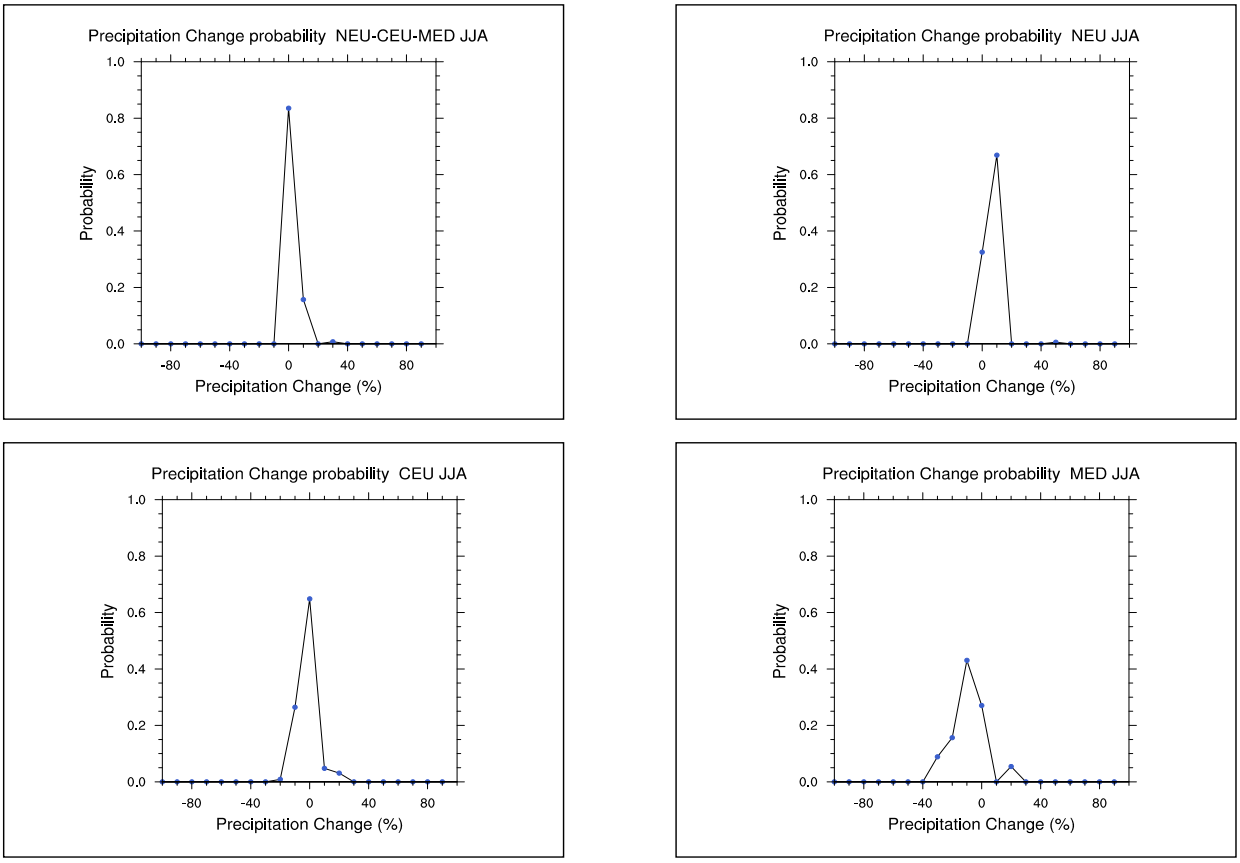


Figure 5: Precipitation Change (%) probability density functions for the whole European region (NEU-CEU-MED), and for the NEU, MED, CEU areas in the JJA season.

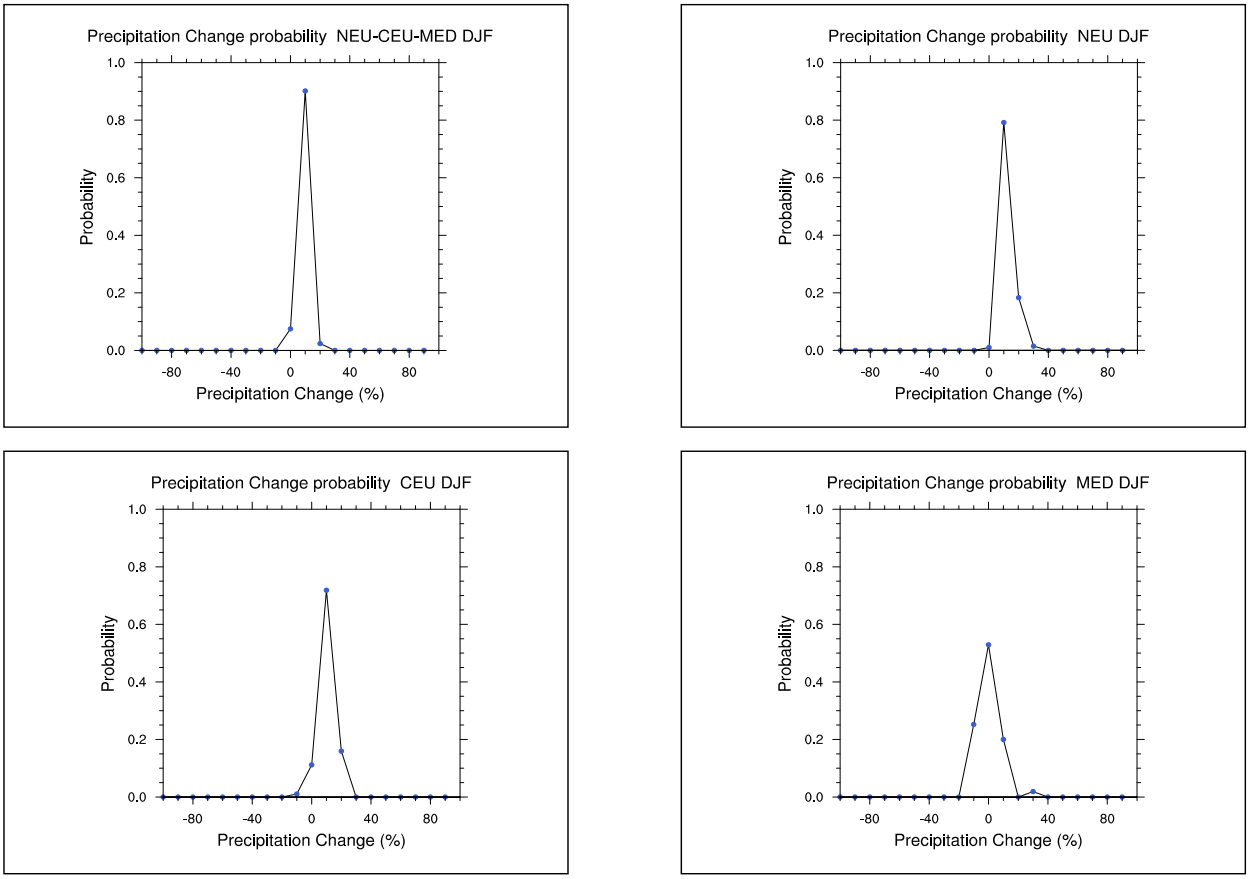


Figure 6: Precipitation Change (%) probability density functions for the whole european area (NEU-CEU-MED), and for the NEU, MED, CEU regions in the DJF season.

B.3 Allen-Stott-Kettleborough approach (ASK) - UEDIN

Author: Sabine Undorf

Short summary of the method

The ASK method derives constraints on the best estimate and the uncertainty range of future change using regression-based optimal detection and attribution methods to observed climate change. The range of factors by which the model simulated response to external forcing can be scaled and still be consistent with observations given internal variability is calculated and applied to future simulations. The approach relies on a forced signal being detectable in observations, and is therefore applicable where the forced signal has emerged, and practically limited by model ensemble size, the availability of observations, and a large enough target region given internal variability.

Key references

Allen et al., 2000; Stott and Kettleborough, 2002; also Kettleborough et al., 2007

Data

We apply the ASK (Allen et al., 2000; Stott and Kettleborough, 2002; Kettleborough et al., 2007; Stott and Forest, 2007) method to constrain the CMIP5 (Taylor et al, 2012) model range of European temperature change using the E-OBS v19.0e (daily, 0.25° regular grid) dataset (Haylock et al., 2008). All models/ensembles are used that provide future RCP8.5 simulations and historical all-forcing and either natural-only or GHG-only runs covering 1950-2012; the analysis is repeated as a sensitivity study for the larger ensemble covering 1950-2005. For E-OBS, monthly values are calculated from the daily data. For the model data, the model's native land/sea mask is applied to retain only grid boxes with a land fraction of ≥50%. To determine the scaling factors, all data is then regridded with an area-weighted scheme to a regular 2.5°×2.5° grid; monthly anomalies are calculated with respect to 12/1950-11/2005; summer (JJA)-mean values taken; and area-weighted area means taken for the NEU, CEU, MED, and total EU region.

Details of method

The ASK method (for Allen et al., 2000; Stott and Kettleborough, 2002) derives constraints on the best estimate and the uncertainty range of future change using regression-based detection and attribution methods. Past observed climate variations are a combination of internal variability and the response to natural (volcanic, solar) and anthropogenic forcings; the range of future model projections will include internal variability and the response to anthropogenic forcings. The forced component of these future projections -estimated as the (typically multi-model) ensemble mean- is scaled by the factor by which models under/overestimate the response to the forcing in the past. Uncertainty in the projection arises from the uncertainty of the estimate of the forced signal. Uncertainty in future natural forcings (Stott and Kettleborough, 2002; Kettleborough et al., 2007) as well as internal variability can further be added to give the full range in which future observations may be expected to lie. The fundamental assumption is that as the *magnitude* of the response is uncertain due to uncertain feedbacks, estimating it from observations of change is important, while the *pattern* of response, in contrast, is governed by climate physics. The presence of aerosol forcing in addition to greenhouse gases complicates this assumption.

The method can be applied in various ways to account for the different sources of external forcing. The simplest option is to scale future projections using the scaling factor derived by regressing "historical" model simulations which contain all forcings onto the observations to scale future projections (as in, e.g., Shiogama et al., 2016). This will be problematic, however, if the models under/overestimate the response to different forcings to different degrees while the ratio between these forcings changes over time. Ideally, future projections would therefore be available for different forcings individually, so that the best estimate of future change could be composed of the scaled response to the individual forcings, with scaling factors derived from historical single-forcing runs. Such future single-forcing runs are planned for a range of CMIP6 models as part of the Detection and Attribution Model-Intercomparison project (DAMIP; Gillett et al, 2016), but for most models not available as part of CMIP5. We can instead scale the future with the scaling factor for the response to anthropogenic forcing only (as in, e.g., Li. et al, 2017), with the caveat of time-varying aerosol-GHG ratios in past and future (Vuuren et al., 2011; Gidden et al., 2019), which limits the method's constraining power (Shiogama et al., 2016). Alternatively, we can derive the scaling factor for the response to GHG forcing only, with the caveat that the future projections also include aerosols and other anthropogenic forcings. Both options are better than using the all-forcing scaling factor since volcanic forcing is not explicitly included in future projections. We expect β_{ANT} to be generally better constrained than β_{GHG} due to the higher uncertainties associated with the relative magnitudes of GHG warming and aerosol cooling in the past.

Here, we derive the scaling factors as follows. For each region, the CMIP5 multi-model mean (MMM) area-mean time series from the historical all-forcing and single-forcing simulations are smoothed with 5-year running means (see Fig 3 for a sensitivity test) and anomalies taken to give the ALL, NAT, and GHG fingerprints. They are optimised with the covariance matrix of internal variability estimates from pre-industrial control simulations by PCA-whitening with truncation (Allen and Tett, 1999; Fig 3). Best-guess scaling factors are then derived by regressing equally smoothed anomalies from E-OBS data simultaneously onto the (a) ALL and NAT fingerprints or (b) ALL and GHG fingerprints using total least-squares regression (Allen and Stott, 2003; implemented as in Polson et al. (2013)). Uncertainty in the scaling factors due to internal variability both in the observations and the forced fingerprints (due to finite model ensemble size) is accounted for by repeating this 10000 times with added climate noise -equally derived samples from pre-industrial control runs from the same models- to both fingerprints and observations. This gives distributions of the scaling factor (β) for each fingerprint from which that for (a) ANT (all but natural) and (b) GHG is derived. For more information on optimal fingerprinting see e.g., Bindoff et al., 2013, and note that other variants, e.g., using a regularisation approach for the optimisation (Ribes and Terray, 2013), could equally have been used to derive the scaling factors.

The best-estimate projection as well as the constrained range is obtained by scaling the MMM projection with the best estimate and percentiles of the distribution of scaling factors, respectively. Here, we scale the projected changes over all regions and grid boxes with the scaling factor range for temperature and precipitation changes, respectively, that are derived from analyis of the combined CEU-MED-NEU region. Note that the constrained projections describe the estimated spread in the forced response only. The additional uncertainty from future internal variability on projected time series may additionally be included e.g. by adding the sigma range corresponding to a chosen confidence level from the standard deviation of area pre-industrial control simulations processed in the same way as the historical simulations, as done in Fig. 2. The additional uncertainty

from future natural forcing (volcanic and solar) may further be added as estimated from the historical natural-only simulations.

The constrained change in temperature and precipitation between baseline and future period is derived from the scaled MMM projection, and compared to the spread in the raw multi-model ensemble. Scaled projections that are wider than the raw MM range indicate that the latter is overconfident in constraining the likely range of future observations.

Observational constraints applied

Each target variable is scaled by the factors derived from the comparison between the change in observations and in models of that same variable. This is done either for each region separately, or for the three subregions combined so that spatial information of the response pattern is included.

Key assumptions

- A model's over/underestimation of a climate variable's response to a specific forcing (combination) is the same in the past and in the future.
- The responses to the single forcings used are linearly additive.
- The magnitude of the climate system's true internal variability is captured reasonably well by the models; if under/overestimated, the constrained projection will be over/underconfident.
- The models' spatio-temporal pattern of response is correct and the observations free of uncertainty; these assumptions may be relaxed by accounting for model error and using observational ensembles, respectively (Hannart et al., 2014; Jones and Kennedy, 2017; Schurer et al., 2018).

Limitations

- The forced signal in the target variable must be detectable over internal variability in the observations. This depends on the magnitude and distinctiveness (e.g., of anthropogenic vs. natural) of the forced change given the length of the observational record and the magnitude of internal variability as well as a good enough estimate of the forced response i.e. a large enough model ensemble size.
- Availability of historical single-forcing simulations with the same models as those used for future projections.
- Availability of (e.g., CMIP5) model runs only until 2005 or 2012 loses the information contained in the observed change over the last 14 or 7 years, respectively.
- Availability of future single-forcing simulations i.e. single-forcing projections for improved performance.

Results

Both the combined anthropogenic and the GHG-only signal separately are detectable in the observed temperature changes in all cases tested for the European (EU) region and the combined European region, and in all but one cases for the sub-regions, too (Fig. 1). The respective scaling factor ranges are all consistent with 1, meaning that the unscaled MMM may also explain the observations. The response to the combined anthropogenic (OTH) forcing is not detectable in the observed temperature change until 2005 over Northern Europe (NEU), where internal variability is higher compared to CEU and MED, while the response to GHG-forcing alone is just detected.

Repeating the analysis for the shorter period of 1950-2005 (not shown) using the larger ensemble of models covering that period tends to give slightly larger scaling factors , since it excludes the most recent, since stopped, period of stalled warming (e.g., Fyfe et al., 2013; Flato et al., 2013; Hu and Fedorov, 2017; see also Stott and Jones, 2012), and gives a slight underestimation of the response over the Mediterranean (MED) region.

