npj | climate and atmospheric science

Article

Published in partnership with CECCR at King Abdulaziz University

https://doi.org/10.1038/s41612-024-00648-8

Assessing observational constraints on future European climate in an out-of-sample framework

Christopher H. O'Reilly[1,2] ✉, Lukas Brunner[3,4], Saïd Qasmi[5], Rita Nogherotto[6,7], Andrew P. Ballinger[8], Ben Booth[9], Daniel J. Befort[2,10], Reto Knutti[4], Andrew P. Schurer[8], Aurélien Ribes[5], Antje Weisheimer[11], Erika Coppola[6] & Carol McSweeney[9]

Observations are increasingly used to constrain multi-model projections for future climate assessments. This study assesses the performance of five constraining methods, which have previously been applied to attempt to improve regional climate projections from CMIP5-era models. We employ an out-of-sample testing approach to assess the efficacy of these constraining methods when applied to "pseudo-observational" datasets to constrain future changes in the European climate. These pseudo-observations are taken from CMIP6 simulations, for which future changes were withheld and used for verification. The constrained projections are more accurate and broadly more reliable for regional temperature projections compared to the unconstrained projections, especially in the summer season, which was not clear prior to this study. However, the constraining methods do not improve regional precipitation projections. We also analysed the performance of multi-method projections by combining the constrained projections, which are found to be competitive with the best-performing individual methods and demonstrate improvements in reliability for some temperature projections. The performance of the multi-method projection highlights the potential of combining constraints for the development of constraining methods.

Projections of future climate are important for policymakers and stakeholders to make informed decision for climate-related policy and adaptation strategies (e.g. ref. 1). Of particular value are reliable climate projections on regional scales, however, projections on these spatial scales are often highly uncertain[2,3]. For climate model projections over the next 30–50 years, the uncertainty stemming from different models and internal climate variability on regional scales is comparable to the uncertainty from different forcing scenarios, particularly outside the tropics[4].

Attempting to reliably narrow the uncertainty and provide more accurate projections has been a significant focus for the climate science community. A high-profile example of this is the recent IPCC Sixth Assessment Report, in which observational constraints were used to modify the raw climate model projections in order to provide the best estimates of future climate[5]. One constraining approach that has been employed, in various forms, is the application of different weights to individual model realisations that make up the climate model projections. These weights have been determined by assessing the model independence[6–8] as well as by determining how well the model simulations perform relative to observational data[9–14]. Other approaches involve estimating changes from external forcing by scaling climate model responses to optimally match the changes seen in the observational record[15–17]. In a recent study, part of the "European Climate Prediction system (EUCP)" project[18], six different constraining methods developed by different research groups were all applied to observational datasets to constrain climate projections of different European regions using standardised baselines, regions and projection periods and the results compared[19]. Whilst the different constraining methods demonstrate

[1]Department of Meteorology, University of Reading, Reading, UK. [2]Department of Physics, University of Oxford, Oxford, UK. [3]Institute of Meteorology and Geophysics, University of Vienna, Vienna, Austria. [4]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland. [5]CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France. [6]The Abdus Salam International Centre for Theoretical Physics, Trieste, Italy. [7]Istituto Nazionale di Oceanografia e di Geofisica Sperimentale (OGS), Trieste, Italy. [8]School of GeoSciences, University of Edinburgh, Edinburgh, UK. [9]Met Office Hadley Centre, Exeter, UK. [10]European Centre for Medium-Range Weather Forecasts (ECMWF), Bonn, Germany. [11]NCAS, Department of Physics, University of Oxford, Oxford, UK. ✉e-mail: c.h.oreilly@reading.ac.uk

some similarities in their constrained projections, particularly for median changes, there are also some distinct differences. In particular, the different methods produce substantial discrepancies in the uncertainty ranges of predicted future changes, which clearly presents significant problems for the interpretation and use of the resulting climate projections.

In addition to presenting issues for their interpretation, the uncertainty range of the different constrained projections raises fundamental scientific questions: do these constraining methods produce more accurate and reliable climate projections? And if so, which methods provide the most accurate projections? To address these questions, we designed an out-of-sample "imperfect model" experiment. Out-of-sample testing has previously been identified as a powerful tool for testing emergent observational constraints on future climate change (e.g. ref. 20) and has previously been applied in studies assessing constraints on global climate sensitivity[21–23] as well as some specific aspects of large-scale atmospheric circulation and North American hydroclimate[24]. Here, we test five different methods of constraining future regional climate over Europe by applying them to the same out-of-sample 'pseudo-observational' datasets within a blind testing framework; this is described in more detail in the following section.

Results

Out-of-sample testing using pseudo-observations

In this study, we took advantage of the recently available CMIP6 archive of coupled climate simulations to act as 'pseudo-observations' to test the different methods of constraining regional climate projections, which use CMIP5-era model data (see Methods for full details); for this reason, we refer to this as an out-of-sample test. The pseudo-observational datasets were produced by regridding the required variables from 125 different CMIP6 model simulations (taken from the historical/SSP5-8.5 simulations), anonymising the data and restricting them to a common reference period 1860–2014. These pseudo-observational datasets were then uploaded to the Zenodo online repository[25], and five different groups used these pseudo-observations (in place of real observations) to constrain CMIP5-era historical/RCP8.5 projections. In the present study, each method was applied independently to the pseudo-observations in an attempt to individually provide to best constraint on the future projections for three European climate regions. The utility of this approach is that we can then determine how the constrained projection compares with the actual future change in the pseudo-observation realisation. Details of the five different methods are described in Methods and Supplementary Information; the methods are referred to as Methods A–E in the analysis presented here.