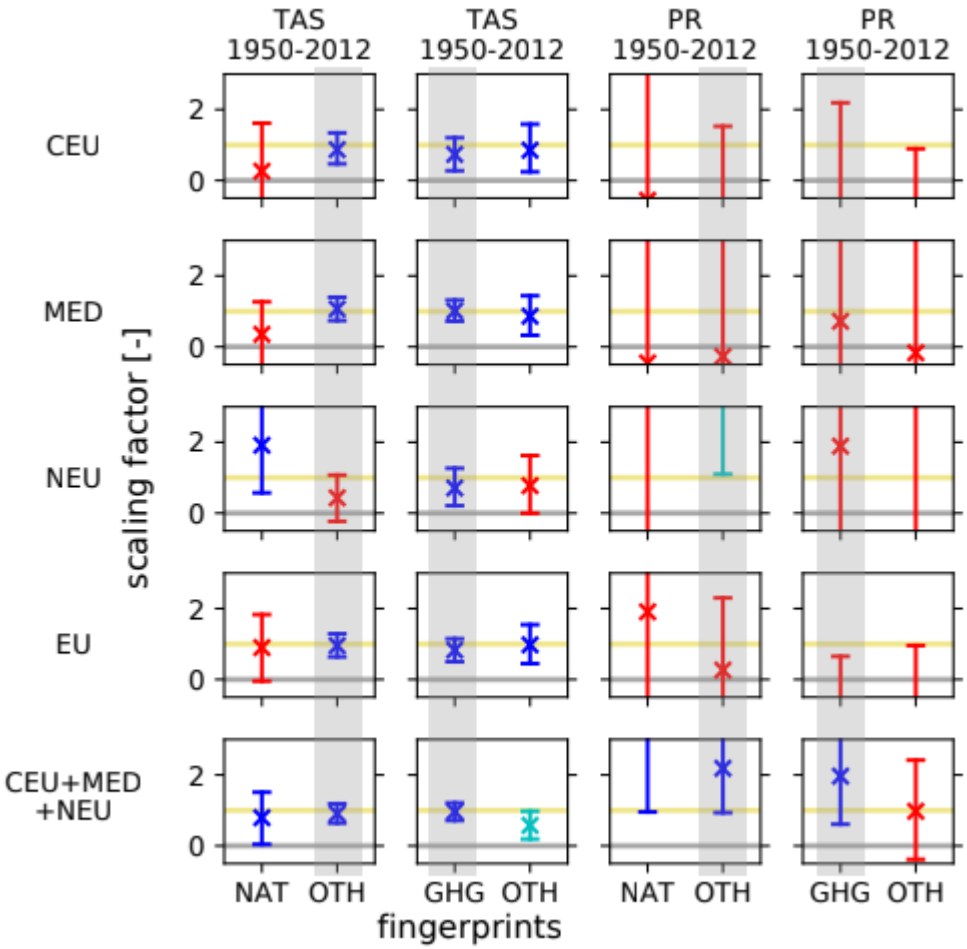


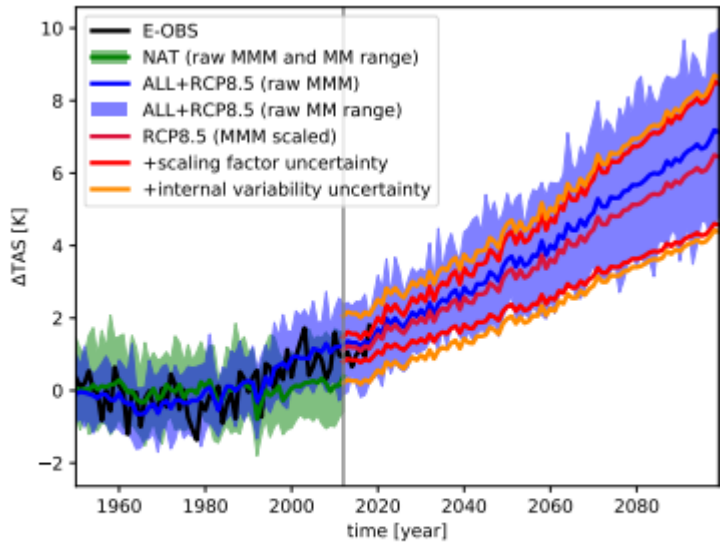
Figure 1: Scaling factors for simulated summer (JJA) (left) near-surface temperature (TAS) and (right) precipitation (PR) over the regions (top to bottom) CEU, MED, NEU, EU, and CEU+MED+NEU resulting from a two-signal detection & attribution analysis with respect to E-OBS observations. Shown are the scaling factors for the CMIP5 multi-model mean (MMM) response to natural (NAT) and anthropogenic (OTH) forcing from simultaneously regressing NAT and ALL on the observations, and those for greenhouse-gas (GHG) and non-GHG (OTH) forcing from simultaneously regressing GHG and ALL, for the time period 1950-2012. Crosses show the best-guess scaling factor, thick lines are the 5-95% range. Lines are blue when detected ($\beta > 0$) at the 95% confidence interval, red when not, and turquois when detected observed changes are inconsistent in magnitude with the MMM (i.e. significantly smaller when below 1, larger when above). The scaling factors for the anthropogenic/GHG forcing are highlighted in grey. The multi-model ensembles are of size 29 (NAT/OTH) and 28 (GHG/OTH) owing to the availability of the respective model runs.

Estimating scaling factors for each region individually assumes independence between these regions, and differing factors would be interpreted to mean model error in the simulated pattern of temperature change; we are cautious to draw this conclusion, however, without further analysis on robustness and an understanding as to why this might physically be the case. Deriving scaling factors for the combined region (CEU+MED+NEU) instead has a better signal to noise ratio while still

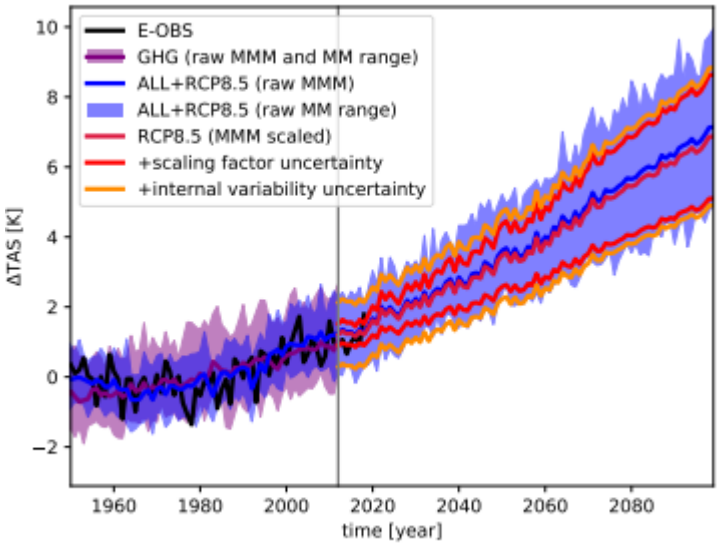
accounting for regional differences in the magnitude of change (as opposed to using area-means over the total region, EU). The scaling factors for GHG-only forcing and for the combined anthropogenic forcing are similar (Fig. 1), reflecting the dominance of past GHG forcing and/or suggesting similar scaling factors for GHG and other anthropogenic forcings in the past.

Applying the scaling factors, either for the response to anthropogenic forcing (β_{ANT}) or that to GHG forcing only (β_{GHG}) and both derived from the combined region, to temperature predictions over Europe (EU) gives the time series shown in Fig. 2. Scaling factor uncertainty leads to uncertainties in the future forced response about as wide as the raw multi-model (MM) range throughout the 21st century. Additional uncertainty related to internal variability is low compared to scaling factor uncertainty (Fig. 2), and that from future natural forcing even lower (not shown); note however, that for lower confidence ranges, this is different given the temperature response to very large volcanic eruptions.

An anthropogenically forced change in precipitation is detected when regressing on observations of the regions combined (Fig. 1), with a larger range of scaling factors compatible with the observed changes than for temperature. Precipitation is subject to high internal variability as well as model uncertainty, and even detecting a forced response in this case may be considered a success, given the difficulty to detect it even in monsoon regions (Undorf et al., 2018) that are thought most susceptible to anthropogenic forcing. This result adds a note of caution to apparent constraints on European precipitation projections based on any other method that constrains to observed changes using the same data over individual regions, and suggests that larger changes in rainfall may occur compared to those predicted in models.



(a) scaled from NAT, ALL



(b) scaled from GHG, ALL

Figure 2: European-mean JJA-mean temperature change relative to 1950-2005. Shown are the time series for (black) observations and as simulated with CMIP5 (blue line and shading) all-forcing (ALL) simulations and (a) (green line and shading) natural-only (NAT) or (c) (purple line and shading) GHG-only simulations for the historical period as well as, for the future, (blue line with dark blue shading) raw multi-model mean (MMM) and 5-95% range. The constrained future projections are shown as (crimson) MMM scaled with the best-guess scaling factor and its (red lines) uncertainty range, along with the additional uncertainty due to unknown future (orange lines) internal variability calculated as 1.66σ in pre-industrial control simulations added in quadrature. The future projections are scaled with the scaling factor for the response to (a) anthropogenic forcings and (c) GHG forcing, both constrained from 5-year running mean observations during 1950-2012 and keeping spatial information in the regression by using the three subregions CEU, MED, and NEU combined. The vertical grey line indicates where past -from which scaling factors are derived- ends and the future -to which the scaling factors are applied- begins. Note that the uncertainty cone for the scaled projection would be zero in 2012, and then the relative contribution from internal variability be larger, if the change with respect to that year was shown. Shown are annual time series.

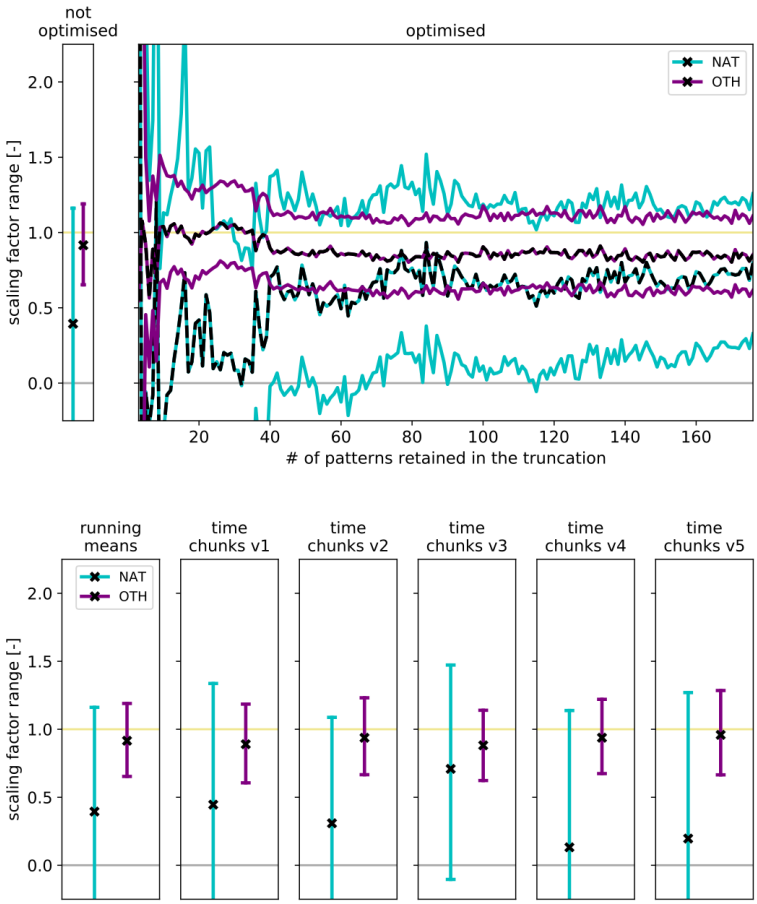


Figure 3: Example of the robustness of the scaling factor range with regard to details of the (top) optimisation in the TLS and (bottom) dimension reduction of the fingerprints. Shown are the results for a 2-signal analysis using all-forcing and natural-only simulations for near-surface temperature. (Top) The (left) non-optimised results are shown as well as those when (right) optimised using truncated PCA-whitening, shown as a function of the number of patterns retained. (Bottom) Results from a non-optimised analysis when the fingerprints are either (left) smoothed by 5-year running means prior to the TLS, or (right) split into time chunks of 5-year means, starting with the (left to right) first, second, …, fifth data point (year).

References

Allen, M. R., and P. A. Stott (2003), Estimating signal amplitudes in optimal fingerprinting, part I: theory, Clim. Dynam., 21, 477–491, doi:10.1007/s00382-003-0313-9.

Allen, M. R., P. A. Stott, J. F. Mitchell, R. Schnur, and T. L. Delworth (2000),
Quantifying the uncertainty in forecasts of anthropogenic climate change, Nature, 407 (6804), 617–620, doi:10.1038/35036559.

R. Allen, M., and Tett, S. (1999). Allen, M. R. & Tett, S. F. B. Checking for model consistency in optimal fingerprinting. Clim. Dyn. 15, 419-434. Climate Dynamics. 15. 419-434. 10.1007/s003820050291.

Bindoff, N.L., P.A. Stott, K.M. AchutaRao, M.R. Allen, N. Gillett, D. Gutzler, K. Hansingo, G. Hegerl, Y. Hu, S. Jain, I.I. Mokhov, J. Overland, J. Perlwitz, R. Sebbari and X. Zhang, 2013: Detection and Attribution of Climate Change:

from Global to Regional. 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 [Stocker, T.F., D. Qin, G.-K.
Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Flato, G., et al. (2013), Evaluation of climate models, 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, pp. 741-882, Cambridge Univ. Press, Cambridge, U. K., doi:10.1017/CBO9781107415324.020.

Fyfe, J. C., N. P. Gillett, and F. W. Zwiers (2013), Overestimated global warming over the past 20 years, Nat. Clim. Change, 3 (9), 767-769, doi:10.1038/nclimate1972.

Gidden, M. J., K. Riahi, S. J. Smith, S. Fujimori, G. Luderer, E. Kriegler, D. P.Van Vuuren, M. Van Den Berg, L. Feng, D. Klein, K. Calvin, J. C. Doelman,S. Frank, O. Fricko, M. Harmsen, T. Hasegawa, P. Havlik, J. Hilaire, R. Hoesly,J. Horing, A. Popp, E. Stehfest, and K. Takahashi (2019), Global emissionspathways under different socioeconomic scenarios for use in CMIP6: A datasetof harmonized emissions trajectories through the end of the century, Geoscien-tific Model Development, 12 (4), 1443–1475, doi:10.5194/gmd-12-1443-2019.

Gillett, N. P., H. Shiogama, B. Funke, G. Hegerl, R. Knutti, K. Matthes, B. D.Santer, D. Stone, and C. Tebaldi (2016), The Detection and Attribution ModelIntercomparison Project (DAMIP v1.0) contribution to CMIP6, GeoscientificModel Development, 9 (10), 3685–3697, doi:10.5194/gmd-9-3685-2016.

Hannart, A., A. Ribes, and P. Naveau (2014), Optimal fingerprinting under multiple sources of uncertainty, Geophys. Res. Lett., 41, 1261–1268, doi:10.1002/2013GL058653.

Haylock, M. R., N. Hofstra, A.M.G. Klein Tank, E.J. Klok, P.D. Jones, and M. New. (2008), A European daily high-resolution gridded dataset of surfacetemperature and precipitation for 1950-2006, Journal of Geophysical Research-Atmospheres, 113, D20,119, doi:10.1029/2008JD010201.

Hu, S., & Fedorov, A. V. (2017). The extreme El Niño of 2015–2016 and the end of global warming hiatus. *Geophysical Research Letters*, *44*(8), 3816–3824. https://doi.org/10.1002/2017GL072908

Jones, G.S. and J.J. Kennedy, 2017: Sensitivity of Attribution of Anthropogenic Near-Surface Warming to Observational Uncertainty. *J. Climate,* **30**, 4677–4691, https://doi.org/10.1175/JCLI-D-16-0628.1

Kettleborough, J. A., B. B. Booth, P. A. Stott, and M. R. Allen (2007), Estimates of uncertainty in predictions of global mean surface temperature, Journal ofClimate, 20 (5), 843–855, doi:10.1175/JCLI4012.1.

Li., C., Zhang, X., Zwiers, F., Fang, Y., & Michalak, A. (2017). Recent Very Hot Summers in Northern Hemispheric Land Areas Measured by Wet Bulb Globe Temperature Will Be the Norm Within 20 Years. Earth's Future, 5, 1203–1216,http://doi:10.1002/2017EF000639.

Polson, D., G. C. Hegerl, X. Zhang, and T. J. Osborn (2013), Causes of robustseasonal land precipitation changes, Journal of Climate, 26 (17), 6679–6697,doi:10.1175/JCLI-D-12-00474.1.

Ribes, A., and L. Terray, 2013: Application of regularised optimal fingerprint analysis for attribution. Part II: Application to global near-surface temperature, Climate Dynamics, 41, 2837–2853, doi:10.1007/s00382-013-1736-6.

Schurer, A., G. Hegerl, A. Ribes, D. Polson, C. Morice, and S. Tett, 2018: Estimating the Transient Climate Response from Observed Warming. *J. Climate,* **31**, 8645–8663, https://doi.org/10.1175/JCLI-D-17-0717.1

Shiogama, H., D. Stone, S. Emori, K. Takahashi, S. Mori, A. Maeda, Y. Ishizaki,and M. R. Allen (2016), Predicting future uncertainty constraints on globalwarming projections, Scientific Reports, 6 (January 2016), 1–7, doi:10.1038/srep18903.

Stott, P. A., and C. E. Forest (2007), Ensemble climate predictions using climate models and observational constraints, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 365 (1857), 2029–2052, doi:10.1098/rsta.2007.2075.

Stott, P. A., and G. S. Jones (2012), Observed 21st century temperatures further constrain likely rates of future warming, Atmos. Sci. Let., 13, 151–156, doi:10.1002/asl.383.

Stott, P. A., and J. A. Kettleborough (2002), Origins and estimates of uncertainty in predictions of twenty-first century temperature rise, Nature, 416 (2002), 723–727.

Taylor, K. E., Stouffer, R. J. and Meehl, G. A. (2012) 'An overview of CMIP5 and the experiment design', *Bulletin of the American Meteorological Society*, 93(4), pp. 485–498. doi: 10.1175/BAMS-D-11-00094.1.

Undorf, S., Polson, D., Bollasina, M. A., Ming, Y., Schurer, A., & Hegerl, G. C. (2018). Detectable impact of local and remote anthropogenic aerosols on the 20th century changes of West African and South Asian monsoon precipitation. Journal of Geophysical Research: Atmospheres, 123, 4871–4889. https://doi.org/10.1029/2017JD027711

Vuuren, D. P. V., J. Edmonds, M. Kainuma, K. Riahi, N. Nakicenovic, S. J.Smith, and S. K. Rose (2011), The representative concentration pathways: anoverview, Clim. Change, 109, 5–31, doi:10.1007/s10584-011-0148-z.

B.4 Historically constrained probabilistic projections (HistC) - CNRM

Authors: Saïd Qasmi, Aurélien Ribes, Hervé Douville

Method

The following method aims at providing a probabilistic projection of future changes of the European climate by using the information from the observed warming during the historical period within CMIP5 models.

Recipe

We consider a state vector X describing changes in temperature (e.g. temperature at different time steps), and other useful variables. Details on how X is defined are given below.

As a first step, we estimate X in each climate model considered. The forced response is estimated through using a Generalized Additive Model, where the response to natural forcings is derived from an Energy Balance Model and the anthropogenic influence is assumed to be smooth in time. Then, we derive a multi-model distribution $\prod(X)$ for X. $\prod(X)$ is calculated considering the "models are statistically indistinguishable from the truth" paradigm. In a Bayesian perspective, where the truth is treated as non-deterministic, this paradigm simply assumes that the models and the truth are taken from the same distribution.

As a second step, we want to constrain the model simulated changes by observations Y. For this purpose, the multi-model distribution $\prod(X)$ is considered as a prior distribution, and we want to derive the posterior distribution $p(X|Y)$. As a main statistical model, we assume:

$Y = HX + \varepsilon_Y$, where $\prod(X) \sim N(\mu, \sum_X)$ and $\varepsilon_Y \sim N(0, \sum_Y)$

where H is an observation operator, $\sum_Y$ describes the noise in observations (i.e. the deviation between the forced response and the observed climate). Two main factors contribute to $\sum_Y$: measurement error and internal variability. $\prod(X)$ and ε_Y are assumed to follow normal law *N*, with the multimodel mean μ and zero as respective means and known variance-covariance matrices $\sum_X$ and $\sum_Y$ (see Ribes et al. (2017) for the method of estimation of these matrices).

In this model, the posterior distribution $p(X|Y)$ provides confidence regions for X given observations Y and can be derived in closed form using a conventional Bayesian technique:

$X|Y \sim N(\mu + \sum_X H'(H\sum_X H'+\sum_Y)^{-1}(Y-H\mu), \sum_X - \sum_X H'(H\sum_X H'+\sum_Y)^{-1}H\sum_X)$

This is a new method and there is currently no published paper providing a detailed description of the method.

Practical implementation

As a state vector X, we use:
$X = (T^{reg}_{1870}, T^{reg}_{1871}, \dots, T^{reg}_{2100}, T^{glo}_{1870}, T^{glo}_{1871}, \dots, T^{glo}_{2100})$,

where T^{glo}_{1870} is the global mean temperature in year 1870, T^{reg}_{1870} is the summer mean temperature (June-July-August, JJA, as the focus is on summer only) over the considered region (which can be EU, NEU, CEU, MED). X is then a vector of dimension 2*n*, in which the first *n* elements are the temperature time series calculated from 1870 to 2100 for the given region.

The vector of observations Y is written as follows:
$Y = (T^{reg}_{obs,1870}, T^{reg}_{obs,1871}, \dots, T^{reg}_{obs,2100}, T^{glo}_{obs,1870}, T^{glo}_{obs,1871}, \dots, T^{glo}_{obs,2100})$,

where $T^{reg}_{obs,i}$ ($T^{reg}_{obs,i}$) is the observed global (regional) temperature for each observed year *i*.

The observation operator H is then a sparse matrix of dimension 2*n* x 2*n*, in which $H_{i,j} \in \{0,1\}$.

Data

Simulated surface air temperature are taken from historical runs and RCP8.5 scenarios in 34 CMIP5 models over the 1870-2100 period. The observed air surface temperature (precipitations) comes from the HadCRUT4 (GPCC) dataset over the 1870-2018 (1901-2013) period.

Key assumptions

- The response to anthropogenic forcings is smooth over time,
- Each climate model is independent from the others,
- All variables follow Gaussian distributions,
- The real-world response to external forcings lies within the range of (or population of) model responses, consistent with the "models are statistically indistinguishable from the truth" paradigm.

Limitations

- The estimation of $\sum_Y$ is based on a pre-industrial control simulation, in which the interannual and decadal internal variabilities are not necessarily identical to the observed internal variability,
- The real-world response to external forcings (past and future) lies within the range of model responses,
- $\sum_X$ and $\sum_Y$ are difficult to estimate, and are treated as known quantities in the statistical procedure.

B.5 A Bayesian Method for Probabilistic Climate Projections for Europe (UKCP) - Met Office

Authors: G.R. Harris, J.M. Murphy, D.M.H. Sexton, B.B.B. Booth

Short summary of method

A Bayesian approach has been used to produce probabilistic projections of the European climate response for the RCP85 emission-driven scenario. To achieve this, a large Monte-Carlo integration over defined ranges of uncertain model parameters is performed, weighting the sampled outcomes in the integration by relative likelihood (estimated using a set of specified observables) to produce observationally constrained pdfs. In order to make the estimation computationally feasible, predicted outcomes are sampled from a fast statistical emulation of the equilibrium response to doubled CO_2, calibrated to the response of a large perturbed parameter ensemble (PPE) of simulations based on a single model. A scaling approach calibrated to the transient response of a second PPE of earth system model variants allows the transient response to be inferred, while also enabling sampling of other uncertainties associated with the carbon cycle, ocean and aerosol components of the climate system. A third ensemble of CMIP5 simulations is also used to sample the structural error component associated with the PPE base model.

Key references

Murphy *et al.* (2018), Harris *et al.* (2013), Sexton *et al*. (2012), Sexton and Harris (2015).

Details of Method

The methodology used here is based on the Bayesian statistical framework that underpinned a previous set of UK Climate Projections (UKCP09, Murphy *et al*. 2009). Details of the approach are given in Sexton *et al.* (2012) and Harris *et al*. (2013). This methodology has been subsequently updated and used to produce a new set of probabilistic climate projections for the United Kingdom (Murphy *et al*. 2018). Key elements of the method, including recent updates are summarized below (see preceding references for additional detail).

Modeling uncertainty is explored using the Perturbed Parameter Ensemble (PPE) approach (Collins *et al*. 2006, Murphy *et al*. 2007), where key parameters controlling processes in the atmosphere, surface, ocean, aerosol and land carbon cycle components of the climate system are varied within expert-defined "prior" parameter spaces for a single climate model. Here, variation in historical climate and the equilibrium response to a doubling of CO_2 concentrations is explored using a relatively large 280 member PPE based on the atmosphere-mixed-layer (hereafter SLAB) configuration of the HadCM3 model (Pope *et al*. 2000, Johns *et al*. 2003). Multivariate regression relationships trained on the 280 PPE variants are then used to construct a statistical emulator for the historical and future equilibrium responses. In contrast to use of multi-model ensemble (MME) output, emulation based on PPE provides a systematic and comprehensive sampling of climate response for untried parameter combinations, and makes use of a Bayesian approach both possible and practicable.

Users are impacted by the transient climate response to scenarios of climate forcing rather than the equilibrium response, so a second emulation stage is implemented. Assuming the climate response is proportional to global mean surface temperature (GMST) change (Santer *et al*. 1990), scaling techniques are then used to translate the emulated equilibrium response into estimates of past and

future climate response, for any point in parameter space. Given emulated predictions for the climate feedbacks, a Simple Climate Model (SCM) (Harris *et al*. 2013, Suppl. Mat.) is used to predict the transient global temperature response and perform the scaling.

Uncertainty in modeling the rate of CO_2 uptake by the land and ocean biogeochemical systems contribute substantially to uncertainties in projections of GMST, which in turn influence regional response (Knutti *et al*. 2008, Booth *et al*. 2012). Here we therefore use an emission-driven approach (rather than concentration-driven) in order to take fuller account of known limitations in the current modeling of earth system processes. To this end, a 57-member PPE, based on variants of the HadCM3C Earth System Model (ESM) has been used to explore modeling uncertainty in response to RCP85 forcing, accounting for interactions between different earth system components (Lambert *et al*. 2013, Murphy *et al*. 2014). This ensemble (hereafter the ESPPE) is used to calibrate the SCM and provide prior distributions for key earth system processes, such as ocean heat uptake and climate-carbon feedbacks, as well as providing adjustment of potential differences between SLAB and transient regional response.

To produce probability distributions for climate response, Monte-Carlo integration is performed over the prior space of model and SCM parameters. Emulation is used to estimate response, allowing large sample sizes of order 10^6 to be produced, hence improving the coverage of parameter space. Some parameter choices perform better than others when comparing, for example, their predictions of historical climate. Within the Bayesian framework, we include a specified multivariate set of observables in the set of predicted variables, and estimate a "likelihood" for each parameter set (model variant), given the observed data. Each variant is weighted by relative likelihood in the Monte-Carlo integration, to produce updated posterior predicted probability distributions for climate response (Harris *et al*. 2013). The observational data used for likelihood estimation is specified below.

One important component of the method is the recognition that models are imperfect. Given that some parameter choices are better than others at reproducing observations, it is reasonable to assume that there exists a "best input" set of parameter choices that provides the best simulation of true climate. However, due to model imperfection, even the best-input models will possess an irreducible structural error component (termed "discrepancy" in Sexton *et al*. 2012) that cannot be eliminated. We estimate the effects of such structural errors by using output from other independent climate model simulations, searching the prior space for best-input parameter sets that best reproduce output for selected CMIP5 models (Taylor *et al*. 2012). Since we account for carbon-cycle modeling uncertainty, emission-driven CMIP5-ESMs are used for this purpose. We do not use observations to estimate structural error because (a) we don't have observations of the future, and (b) in order to avoid any issues associated with double-counting, since the observations are also used to constrain the projections.

We are obliged to develop the complex projection methodology described above since there is insufficient computational resource to produce sufficiently large, fully-coupled transient ensembles of scenarios projections to provide robust, comprehensive and yet plausible samples of projections for future climate change. We are careful therefore to validate the components of the method, and include the additional statistical uncertainties that arise from the different steps. These include: equilibrium response emulation error, error in converting from equilibrium to transient response, time-scaling error (including inherent model internal variability), and structural error estimates.

Developments going from ENSEMBLES to EUCP

The methodology described in Harris *et al*. (2013) was developed for the UKCP09 projections released in 2009. Subsequently, the same methodology was used to produce European projections (Harris *et al*. 2010) as part of the ENSEMBLES project (van der Linden and Mitchell 2009). More recently, the method has been developed to provide the UKCP18 projections (Murphy *et al*. 2018), and now projections for EUCP. It is useful to review and highlight the main differences in methodology and scope compared to these earlier projections.

- A new 57-member ESPPE ensemble of HadCM3C earth system models has been used to calibrate the projections. This simplifies the method by allowing the number of PPE inputs to be reduced from seven to three, and allows the effects of uncertainties in ocean and carbon cycle processes on spatial patterns of climate change to be considered, alongside influences of land surface and atmospheric processes.
- The ESPPE ensemble allows full incorporation of the land carbon cycle into the Bayesian methodology, with definition of a prior parameter space, and likelihood estimation that includes the carbon cycle response.
- Rather than using the older slab-ocean CMIP3 simulations (Meehl *et al*. 2007) to estimate the structural error component, we now use 12 more recent CMIP5-ESM simulations (Taylor *et al*. 2012).
- Structural errors in predicting non-linear aspects of the transient GMST response (such as temporal effects which cannot be explained by use of a fixed climate feedback parameter) have additionally been accounted for.
- Additional observational constraints are used to weight projections from different points in parameter space, by adding metrics of historical change in upper ocean heat content and CO_2 concentration (the latter to constrain carbon cycle feedbacks, following Booth *et al*. 2017). The use of historical surface temperature changes is also updated to consider changes up to 2017 rather than 2000, thus including the recent "warming hiatus" period (e.g. Trenberth, 2015).
- The representation of historical changes in external forcing has been improved, by using a probability distribution for anthropogenic aerosol forcing provided by AR5 (Myhre *et al*. 2013), and accounting for uncertainties in fossil fuel and land-use carbon emissions (Booth *et al*. 2017).
- The methodology has been extended to present the probabilistic projections for individual years, rather than the 20 or 30-year averages of ENSEMBLES and UKCP09. This is based on the method of Sexton and Harris (2015).
- For UKCP18, we scaled percentage change in precipitation to predict the transient response. For the southern regions of Europe that are predicted to experience stronger summer drying compared to the UK (especially under RCP85 forcing toward the end of the century), this approach in some instances leads to unphysical scaled projections of less than -100%. We have therefore employed a mixed approach here, using the log transform for realizations with strong drying, while for realizations with weak drying or increased precipitation, linear scaling of percentage change is used (Watterson 2008).
- Unlike the UKCP09 and UKCP18 projections, here we do not statistically downscale the projections to a higher resolution of 25km. In this respect, the EUCP projections are like the ENSEMBLES projections, with a finest grid-point resolution of 2.5°×3.75°.

Data

Three ensembles are used to calibrate the methodology:

- 280 1×CO_2 and 2×CO_2 PPE equilibrium simulations with the HadSM3 model (Pope *et al*. 2000, Sexton *et al*. 2012)
- 57 PPE simulations with the coupled atmosphere-ocean earth system model HadCM3C (Lambert *et al*. 2013, Murphy *et al*. 2014), simultaneously varying a total of 54 parameters in the atmosphere/land

surface, ocean, sulfur cycle and terrestrial carbon cycle components. Flux-adjustments are used (Collins *et al*. 2011), and emission-driven RCP85 scenarios from 1860 to 2100 have been produced.

- 12 CMIP5-ESM emission-driven RCP85 simulations. A total of 15 models in the CMIP5 database were initially considered, but three were omitted due to lack of data for some variables, or due to a lack of independence. The 12 models are: bcc-csm1-1, bcc-csm1-1-m, BNU-ESM, CanESM2, CESM1-BGC, GFDL-ESM2G, HadGEM2-ES, inmcm4, IPSL-CM5A-LR, MIROC-ESM, MPI-ESM-LR, and MRI-ESM1. All output is regridded to the 2.5°×3.75° HadCM3 grid, and this is the highest resolution for which we can produce probabilistic projections for EUCP.

Observational data from a variety of sources is employed (see next section). The SCM used to perform the scaling is described in the Supplementary Material for Harris *et al*. (2013).

Observational Constraints Applied

Observational constraints are applied by weighting sampled outcomes by likelihood weights calculated from the multivariate distance between emulated estimates of a set of historical variables and verifying observations. These observations include the same set of seasonal climatological spatial fields used for UKCP09 (Sexton *et al*. 2012) comprising of twelve climate variables: sea surface temperature, screen temperature, precipitation, TOA outgoing shortwave flux, TOA outgoing longwave flux, TOA shortwave cloud radiative effect, TOA longwave cloud radiative effect, sea-level pressure, relative humidity, total cloud, surface sensible heat flux, and surface latent heat flux. Data sources and references for each of these are listed in Table B.1 of Murphy *et al*. (2018). The data amounts to about 175,000 observables and it is necessary to reduce its dimensionality, in order to remove dependencies between variables, and to make the multivariate statistical calculations computationally feasible. This is done by identifying the six leading eigenvectors of these climatological variables in the SLAB ensemble, and emulating the amplitudes for these. They are then compared with amplitudes for these observables projected onto the same set of eigenvectors, and used to estimate associated likelihood weights.
In addition, historical trends for several climate indicators are also included in the set of observational constraints. These include the Braganza indices based on GMST (Braganza *et al*. 2003), heat content change in the top 700m of the oceans, and change in atmospheric CO_2 concentration over a recent 45 year period (Booth *et al*. 2017). Observational data sources and references are listed in Table B.2 of Murphy *et al*. (2018).

Key Assumptions

- The Bayesian statistical framework as presented in Rougier (2007) is assumed as the basis for our methodology.
- The true climate is assumed to lie within the spread of sampled prior outcomes, constructed from PPE and MME output.
- It is assumed that the structural error of the base model for the PPE can be estimated by taking MME simulations as proxies for the true climate and using our best input emulations for these models to specify structural error.
- There is a degree of subjectivity and judgment in specifying the observational data used to constrain the projections. Alternative variables, data sources and processing of the observational data could be used or implemented.

- It is assumed that models that are good at simulating historical climate and historical trends will be good at predicting future climate. Weighting by likelihood tends to support projections that match these observations (although imperfectly since constraints are multivariate).
- The patterns of equilibrium response are assumed to be representative of the fully-coupled response patterns. This assumption is validated in a subset of cases, and adjustment applied. Note that in the case of 10m UK wind-speed response, this assumption was found not to apply in a substantial number of cases, thus invalidating use of the method for this particular variable.
- Transient responses are assumed to scale linearly with global temperature response. This assumption is tested for the subset of ESPPE simulations, and any potential non-linearity in the residual error is resampled and included in the projections.
- For the purposes of time scaling, climate feedbacks are assumed to be constant and not evolve with climate state.
- The variability in the ESPPE, which is used to specify variability in the probabilistic projections, is assumed to be a good representation of variability in the true climate. Sensitivity to this assumption was tested in Sexton and Harris (2015) for the UK regions, by implementing a rescaling of variability to match observed variability. Conclusions were not altered substantially by this. This approach has not been implemented yet for the EUCP projections, but could potentially be applied later in the project.

Limitations

- The predicted pdfs can only cover the range spanned by the prior distribution, so if the real world response is outside this, it will not be captured in the pdfs (see next bullet).
- The inclusion of the structural error component can mitigate against overconfidence and lack of spread, but does not account for the unknown effects of errors common to all models (e.g. due to resolution, or missing processes). Note though that this issue of common model biases applies to all model-based projections, and not just our Bayesian methodology.
- If the overlap between the PPE and MME range of response is small, then our best-input emulations are likely to be poor, leading to large structural error. Structural error adjustments may then dominate the modeling uncertainty component, leading to sensitivity in the pdfs to the small sample of MME outcomes (i.e. a lack of robustness). Even if structural errors are relatively small, since the number of MME simulations is limited (12 in this study), the structural error estimation may be sensitive to choice of models.
- Large structural error adjustments along with accumulated statistical uncertainty may lead to broad pdfs and outcomes in the tails that are not supported by any simulations, reducing their credibility. For example, for UKCP18 a large majority of ESPPE variants predicted reductions in UK relative humidity, while most CMIP5-ESMs predicted the converse, an increase. The resulting large structural error adjustment led to relative humidity pdfs that lacked credibility, so these were not provided (Murphy *et al*. 2018). Similar screening checks were done for all UKCP18 variables, and most passed the tests, including all precipitation and temperature variables. Corresponding checks will be carried out subsequently for the EUCP precipitation and temperature projections provided here.
- Not all sources of uncertainty have been explored in our PPE, including modeling uncertainty in the ocean biogeochemistry component that influences ocean CO_2 uptake, uncertainty associated

with forcing from minor gases (methane, nitrous oxide, ozone, CFCs), and variations in solar and volcanic forcing.

- The pdfs are conditional, and depend on many things, including the models and understanding available at the time of their construction, prior assumptions regarding choice and range of uncertain model parameters, limited sample sizes for the PPE and MME ensembles, the choice of observational constraints, choices for input emissions and forcings, and methodological assumptions (e.g., emulation and scaling techniques, use of flux adjustments in the ESPPE, method for structural error estimation, etc). New models and observations, and updated methods will lead to evolution in the predicted pdfs.

Example Projections

The methodology has been used to produce all Tier I projections (large medium and small spatial aggregations for the JJA season). In addition, a set of Tier II projections has been made for the additional WP3 domains, for all four seasons, although no projections are available yet for tasmin and tasmax. Since projections are made on annual timescales, alternative baselines and future time-averaging periods are straightforward to produce. These projections are still provisional and require further calibration and validation, so although close to final, they are still subject to future potential revision. The final output of the methodology is a large sample (3000 here) of realizations constrained by observations that combine modeling uncertainties in the mean signal with internal variability from the input ESPPE simulations. The sample realizations can then be combined to produce probability distributions for projected climate response.

Figure 1 below shows an example of the projections: the summer surface air temperature response to the RCP85 emissions scenario for the three European SREX regions, plus the combined European region. The three coloured realizations are individual examples of the 3000 realizations in the full sample, which are represented in the Figure by selected quantiles of the distribution of response, evolving as grey plumes of uncertainty through the 21st century. There is nothing special about the three coloured realizations, except they have been selected to show a spread in response, picking EUR-SREX realizations close to the 10%, 50% and 90% quantiles at the end of the century. Summer warming is greater in southern regions of Europe compared to the north; for example, a median warming of 7.8°C is predicted for MED-SREX for RCP85, compared to 5.6°C for NEU-SREX.

Figure 2 presents a second example of the probabilistic projections, in this case pdfs for summer precipitation change for the four small WP2 regions for the 20 year mean period 2041-2060, relative to the 1995-2014 baseline, in response to RCP85 forcing. The region aggregates here actually represent 2.5°×3.75° grid-boxes (resolution of our input PPE) closest to the 2.5°×2.5° aggregate regions defined for WP2. For each variable the 3000 realizations are averaged over the selected 20-year future period, and a pdf is fitted to the resulting sample data using kernel density estimation (KDE) techniques to reduce the effects of sampling noise. The probability distributions are fitted here using the python scipy package 'gaussian_kde' (Jones *et al*. 2001), assuming a normal density function for the kernel and the bandwidth specified using Scott's Rule (Scott, 2015). The pdfs In Figure 2 are compared with climate model data from the SLAB, ESPPE and CMIP5-ESM ensembles that are input to the methodology. The SLAB equilibrium responses have been converted to equivalent transient response by scaling (randomly sampling carbon cycle feedbacks from the ESPPE prior parameter space).

We note that the ESPPE typically predicts a drier response (e.g. for Transylvania) compared to the CMIP5-ESM distribution of responses. The structural error implied by this adjusts the posterior pdfs

toward a somewhat less dry response than is obtained in the SLAB and ESPPE ensembles. The correspondence between the GCM data and the pdfs helps confirm that methodological assumptions have not lead to excessive statistical inflation (although some is unavoidable), providing confidence that the pdfs provide a reasonable expression of the uncertainties implied by the underlying set of model simulations. It can be noted too that the spread in response in the 12 member CMIP5-ESM ensemble is substantially smaller than that sampled in the two PPEs. There is a clear signal toward greater summer drying of climate in the southern regions of Europe compared to Northern regions; e.g., the median response for Madrid is -24%, compared a median increase of 5% for Svealand. Uncertainty is large however, and although a majority of realizations give reductions in rainfall, increases cannot be ruled out.

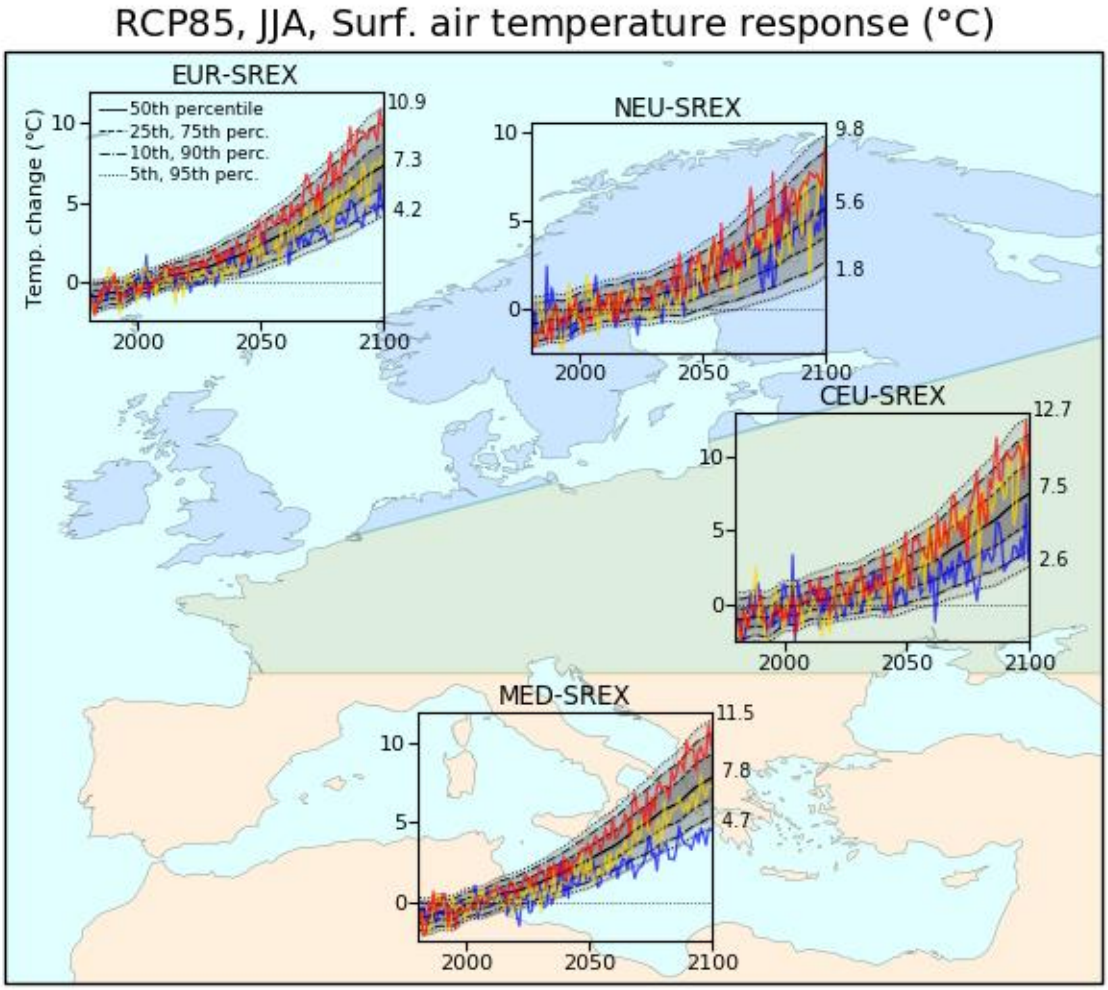


Figure 1: Projections for individual seasons in response to historical emissions followed by the RCP85 emission-driven scenario for summer temperature change for the SREX regions of Northern Europe (NEU-SREX, blue region), Central Europe (CEU-SREX, green region), the Mediterranean (MED-SREX, orange region) ,and the combined all-Europe region (EUR-SREX). Grey shading and lines show percentiles of anomalies in the variables relative to 1995–2014, calculated from 1-year mean PDFs constructed from 3000 sample realizations. Coloured lines show three individual realizations of year-to-year variation sampled from the 1-year pdfs so that simulated temporal correlations are captured. The three colours represent typical warm (red), medium (gold) and cool (blue) cases, and correspond to the same realization in each of the different plots.

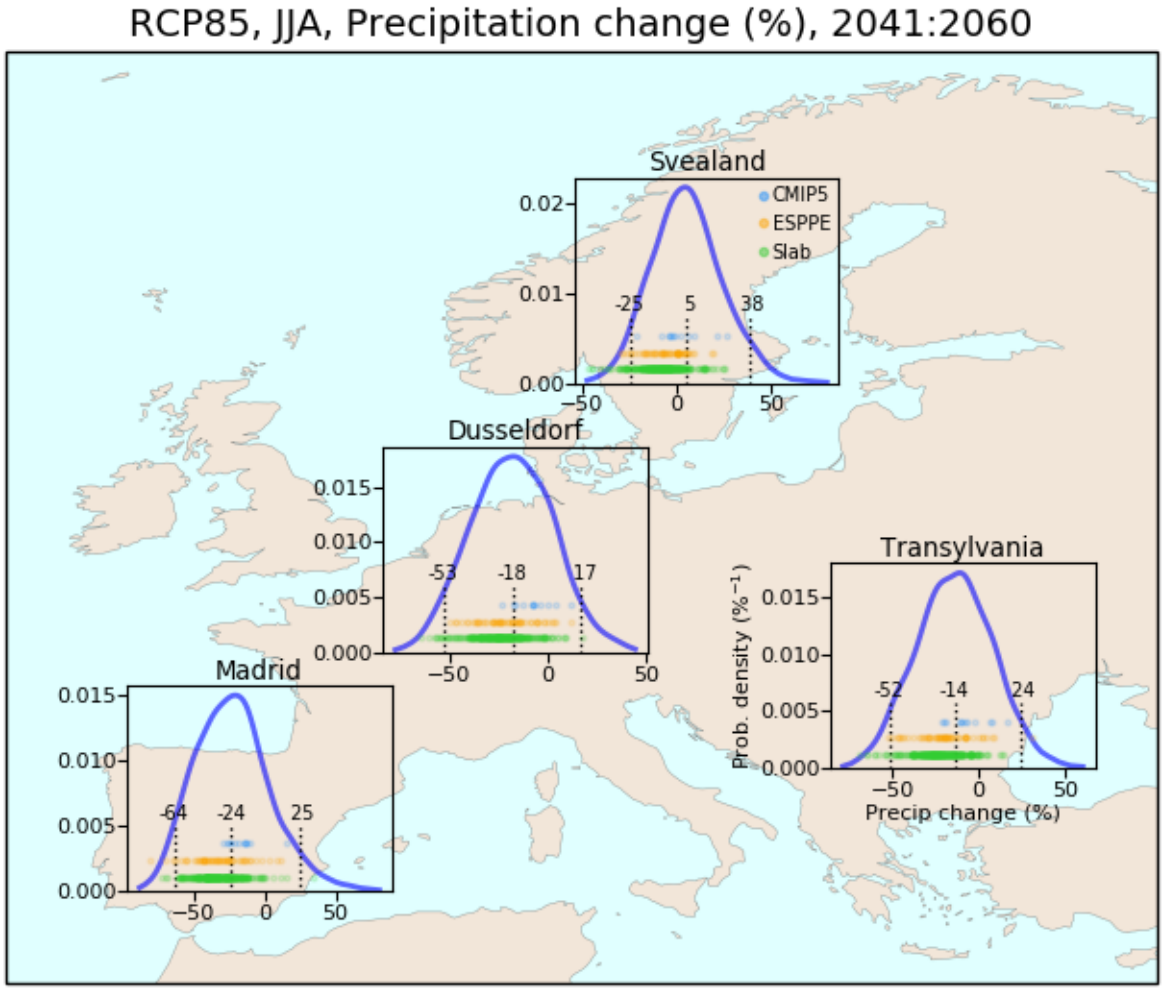


Figure 2: Probability density functions (blue curves) for the 20-year mean percentage change in precipitation relative to 1995-2014 for the RCP85 emission-driven scenario for the four small WP2 regions. Here we are limited to the by PPE resolution to the closest 2.5°×3.75° grid boxes, rather than the 2.5°×2.5° definitions. Positions and values of the 5th, 50th and 95th percentiles are indicated by the labeled dotted vertical lines. Coloured points correspond to data from the SLAB, ESPPE and CMIP5-ESM ensembles that are input to the methodology (see text).

References

Booth BBB, Jones CD, Collins M, Totterdell IJ, Cox PM, Sitch S, Huntingford C, Betts RA, Harris GR, Lloyd J (2012) High sensitivity of future global warming to land carbon cycle processes, *Environ. Res. Lett.*, 7, 24002, doi:10.1088/1748-9326/7/2/024002

Booth BBB, Harris GR, Murphy JM, House JI, Jones CD, Sexton DMH, Sitch S (2017), Narrowing the range of future climate projections using historical observations of atmospheric CO_2. *J. Clim.* 30, 3039-3053, https://doi.org/10.1175/jcli-d-16-0178.1

Braganza K, Karoly DJ, Hirst AC, Mann ME, Stott P, Stouffer RJ, Tett SFB (2003). Simple indices of global climate variability and change: Part I — variability and correlation structure. Clim. Dyn. 20:491–502.

Collins M, Booth BBB, Harris GR, Murphy JM, Sexton DMH, Webb MJ (2006), Towards Quantifying Uncertainty in Transient Climate Change *Clim. Dyn.*, 27, 127-147. doi:10.1007/s00382-006-0121-0.

Collins M, Booth BBB, Bhaskaran B, Harris GR, Murphy JM, Sexton DMH, Webb MJ (2011) Climate model errors, feedbacks and forcings: A comparison of perturbed physics and multi-model ensembles, *Clim Dyn*, 36, 1737-1766, doi 10.1007/s00382-010-0808-0.

Harris GR, Collins M, Sexton DMH, Murphy JM, Booth BBB (2010) Probabilistic Projections for 21st Century European Climate, *Nat. Hazards Earth Syst. Sci.*, 10, 2009-2020, https://doi.org/10.5194/nhess-10-2009-2010

Harris GR, Sexton DMH, Booth BBB, Collins M, Murphy JM (2013), Probabilistic Projections of Transient Climate Change, *Clim Dyn* 40: 2937. https://doi.org/10.1007/s00382-012-1647-y

Johns TC, Gregory JM, Ingram WJ, Johnson CE, Jones A, Lowe JA, Mitchell JFB, Roberts DL, Sexton DMH, Stevenson DS, Tett SFB, Woodage MJ (2003) Anthropogenic climate change for 1860 to 2100 simulated with the HadCM3 model under updated emissions scenarios. *Clim Dyn* 20:583–612

Jones E, Oliphant E, Peterson P, *et al*. SciPy: Open Source Scientific Tools for Python, 2001-, http://www.scipy.org/ [Online; accessed 2019-05-13]. For Gaussian kernel density estimation: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.gaussian_kde.html

Knutti R, Allen MR, Friedlingstein P, Gregory JM, Hegerl GC, Meehl GA, Meinshausen M, Murphy JM, Plattner G-K, Raper SCB, Stocker TF, Stott PA, Teng H, Wigley TML (2008). A review of uncertainties in global temperature projections over the twenty-first century. *J. Clim.* 21: 2651-2663.

Lambert FH, Harris GR, Collins M, Murphy JM, Sexton DMH, Booth BBB (2013) Interactions between perturbations to different Earth system components simulated by a fully-coupled climate model. *Clim Dyn* 41: 3055. https://doi.org/10.1007/s00382-012-1618-3

van der Linden P., and J.F.B. Mitchell (eds.) 2009: ENSEMBLES: Climate Change and its Impacts: Summary of researchand results from the ENSEMBLES project. Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK. 160pp. http://ensembles-eu.metoffice.com/docs/Ensembles_final_report_Nov09.pdf

Meehl GA, Covey C, Delworth T, Latif M, McAvaney B, Mitchell JFB, Stouffer RJ, Taylor KE (2007) The WCRP CMIP3 multi-model dataset: a new era in climate change research. *Bull Am Meteorol Soc* 88:1383–1394

Murphy JM, Booth BBB, Collins M, Harris GR, Sexton DMH, Webb MJ (2007) A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. *Philos Trans R Soc A* 365:1993–2028

Murphy, J.M., Sexton, D.M.H., Jenkins, G.J., Boorman, P.M., Booth, B.B.B., Brown, C.C., Clark, R.T., Collins, M., Harris, G.R., Kendon, E.J., Betts, R.A., Brown, S.J., Howard T.P., Humphrey, K.A., McCarthy, M.P., McDonald, R.E., Stephens, A., Wallace, C., Warren, R., Wilby, R., Wood, R.A. (2009). UK Climate Projections Science Report: Climate change projections. Met Office Hadley Centre, Exeter, U.K., https://webarchive.nationalarchives.gov.uk/20181204111026/http://ukclimateprojections-ukcp09.metoffice.gov.uk/22530#projections

Murphy JM, Booth BBB, Boulton CA, Clark RT, Harris GR, Lowe JA, Sexton DMH (2014), Transient climate changes in a perturbed parameter ensemble of emissions-driven earth system model simulations, *Clim Dyn* 43: 2855. https://doi.org/10.1007/s00382-014-2097-5

Murphy, J.M., Harris, G.R., Sexton, D.M.H., Kendon, E.J., Bett, P.E., Clark, R.T., Eagle, K.E., Fosser, G., Fung, F., Lowe, J.A., McDonald, R.E., McInnes, R.N., McSweeney, C.F., Mitchell, J.F.B., Rostron, J.W., Thornton, H.E., Tucker, S., Yamazaki, K. (2018) UKCP18 Land Projections: Science Report, Met Office Hadley Centre, Exeter, U.K., https://www.metoffice.gov.uk/pub/data/weather/uk/ukcp18/science-reports/UKCP18-Land-report.pdf

Myhre G, Shindell D, Bréon F-M, Collins W, Fuglestvedt J, Huang J, Koch D, Lamarque J-F, Lee D, Mendoza B,Nakajima T, Robock A, Stephens G, Takemura T, Zhang H (2013). Anthropogenic and Natural Radiative Forcing. 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 [Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 659–740, doi:10.1017/CBO9781107415324.018.

Pope VD, Gallani ML, Rowntree PR, Stratton RA (2000) The impact of new physical parametrizations in the Hadley Centre climate model - HadAM3. Clim Dyn 16:123-146.

Rougier J (2007) Probabilistic inference for future climate using an ensemble of climate model evaluations. *Climatic Change* 81:247, https://doi.org/10.1007/s10584-006-9156-9

Santer BD, Wigley TML, Schlesinger ME, Mitchell JFB (1990) Developing climate scenarios from equilibrium GCM results. Max-Planck Institute for Meteorology Report Number 47, Hamburg

Scott DW, Multivariate Density Estimation: Theory, Practice, and Visualization, Second Edition, John Wiley & Sons, 2015, doi:10.1002/9781118575574

Sexton DMH, Harris GR (2015) The importance of including variability in climate change projections used for adaptation. *Nature Climate Change* 5, 931-936, https://doi.org/10.1038/nclimate2705

Sexton DMH, Murphy JM, Collins M, Webb MJ (2012), Multivariate probabilistic projections using imperfect climate models, Part I: Outline of methodology, *Clim Dyn*, 38: 2513, https://doi.org/10.1007/s00382-011-1208-9

Taylor KE, Stouffer RJ, Meehl GA (2012). An overview of CMIP5 and the experiment design. Bull. *Am. Met. Soc.* 93: 485-498.

Trenberth KE (2015). Has there been a hiatus? *Science* 349:691-692.

Watterson IG (2008) Calculation of probability density functions for temperature and precipitation change under global warming. J Geophys Res 113:D12106. doi: 10.1029/2007JD009254

B.6 Estimating internal variability with the EC-Earth initial-condition ensemble (BNV) - KNMI

Authors: Hylke de Vries, Geert Lenderink

Short summary of method

One of the WP2 deliverables in EUCP is to develop methods that constrain multi-model pdfs of climate variables under climate change. Identification of and accounting for internal variability is important as it influences trend estimates even on longer time scales. Here we describe several methods to estimate internal variability and apply them to a 16-member initial-condition ensemble of climate projections obtained with EC-Earth GCM. Our primary target variables are summer precipitation and temperature, aggregated to three different spatial scales.

Introduction

Internal variability in the climate system is one of the factors limiting the determination of trends and the attribution of extreme events in a world that is warming due to anthropogenic effects. The amplitude of the internal variability depends strongly on the variable of interest and on the spatial and temporal aggregation scale. Here we determine the internal variability present in the climate change signal of summer (JJA) temperature and precipitation over Europe. Three levels of spatial aggregation are considered (see Fig. 1): *European*: Covering the land area of three SREX-regions (NEU, CEU and MED combined). *Regional*: Covering the land area of the three SREX-regions, but now separately. *Local*: Specific locations (Fig. 1). We explore a number of methods to identify the internal variability (more details in the next section). Typical research questions: 1) Are the results sensitive to the choice of method? 2) What is the added value of having a large ensemble? [In other words, how well can we estimate the internal variability from a single realisation?] 3) To what extent does the 'true' mean fall within the natural variability estimated from a single member, and in what way does this information help us in multi-model analyses? And finally 4) Do the results depend on the level of spatial aggregation?

Data and method

We use data from a large initial-condition ensemble of climate simulations obtained with the EC-Earth GCM (16 members that only differ in their initial condition). Temperature and precipitation fields have first been interpolated to a 2.5x2.5 degree regular longitude latitude grid using conservative remapping. In this stage of the project, we consider only two 20-year periods: a reference period (1995-2014) and a mid-of-century period (MOC, 2041-2061). For completeness we also include an end-of-century period (EOC, 2081-2100). More specifically, in the methods below, we assume that only these time slices are available.

To estimate the internal variability at the 20-year time scale, we compare several methods:

1. *m.empirical*. In this method we estimate the percentiles (P05 to P95 in steps of 5%) of the pdf directly from the raw 16-member ensemble. An alternative to this method, referred to as *m.gaussian*, is to fit a Gaussian through the ensemble using the sample mean and standard deviation and compute the percentiles from there. Both methods assume that 16 members are enough to robustly sample the distribution of 20-year means.

2. *m.bootstrapI*. If each simulation would be obtained from a different model (such as for example when reconstructing a pdf from a CMIP5 plume), one could try to estimate the 'missing' internal variability, by bootstrapping individual members. In *m.bootstrapI* we bootstrap each ensemble member *individually* (i.e., by only resampling from that member) and reconstruct the pdf from the total collection of bootstrap samples. Typically we take O(10^4) resamples of the same length as the period under consideration. This method is regarded as our baseline pdf. The pdf obtained using this method is generally wider than that obtained with the previous method.
3. *m.bootstrapN*. Similar to the previous method, but now we explicitly use the entire ensemble, to create a multi-member bootstrap, where years are sampled with replacement from all members. As before, we take O(10^4) resamples of the same length as the period under consideration.
4. *m.bootstrap1*. Take only one member and bootstrap. Years or seasons are considered independent and are sampled with replacement from one single member. This method should be contrasted to *m.bootstrapN* to explore the added value of having a large ensemble.

Some notes on these methods are now given. The first two methods (*m.empirical* and *m.gaussian*) retain all correlations that exist within individual realisations (i.e., the differences are computed per member). Such time-correlation preserving methods make sense in a multi-model ensemble where each member is taken from a different model (with e.g., different implementation of physical processes, a different climate sensitivity), or if there are clear physical mechanisms that relate a bias with respect to the ensemble average in the current climate, to a bias in the response. For a single model ensemble, however, the physical motivation of retaining the time-correlations is not clear, and may even, especially for small ensembles, introduce spurious results. For example a member that - due to internal variability - is relatively warm in the current climate (wrt ensemble average) is less likely to still be warm in a future period (regression to the mean). In time-slice experiments, therefore, one generally finds a negative correlation between present-day temperature and future trend. To avoid this problem, we have included the other methods that are based on bootstrapping. Because bootstrapping relies on sampling with replacement, time-correlations are destroyed in these methods. In the baseline approach (*m.bootstrapI*) the members are bootstrapped individually, similar to what one would do if the ensemble were not derived from a single model, but from an ensemble of models with single realisations (e.g. CMIP5 or CMIP6). In *m.bootstrapN* we use that the ensemble is generated by the same model, and resample from all members.
Finally, note that the methods above have been designed for use with time-slice experiments. They do not use the fact that the time slices (in the case of the EC-Earth simulations) are taken from much longer transient simulations. If the entire transient is available, other methods could be employed to estimate internal variability. An example of such a method is to use regression against time or another reasonably smooth variable such as global temperature.

Results (some examples)

Figures 2 and 3 show the results of the different methods for summer precipitation in the largest domain (EUR). Similar figures are obtained for the other domains (not shown, available upon request). The top-left panel shows histograms of the precipitation change, based on the 16 realisations of the GCM. The second panel shows a scatter plot of precipitation average in the historic period and the subsequent response. Linear regressions are indicated. The remaining panels show the cumulative distributions obtained using the different methods, with the horizontal lines representing the P10, P50 and P90 levels. For each of the periods there is substantial internal

variability (top-left). Even at the largest spatial scale, the differences of individual members with respect to the ensemble mean exceed 10%. At the largest scale (EUR) there is a weak indication of a relation between bias and future response in summer (top-middle), although the regressions are not found to be statistically significant. At smaller spatial scale this relation becomes stronger (not shown). However, we believe that this relation is almost exclusively caused by internal variability. For temperature the relation is even more clearly visible (not shown): warm members in the current climate experience less warming as a result of regression toward the mean. The other panels describe the different methods. Even at the largest spatial scale there are noticeable differences. For example, the 1-member method, (*m.bootstrap1* of panel f), does not produce a meaningful estimate of the 'true' mean (as approximated by the ensemble average) and neither of its change (at least not for precipitation). However, the *m.bootstrap1* method *can* be used to derive a meaningful approximation of the amplitude internal variability (P10-P90 range), although it will underestimate the 'true' amplitude in most cases. The three other methods make use of the entire ensemble. The *m.empirical* method is the most straightforward and estimates the P10-P90 range by ordering and counting. Being non-parametric, it might be not so sensitive to outliers, yet still suffers from the relatively small ensemble size. The parametric method *m.gaussian* fits a Gaussian through the ensemble data using estimations of the mean and standard deviation. The result looks much like a smoothed version of the *m.empirical* method (not shown). Neither of these two methods can account for the fact that the number of realisations (16) is much smaller than the number of possible states. Both methods also preserve the correlations in time (i.e., the future minus present-day differences are computed per member). These time correlations are destroyed in the *m.bootstrapI* and *m.bootstrapN* methods (panels (d and e)). The first is our baseline method and gives a rather broad uncertainty range. The second produces a narrower pdf, by using the fact that all members derive from the same model configuration and therefore should have a similar 'true' model climate. The last method obviously gives the most robust results. It is the preferred method if there is no physically justifiable reason to assume a physical relation between historic climatology and future change. The results are summarised in box-whisker plots (Fig.3 and Figures 4-5 for the other spatial aggregation levels). From these one can also examine whether the true mean generally falls within the uncertainty estimated from the single-member m.bootstrap1 method. For the change of summer precipitation, the chosen random member (we took ensemble member #1) was able to produce a spread wide enough to cover the central value of the full ensemble. However, further research is required to investigate the robustness of this statement. Finally, the results for temperature change are similar to that obtained for precipitation change, yet with much larger signal-to-noise ratios.

Conclusion

In this document we have presented a number of methods to estimate internal variability. If there is no physically motivated reason to assume that present-day climatological values constrain the future response, *m.bootstrapN* method gives a robust estimate of the internal variability (estimated here as the P10-P90 range of 20-year means). A baseline bootstrap method *m.bootstrapI* that one could use if all ensemble members are obtained from different models (such as for example the CMIP5 plume) gives a wider range with possibly a shifted mean. The *m.empirical* and *m.gaussian* method may show spurious results in the change-pdf, especially if the
ensemble has only a limited number of members. The estimate of internal variability that one can obtain from a single-member bootstrap (*m.bootstrap1)* underestimate the internal variability as obtained with any of the other methods, yet at the 20-year time scale under consideration does

provide a meaningful first guess lower bound. However, its central estimate is often quite far from the true model mean as estimated from the entire ensemble.

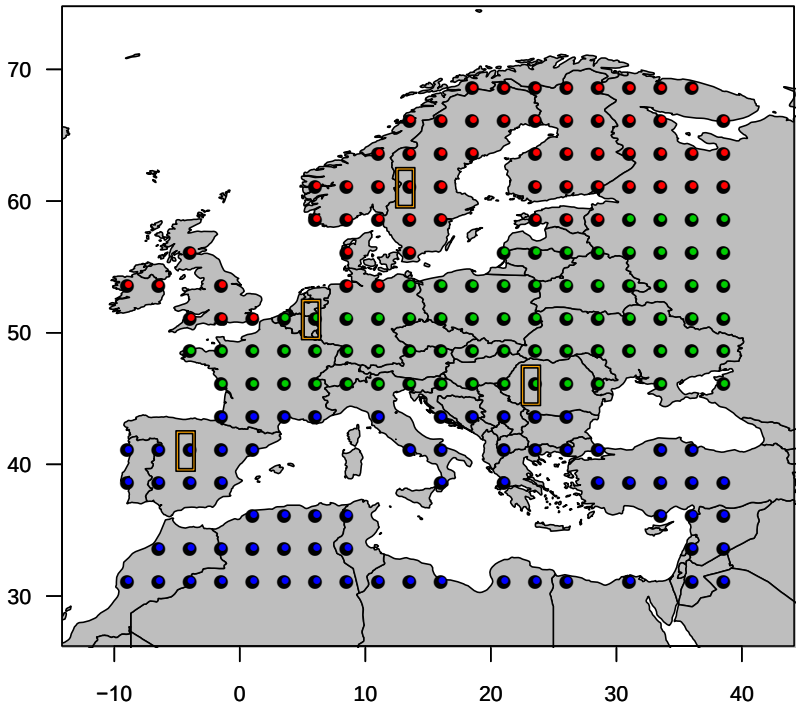


Figure 1: Analysis grid and the three SREX regions (NEU, CEU and MED, a land-sea mask has been applied) and the location of the special points in orange).

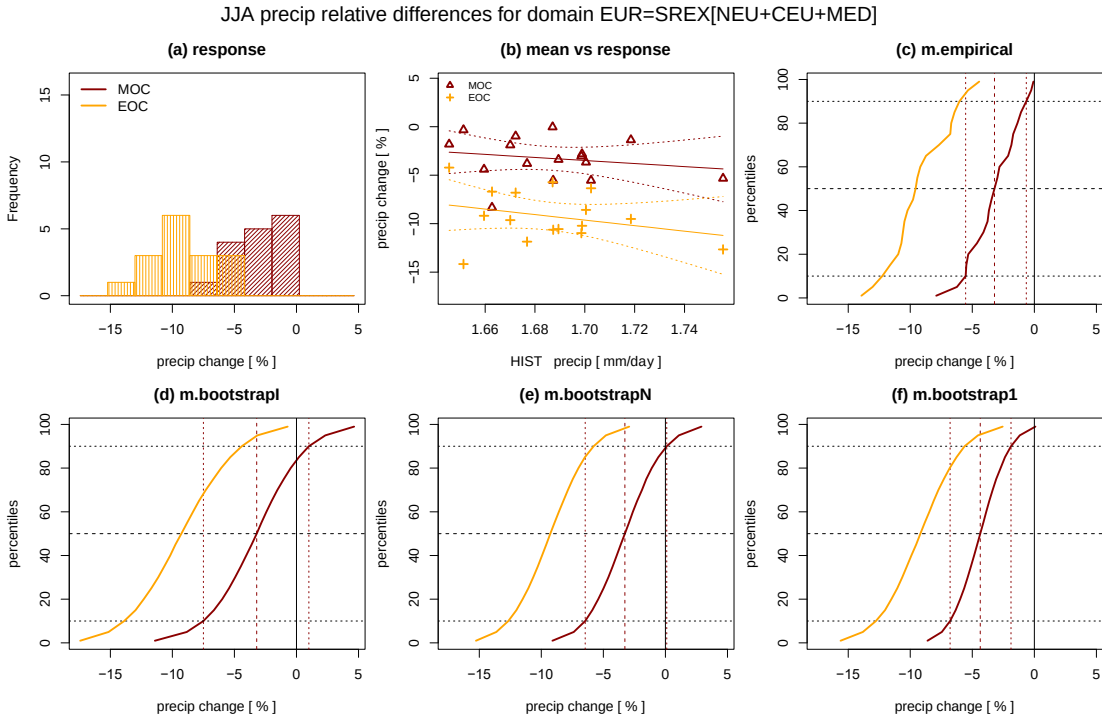


Figure 2. JJA precipitation change over EUR domain. (a) Histogram of precipitation differences for the different time slices. (b) Scatter :lot between current climate mean and future change. Regressions including confidence intervals are also indicated. (c-f) The different methods used in the paper to determine the amplitude of internal variability (P10-P90, horizontal lines).

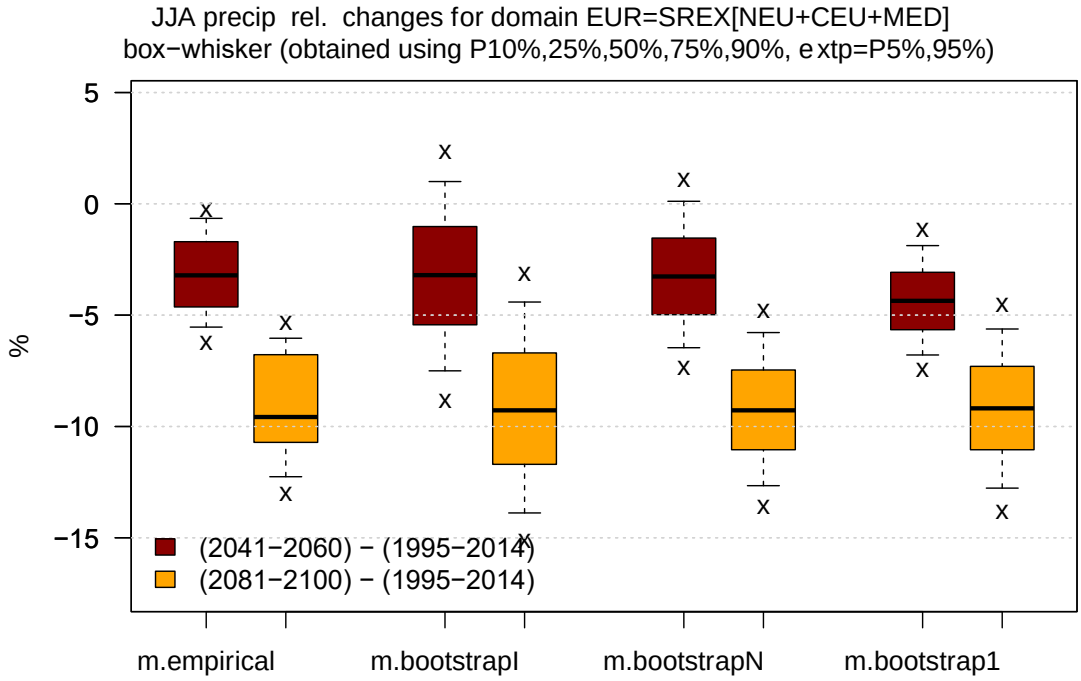


Figure 3: JJA precipitation-change over EUR domain in terms of box-whisker plots. The quantiles on which the box-whisker is based are indicated. The crosses represent the P05 and P95.

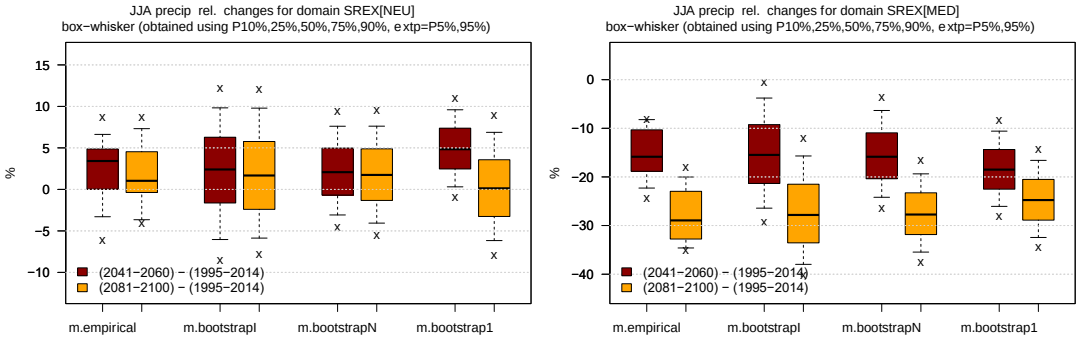


Figure 4: As in Figure 3, but applying the methods to regional scale, SREX NEU (left) and SREX MED (right).

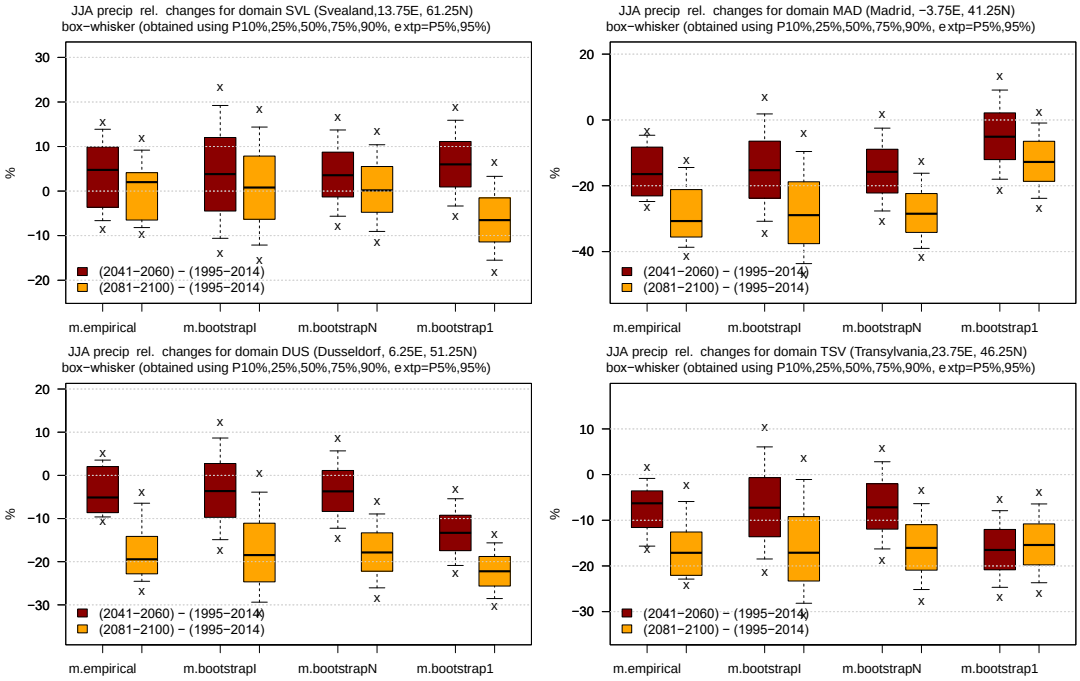


Figure 5: As in Figure 3, but applying the different methods to the local scale.

B.7 Calibrating large ensemble projections using observational data (CALL) - UOXF

Authors: C. H. O'Reilly, Daniel Befort, and Antje Weisheimer.

Short summary of method

A method of calibrating the output of large single model ensembles has been developed. The method involves fitting seasonal ensemble data to observations over a reference period and scaling the ensemble signal and spread so as to optimize the fit over the reference period. This scaling is then applied to the future (or out-of-sample) projections. Several simple calibration methods have been tested and several similar methods give indistinguishable results so the simplest of these methods, namely Homogenous Gaussian Regression, is selected. An extension to this method is that applying to dynamically decomposed data – in which the underlying data is separated into dynamical and residual components – is found to the reliability of the calibrated projections. The calibration methods were tested and verified using an "imperfect model" approach using the historical/RCP8.5 simulations from the CMIP5 archive.

Key references

O'Reilly et al. (*in prep.*) [EUCP WP2 publication]

Data

The ensemble projection data is taken from the CESM Large Ensemble (LENS; Kay et al., 2015). Observational data for temperature and precipitation was taken from the CRU-TS v4.01 gridded surface dataset (Harris et al., 2014). The data used for out-of-sample verification was taken from the CMIP5 archive. All datasets are used between 1920-2060, except for the observations, which are taken over the reference period 1920-2016.

Details of method

Seasonal indices for a given variable (e.g. surface temperature or precipitation) are calculated for various regions over Europe, following the common WP2 protocol. Indices for the LENS dataset, which has 40 ensemble members, is separated into an ensemble mean signal and spread about that signal. Over the reference period, 1920-2016, the first step is to remove the climatological mean over the reference period. The scaling parameters for the ensemble mean and spread are then fit to target reference dataset (e.g. the observational timeseries). The method for the for to the target dataset follows the methodology for *Ensemble Model Output Statistics* (EMOS) or Non-homogenous Gaussian Regression, typically applied to seasonal/decadal forecasting data (e.g. Gneiting et al., 2005; Samson et al., 2016). This fit involves calculating scaling parameters for the ensemble mean and variance using a numerical optimization procedure which maximises the likelihood function over the reference period. Further tests revealed that including information about the varying ensemble spread did not improve the calibration method, so *Homogenous Gaussian Regression* (HGR; Barnston & Tippett, 2008) is used for the final calibration. HGR involves scaling the ensemble mean signal and dressing the scaled ensemble mean with an appropriate Gaussian probability density function. In addition to EMOS and HGR we also tested the *Variance Inflation* method (Doblas-Reyes et al., 2005) which yielded broadly similar results, however, we have chosen to use HGR as it is simpler parameterized approach that is slightly more transparent. To represent

uncertainty in the fitting procedure, a bootstrap-with-replacement method is used over the reference period to fit the HGR scaling parameters. This process is performed 1000 times and for each set of the fit parameters, one ensemble member was produced by sampling at random from the calibrated Gaussian distribution for each year. The number of calibrated timeseries that can be produced is clearly unlimited but 1000 was found to be sufficient to generate clear projections. For comparison, a 1000 member uncalibrated ensemble is calculated from a Gaussian distribution fit to the raw LENS 40-member distribution for each season individually, though this is mainly for presentation purposes as the verification statistics are not significantly changed.

In addition, we also applied these calibration methods to dynamically decomposed timeseries. The rationale for doing this was that it is possible that a climate change signal in the target dataset may be masked by large-scale dynamics and that by decomposing the data we may avoid conflating signal (or noise) from the internal dynamics with a thermodynamical climate change signal. Such a dynamical decomposition involves splitting the full timeseries as *FULL = DYNAMICAL + RESIDUAL.* The dynamical component was calculated for all model ensemble members and observations following the analog method of Deser et al. (2012). In this method, sea-level pressure (SLP) anomaly fields for each month in each are fit using other SLP anomaly fields from the corrensponding month from other years over the reference period. This regression fit yields weights which are then used to computed the associated dynamical surface temperature or precipitation anomaly. Each field can then be separated into a dynamical and residual component, as shown for the LENS decomposition in Figure 1. The regional dynamical and residual timeseries were fit to the corresponding dynamical and residual timeseries separately and these are then combined to give a full calibrated ensemble projection. This method is referred to as HGR-decomp.
The calibrated large ensemble projections produced using the HGR and HGR-decomp methods were tested in an “imperfect model” approach, using the model integrations from the CMIP5 archive. The motivation here is that if the calibration tends to improve the prediction of the future climate in a model in an out-of-sample sense, then it might be expected to improve the reliability of future climate projections when calibrated using observations. The first ensemble member for the 39 CMIP5 models (which have historical/RCP8.5 simulations) were used to calibrate the LENS data using the HGR method over the reference period. The uncalibrated and calibrated ensembles are then verified against the out-of-sample future projections over the period 2017-2060, shown for one cherry-picked CMIP5 model projection in Figure 2. For each model the root-mean-square-error (RMSE) was calculated using all 44 seasons over the verification period. As well as RMSE we also calculated the ensemble spread, the spread/error ratio and the continuous ranked probability score (CRPS). For the RMSE and CRPS, a useful calibration would reduce these, which would indicate a more skilful projection. A spread/error ratio of one indicates that a probabilistic forecast is reliable in a statistical sense, meaning that the spread of the prediction gives useful information about the likely scale of error, whereas spread/error ratios less that one and greater than one indicates, respectively, overconfident and underconfident predictions.
A summary plot of the verification statistics of the uncalibrated and calibrated LENS projections over the out-of-sample period 2017-2060 is shown in Figure 3. The verification against all of the 39 CMIP5 models is shown for 4 metrics. The calibration methods reduce the RMSE and CRPS for temperatures during summer and also reduce errors for precipitation in Northern Europe. During winter there is no clear improvement in skill but the calibration does not degrade the performance of the ensemble. The HGR and HGR-decomp methods are generally fairly similar in terms of projection

skill. However, the HGR-decomp calibration tends to have higher values of spread and this is manifested in spread/error ratios that are generally slightly closer to one than in the HGR calibration. The calibration acts to improve the spread/error for most regions and variables, with the spread/error consistently being close to one in the calibrated ensembles, indicating that the method results in more reliable projections over this period on seasonal timescales.

Example projections

Some example projections are shown in Figure 4 for several European regions. Probability density functions for 2041-2060 means calculated from these projections, relative to the 1995-2014 mean, are shown in Figure 5.

Imperfect model verification for the 2041-2060 means are shown in Figure 6. Here, the verification is calculated across the uncalibrated and calibrated forecasts for the 39 CMIP5 models. The calibration lowers the error for surface air temperature in all regions, however, there is little improvement for precipitation apart from in Northern Europe. The reliability - measure in terms of spread/error ratio - of the projections for surface air temperature are greatly improved using the calibrated ensemble, though again there is no improvement in the reliability for precipitation.

Key assumptions

- This method assumes that past evolution of the observations that can be captured by the ensemble datasets contain meaningful information for the future evolution.
- The calibrations seems to be useful for improving the reliability of future projections in imperfect model tests, but this assumes performance in this setting is a useful basis for applying to observations which may not be the case.

Limitations

- One important limitation is that the calibration is univariate, such that variables that are likely related are not calibrated in a necessarily consistent way. However, the underlying ensemble will to some extent reflect these physical constraints.
- The method is limited by the ability of the underlying large ensemble. For example, if the underlying ensemble has no signal that correlates with the observed variability over the reference period then the future mean changes will be reduced towards zero.
- For this reason the method is likely to perform better on climate signals that have already begun to emerge from internal variability and less useful for other variables, though the calibration is not found to worsen the projections in these instances.

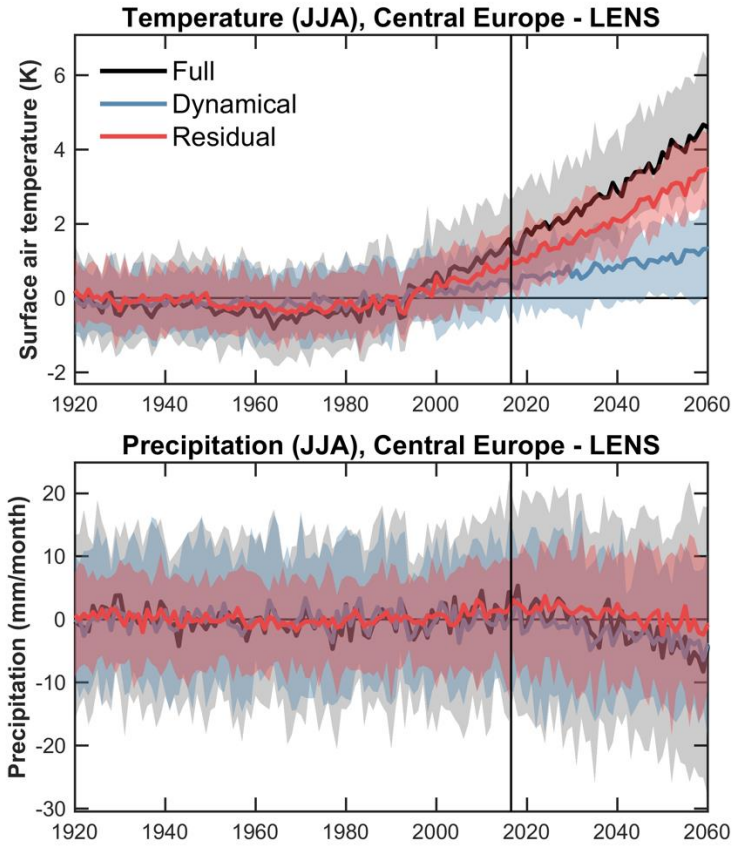


Figure 1: Example dynamical decomposition of the LENS dataset for the Central Europe region for summer (JJA). The shading shows the 90% range of the ensemble and the bold lines shows the median.

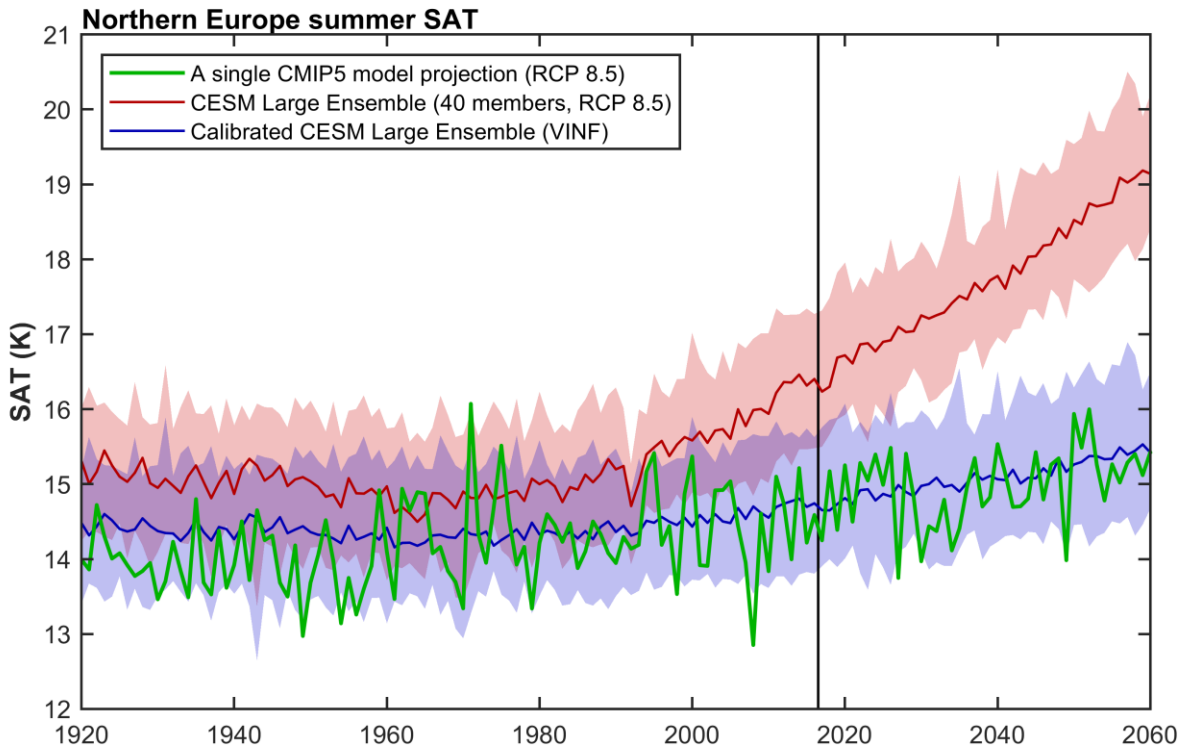


Figure 2: Example calibration for one (cherry-picked) ensemble member from the CMIP5 RCP8.5 dataset. This example shows the Variance Inflation method, which produces similar results to the HGR used to produce the final projections. The shading shows the 90% range of the ensemble and the bold lines shows the median.

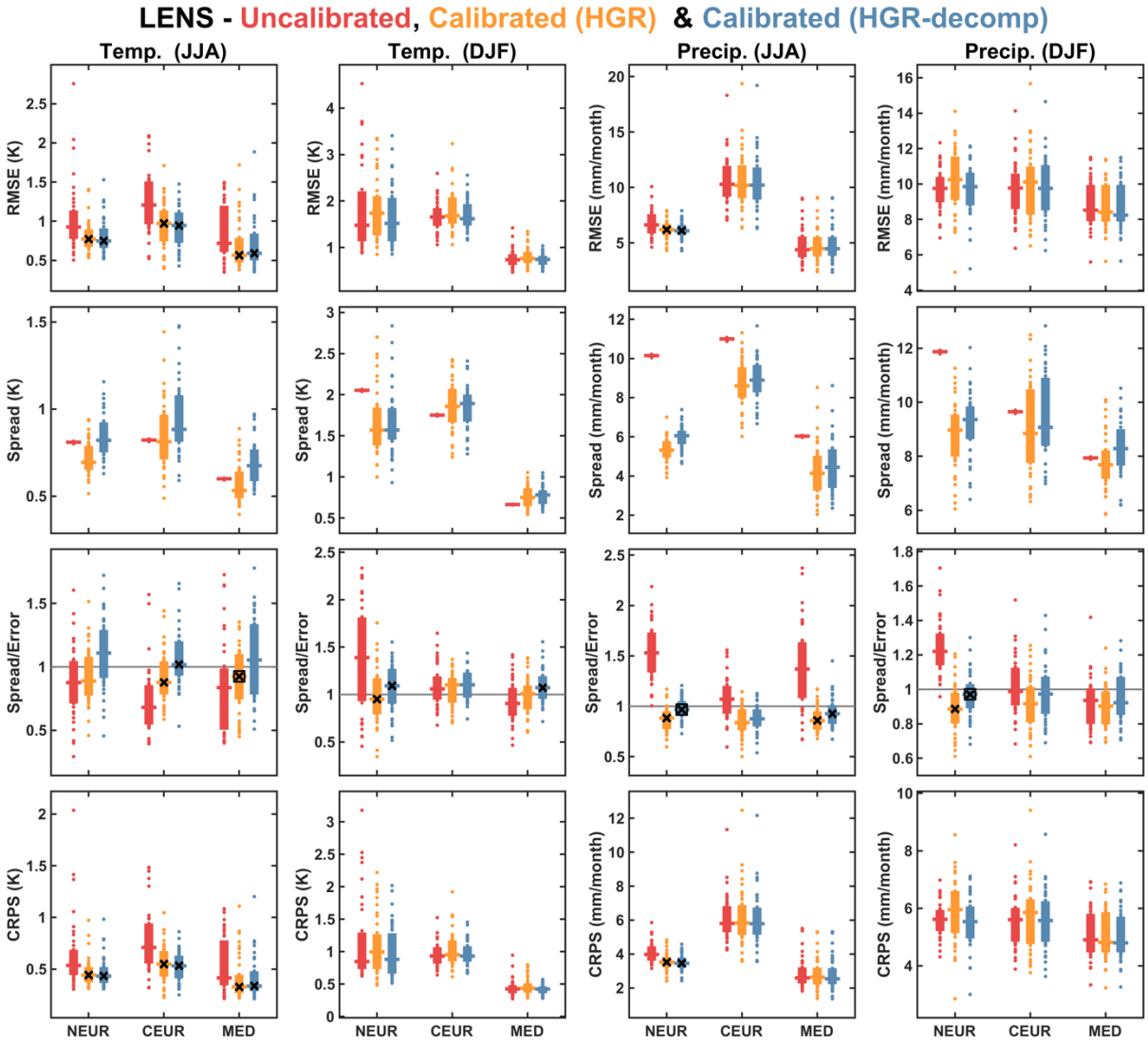


Figure 3: Verification statistics for the uncalibrated and calibrated LENS projections against all 39 CMIP5 simulations. The statistics are calculated for the ensemble against the target CMIP5 dataset over the period 2017-2060. Dots indicate individual CMIP5 verifications, the thick vertical lines show the interquartile range across the CMIP5 models and the horizontal lines indicate the median value. Black crosses show where the calibrated distribution is significantly better than the uncalibrated projection at the 90% significance level. Black squares show where a calibrated ensemble method is significantly better that the other calibration method (i.e. HGR vs. HGR-decomp). Significance values were estimated using the non-parametric Mann-Whitney U-test.

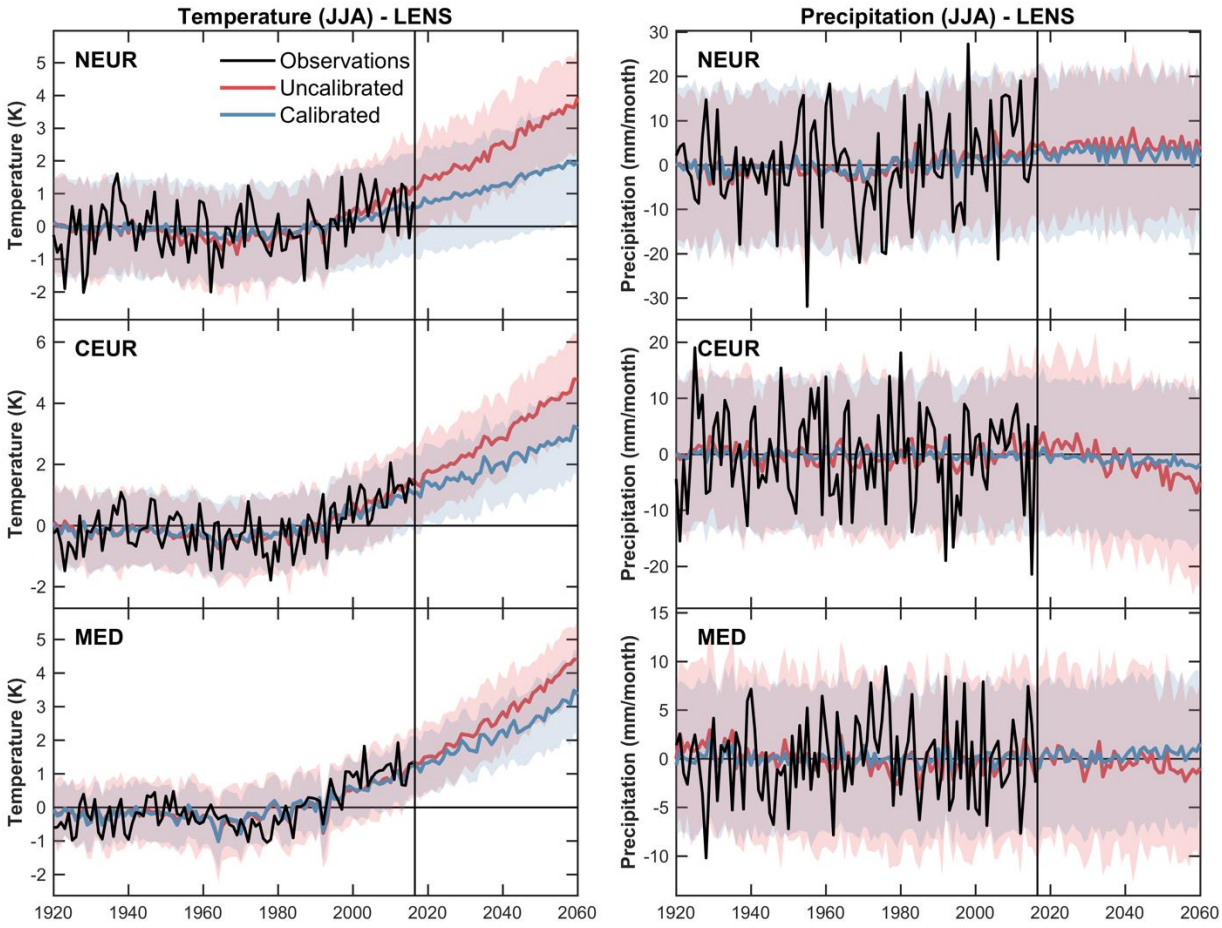


Figure 4: Uncalibrated and calibrated projections, using observational indices from the CRU-TS as the reference dataset. The shading shows the 90% range of the ensemble and the bold lines shows the median. Anomalies are plotted relative to the 1920-2016 reference period.

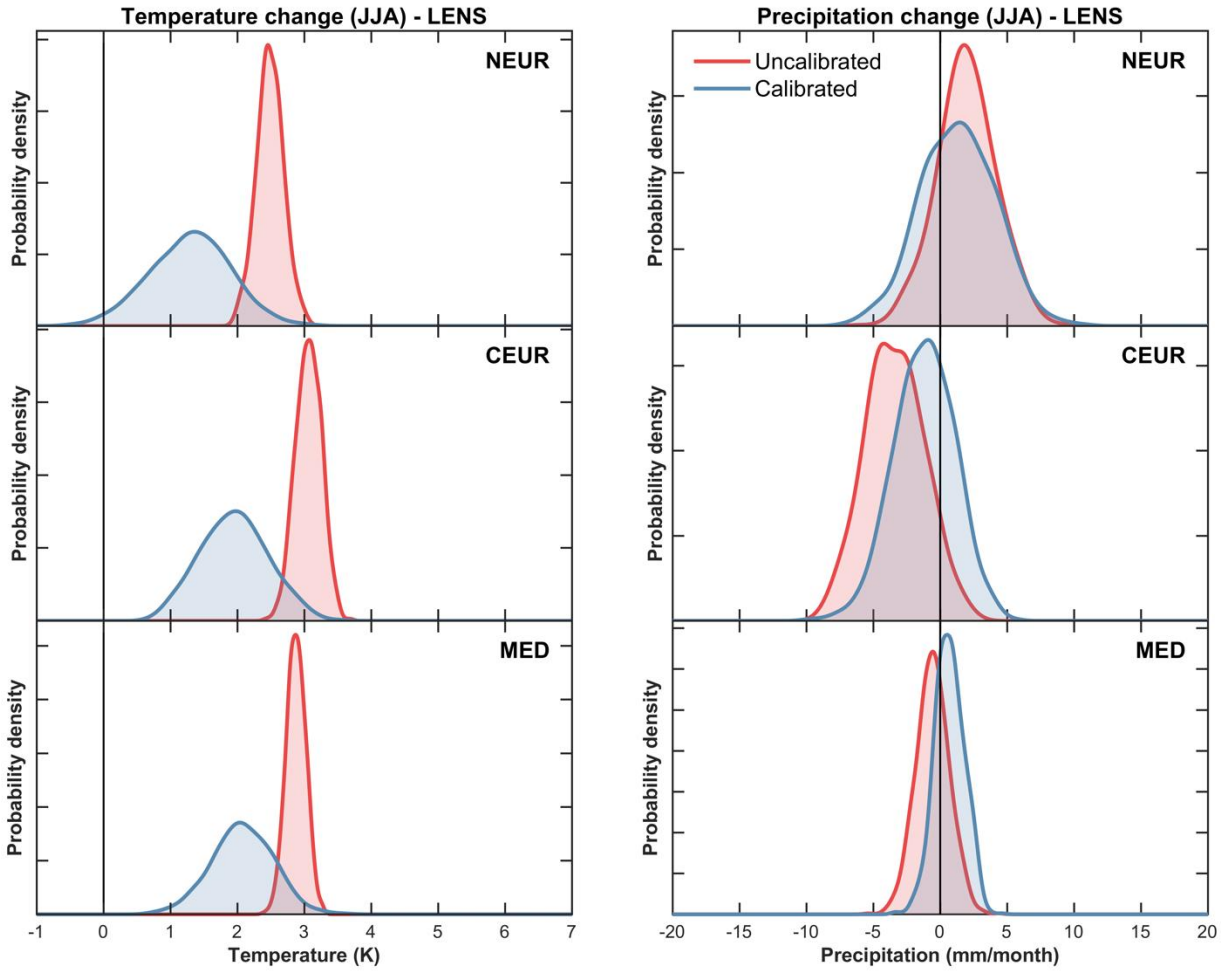


Figure 5: Uncalibrated and calibrated projected mean changes, using observational indices from the CRU-TS as the reference dataset. Projections are for the mean 2041-2060 climate relative to the 1995-2014 mean, following the EUCP WP2 protocol.

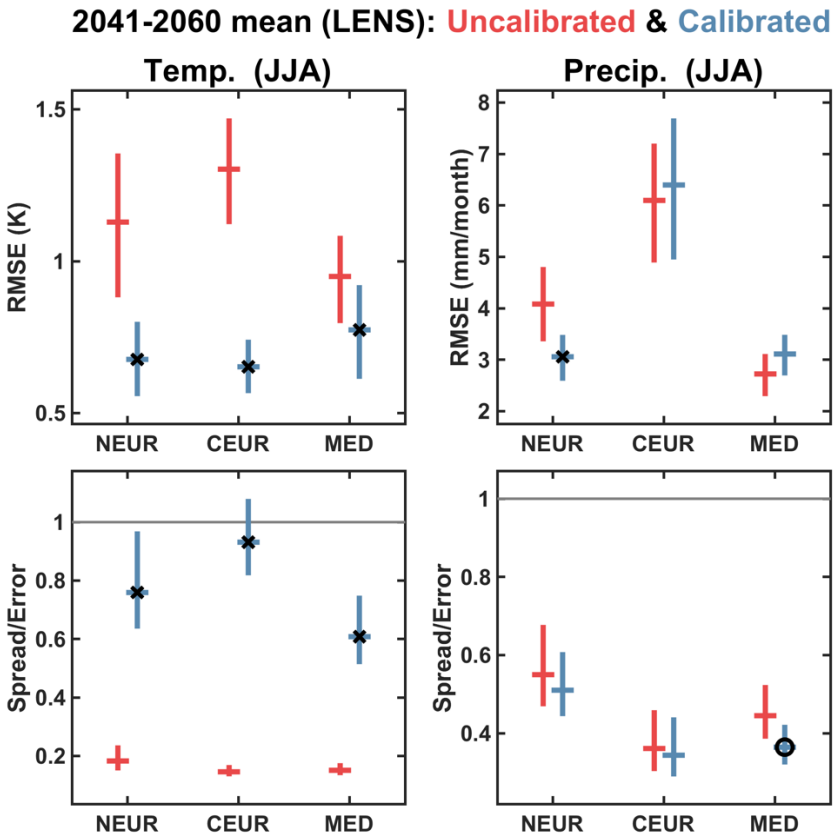


Figure 6: Verification statistics for the 2041-2060 mean projections calculated from the application of the HGR-decomp calibration method to the 39 CMIP5 models. The vertical lines show the 90% confidence interval based on a bootstrap-with-replacement resampling across the 39 model realisations. Black crosses show where the calibrated distribution is significantly better than the uncalibrated projection at the 90% significance level. Black circles show where the calibrated distribution is significantly worse than the uncalibrated projection at the 90% significance level.

References

Barnston & Tippett (2008). Skill of multimodel ENSO probability forecasts. *Monthly Weather Review*, 136(10), 3933-3946.

Deser et al. (2012). Forced and internal components of winter air temperature trends over North America during the past 50 years: Mechanisms and implications. *Journal of Climate*, 29(6), 2237-2258.

Doblas-Reyes et al. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting—II. Calibration and combination. *Tellus A: Dynamic Meteorology and Oceanography*, 57(3), 234-252.

Harris et al. (2014). Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset. *Int. J. Climatol.,* 34: 623-642.

Gneiting et al. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5), 1098-1118.

Kay et al. (2015). The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society*, 96(8), 1333-1349.

Samson et al. (2016). Best practices for postprocessing ensemble climate forecasts. Part I: Selecting appropriate recalibration methods. *Journal of Climate*, 29(20), 7247-7264.

B.8 Ensemble analysis of probability distributions (ENA) - CMCC

Authors: Marianna Benassi, Silvio Gualdi, and Antonio Navarra

Introduction

We want to explore different single model large ensembles, which are generally considered as preferred tools to assess uncertainty in climate signals due to internal variability. In fact, the spread across the simulations included in this kind of ensemble is generally considered as a proxy for internal variability, since by construction differences in the model structure and parametrization are not taken into account. We take advantage of the NCAR Large Ensemble (NCAR-LENS, Kay et al. 2015) and of the MPI Grand Ensemble (MPI-GE, Maher et al. 2019). In the framework of EUCP-WP2 activities, and in particular of the uncertainty quantification method inter comparison, our analysis will provide a benchmark information aiming to emphasize the role of internal variability in climate projections.

One potential application of this kind of ensemble is to estimate the ensemble size needed for a specific propose. In this document, some preliminary results will be presented showing how the estimate of the mean signal and of the related uncertainty varies with the ensemble size.

Data and Method

NCAR-LENS is a 40-member ensemble spanning the historical period from 1920 to 2005, and following the RCP8.5 scenario for the period 2006 to 2100. This set of simulations is run with the Community Earth System Model (CESM) version 1 with the same model configuration exploited in the CMIP5 effort (CESM-CAM5 setup). The MPI-GE is a 100-member ensemble using MPI-ESM1.1 and forced with the CMIP5 simulation protocol in the historical period (1850-2005) and in the scenario one (2006-2099). For consistency, also in the MPI-GE case, we will focus on the scenario RCP8.5.

Following the common protocol defined in the frame of EUCP-WP2 activities, we consider the boreal summer (JJA) temperature and precipitation changes. The changes are computed between the 20-year average over the mid-of-century period (2041-2060) and the present-day period (1995-2014). For temperature, absolute changes are considered, while for precipitation we take into account relative changes.

In Figure 1 the absolute temperature change signal over Europe from NCAR-LENS (in the green box) and from MPI-GE (in the orange box) is presented. For each model the ensemble mean, standard deviation, 10^{th} and 90^{th} percentiles are shown. Despite the prevailing warming signal, some differences across the two models may be found. Generally, MPI-GE shows a less intense warming both in the average and in the extremes of the ensemble distributions, while the variability results most pronounced, in particular in the southern part of the domain. In both cases the most pronounced warming is found over the Balkan region.

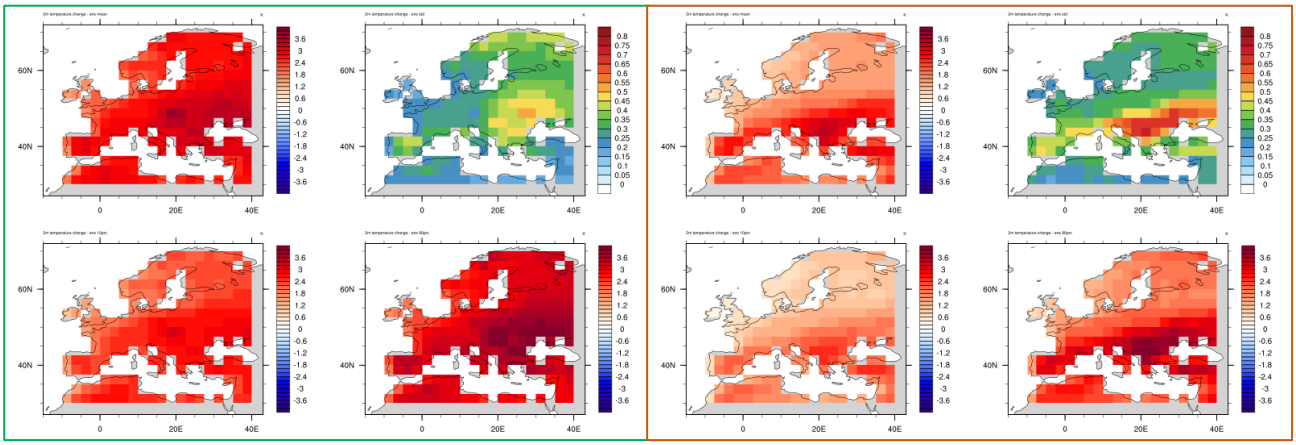


Figure 1: In the green box: NCAR-LENS absolute JJA temperature change (mid-of-century – present-day) ensemble mean (top left), ensemble standard deviation (top right), ensemble 10th percentile (bottom left), and ensemble 90th percentile (bottom right). In the orange box the same plots are presented for MPI-GE.

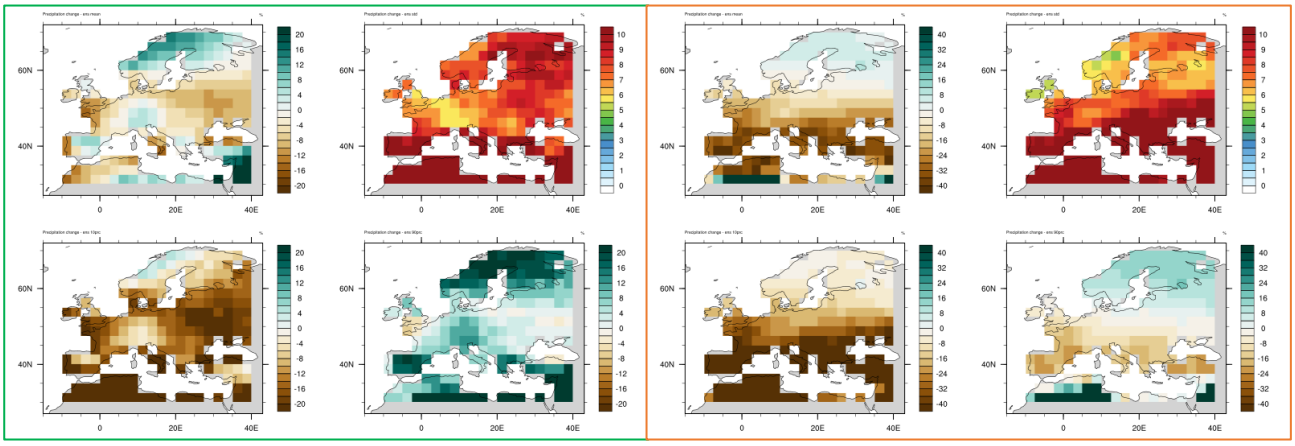


Figure 1: In the green box: NCAR-LENS relative JJA precipitation change (mid-of-century – present-day/present-day) ensemble mean (top left), ensemble standard deviation (top right), ensemble 10th percentile (bottom left), and ensemble 90th percentile (bottom right). In the orange box the same plots are presented for MPI-GE.

In Figure 2 we show the analogous map for the relative precipitation change signal: also in this case, for each model the ensemble mean, standard deviation, 10th and 90th percentiles are displayed. Here the comparison is less straightforward: while for MPI-GE the mean precipitation change shows a clear meridional gradient with a much drier Mediterranean and a slightly wetter Scandinavia peninsula, the pattern in the NCAR-LENS case exhibits some positive anomaly over central Europe, while the negative anomaly over the Mediterranean is less pronounced. Differences may be found also in the extremes, with a drier Mediterranean for MPI-GE in both the extremes, and a nearly opposite pattern for NCAR-LENS.

In this framework, the spatial aggregations of interest consist of the European domain (EUR) - defined as the combinations of the land area over the three SREX regions (Northern Europe NEU, Central Europe CEU, Mediterranean sector MED)-, the three SREX regions taken separately, and four specific locations (*i.e.* some selected grid points). In the pre-processing phase both the datasets have been extrapolated with conservative remapping to a 2.5x2.5 degree regular grid and a proper

masking has been applied to define the different subdomains. For the single location case, the values of the field of interest have been extracted knowing the specific coordinates.
Once the period and seasons have been extracted and the proper time and spatial averages applied, the absolute change for temperature and the relative change for precipitation has been computed, allowing to derive the empirical distributions of the two quantities from the two ensembles. In order to assess the uncertainty in the estimate of the empirical distribution parameters a *bootstrap resampling* have been performed. For both the datasets separately, an extraction with replacement has been repeated 1000 times, allowing to define for different distribution parameters a 95% confidence level interval.
The bootstrap approach has allowed us to characterize also the behavior of different parameters of the distribution (*i.e.* mean and standard deviation) as a function of the ensemble size. For each possible subsample size *n* (*n=2, 3, …, N*) a *bootstrap subsampling* has been performed. In other words, 1000 couples, 1000 triplets... 1000 n-tuples, have been randomly selected from the full ensemble allowing for repetition in order to have for each *n* a sample of 1000 independent elements (couples, triplets, and so on). In order to characterize the approximated distributions, for each of *the 1000 n-tuples* an estimate of the mean and of the standard deviation has been computed. For each *n,* these 1000 estimates allow to define a distribution of the mean and of the standard deviations values, which we interpret as characteristic of that *n*. This approach has been adopted for both the datasets, and some illustrative results are shown in the next section.

Results

For the sake of simplicity just the regional cases will be commented in this document. In the box plots reported in Figure 3 the empirical distributions for the global and regional changes of surface temperature are shown for the two ensembles. The shaded boxes to the right of each box plot represent the 95% confidence level interval for the median, the 5^{th} and the 95^{th} percentiles. The overall features noticed from the spatial pattern may be found also for the average regional changes. The different subdomains are substantially affected by internal variability, showing a much more dispersed distribution than the reference global average case. NCAR-LENS is warmer than MPI-GE, with major differences found for the mean changes in the Northern and Central Europe cases. For Northern Europe, the spread of the distribution is higher for NCAR-LENS than for MPI-GE, vice versa for the other cases. Anyway, for all the domains, the standard deviation values for the two ensembles are generally comparable.
In Figure 4, following the same convention, the empirical distributions for the global and regional precipitation changes with the respective 95% confidence level interval are reported. Some differences may be detected across the two ensembles, both in the mean signal and in the variability. In both the models Northern Europe case shows a neutral or slightly positive change signal, with NCAR-LENS being characterized by a higher dispersion. Both Central Europe and the Mediterranean sector show a mean drying signal, more enhanced in MPI-GE. For the Mediterranean case, despite the huge spread, in MPI-GE also the extremes are characterized by negative values, while in NCAR-LENS the spread of the distribution is such as to include some wet anomalies.

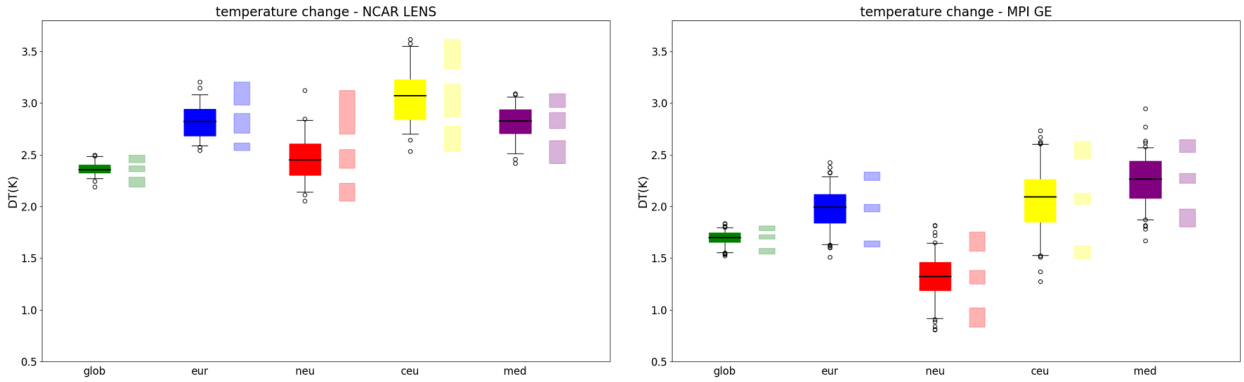


Figure 2: In the left panel: NCAR-LENS absolute JJA average temperature change empirical distribution for the global domain (green), the EUR domain (blue), the NEU domain (red), the CEU domain (yellow), and the MED domain (purple). The boxplot shows the 25th-75th percentile span and the median, while the whiskers cover the 5th-95th percentile interval. The shaded boxes represent the 95% confidence level interval for the median, the 5th, and the 95th percentile computed with a bootstrap extraction. In the right panel the same convention is adopted for MPI-GE data.

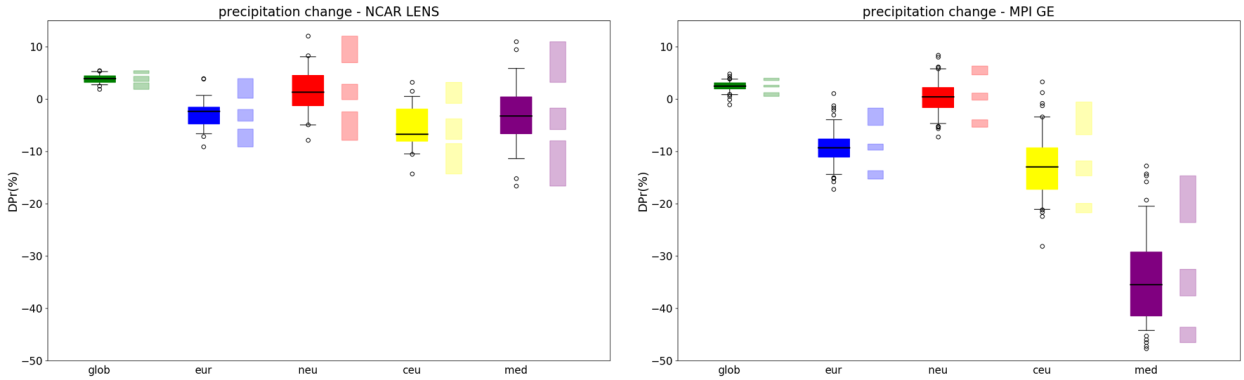


Figure 3: In the left panel: NCAR-LENS relative JJA average precipitation change empirical distribution for the global domain (green), the EUR domain (blue), the NEU domain (red), the CEU domain (yellow), the MED domain (purple). The boxplot shows the 25th-75th percentile span and the median, while the whiskers cover the 5th-95th percentile interval. The shaded boxes represent the 95% confidence level interval for the median, the 5th, and the 95th percentile computed with a bootstrap extraction. In the right panel the same convention is adopted for MPI-GE data.

For the bootstrap subsampling exercise, some results from NCAR-LENS regional temperature change case are presented in Figure 5. In the left panel, it is straightforward to observe the rapid convergence of the mean estimates, which can be interpreted as a direct consequence of the Central Limit theorem. In the right panel the standard deviation analysis is shown. If we consider this parameter as an indicator for internal variability uncertainty, we see how taking few ensemble members obviously causes an underestimation of uncertainty. The distribution of the standard deviation estimates, initially positively skewed, tends to become more symmetric as the size of the ensemble increases. The plateau is reached for $n \approx 20$.

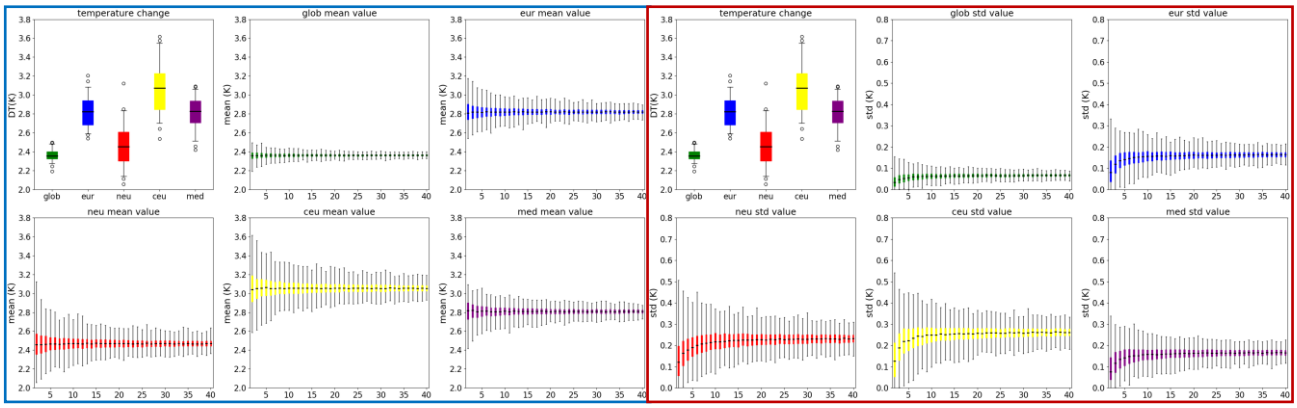


Figure 4: In the blue box: JJA average temperature change distribution for NCAR-LENS and bootstrap estimate of subsample temperature change means. In the top left panel, the empirical distribution of the JJA average temperature change for the global and the SREX domains is reported, with the same convention adopted in the boxplot of figures (3). In the other panels the distributions of the means from the bootstrap subsamples are shown. Here the boxplots span the 25th-75th percentile range, while the whiskers cover from the maximum to the minimum values of the distribution. The black line represents the median of the distribution. In the red box the same convention is adopted, but the distributions of the standard deviations from the bootstrap subsamples are shown.

The bootstrap allows also to empirically define a confidence level interval for the parameters taken into account, in this case the mean and the standard deviation of the regional (or local) temperature/precipitation change distribution. With the subsampling bootstrap this estimate should be repeated for each *n* (as in Figure 6), allowing to evaluate the extra-uncertainty induced by accounting for less ensemble members.

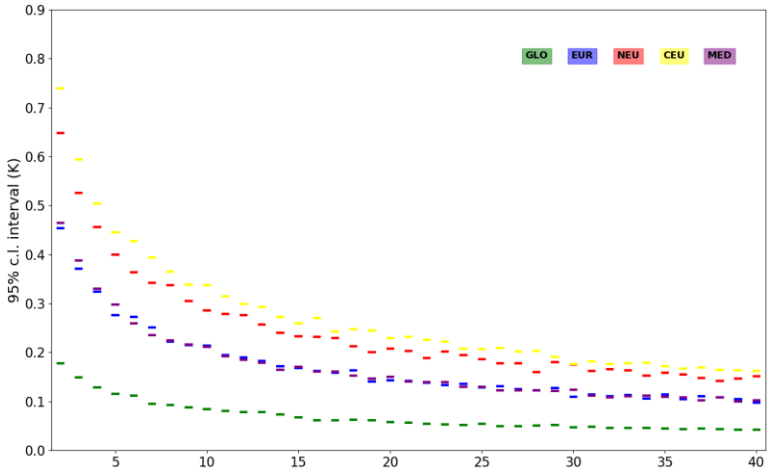


Figure 5: Amplitude of the 95% confidence level interval for the bootstrapped mean JJA average temperature change for each subsample n. The amplitude of the confidence level interval has been computed from the bootstrap distribution of the means of the JJA average temperature change (blue box in figure 5), taking into account the distance between the 2.5th and the 97.5th percentiles. The same color convention of the previous plot has been adopted.

Future Developments

Our analysis is focused on the characterization of single model large ensembles, with the idea of providing a benchmark assessment on the shape of the projected regional and local temperature and precipitation changes, by construction affected only by internal variability. The uncertainty due to internal variability for the different spatial aggregations and for the different variables is represented by the spread of the obtained distributions, which shows a substantial regional and model dependence.

We have included in this analysis some statistical characterization of the two ensembles, in order to determine how different statistics may depend on the ensemble size. Introducing a bootstrap procedure has allowed us to derive an assessment on the role of the ensemble size in estimating the statistics of interest. However, it should be noticed that, at the current stage, the analysis on the ensemble size has been focused on averaged properties. Thus, the obtained results in terms of convergence with the ensemble size are in line with what should be expected from the Central Limit theorem. In the future, the plan is to extend this analysis to examine the sensitivity to the ensemble size of the variability of complex structures both in space and time (*e.g.* teleconnection indices, EOF patterns *etc.*).

References

Kay, Jennifer E., et al. "The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability." *Bulletin of the American Meteorological Society* (2015).

Maher, Nicola, et al. "The Max Planck Institute Grand Ensemble: Enabling the Exploration of Climate System Variability." *Journal of Advances in Modeling Earth Systems* (2019).