This analysis is similar to a perfect model study but, in some senses, represents a more difficult test of the different constraining methods. The pseudo-observations are drawn from CMIP6, an ensemble that includes a better representation of key processes (such as super-cooled cloud droplets and a wider representation of aerosol-cloud interactions, for example, ref. 26) and is also subjected to slightly different external historical and scenario forcing than the CMIP5-era ensembles that the methods use as the basis for their projections[27]. In addition, the pseudo-observations include those from a number of CMIP6 simulations which fall outside (on the warm end) of the CMIP5 ranges[28]. These factors lead to a tougher test of the methods than a perfect model approach usually implies, and is more along the lines of an 'imperfect model' test (e.g. ref. 29). The assessment was done in this way because it goes part of the way to replicating some of the differences between the real world and the necessarily simplified representations that we use in climate projections. At the same time, several of the 'imperfect models' from the CMIP6 ensemble are direct successors of their CMIP5 ancestors and are therefore not entirely independent[13], as was shown to be the case for CMIP3 and CMIP5 generation models[6].

An example of the application of one of the constraining methods (Method D) is shown for projections of summer (JJA) temperature over the Northern European region applied to pseudo-observational dataset #50 is shown in Fig. 1. In this cherry-picked example, the information from the pseudo-observational data over the reference period (i.e. 1860–2014) results in a constrained projection that has a larger warming signal in the future. This difference is clear for the distributions of projected mean temperature changes over the 2041–2060 verification period (Fig. 1b). As the pseudo-observational dataset is from an out-of-sample CMIP6 model integration, data also exists for the time period from 2041 to 2060. This data can now be used to verify to what extent the constrained projection is more or less accurate compared to the unconstrained projection. In this case, the constrained projection displays a higher probability of future change happening and a smaller ensemble mean error than the unconstrained projection.

The process outlined in Fig. 1 was repeated for each of the five methods and applied to the 125 different pseudo-observational datasets. The systematic approach with which we have applied the individual methods enabled us to explore constrained projections by combining output from the five methods to produce multi-method projections. These multi-method projections are a linear combination of the probabilistic projections from all five methods, sampling equally from each of the five individual constrained/unconstrained projections to create constrained/unconstrained multi-

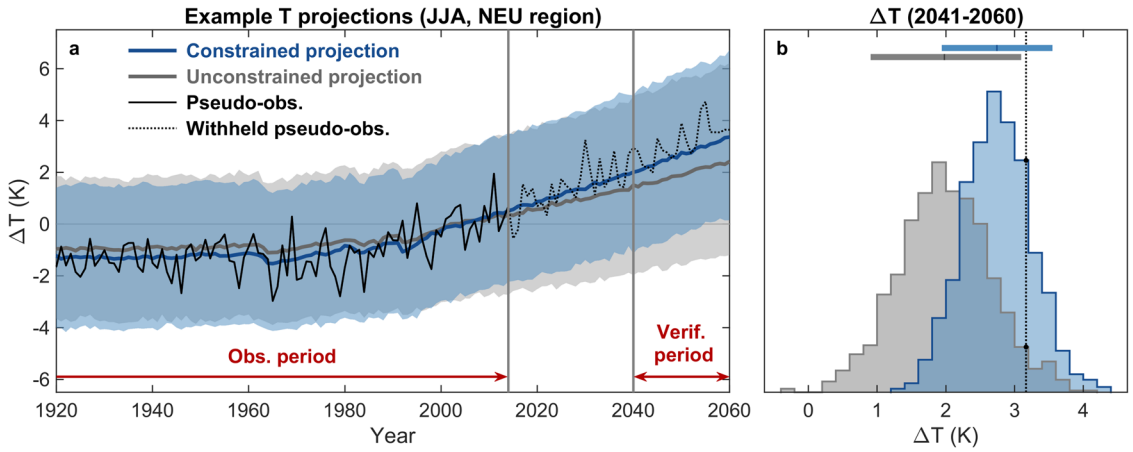

Fig. 1 | An example of one constrained surface-air temperature projection for the Northern Europe (NEU) region (following the RCP8.5 scenario), by Method D from pseudo-observational dataset #50 (c.f. Fig. 2). a The median (thick lines) and 90% range of the unconstrained and constrained projections ensembles are shown along with the pseudo-observational data, all with respect to the 1995–2014 mean. The observational period upto 2014 is shown, along with the future period of which the 2041–2060 change (relative to 1995–2014) is used to verify the projections. **b** The distribution of predicted changes from the unconstrained and constrained projections for Method D applied to pseudo-observational dataset #50; the filled circles and dotted line indicate the actual temperature change in the withheld pseudo-observational dataset. The horizontal lines in (**b**) show the 5–95% range of the constrained and unconstrained projected changes.

Fig. 2 | Projected summer temperature changes for the Northern Europe (NEU) region for each of the Methods A–E, along with the multi-method projections, for the 125 pseudo-observational datasets analysed in this study. The changes are for the mean 2041–2060 temperature relative to the 1995–2014 mean. The box and whiskers show the 5th, 25th, 75th and 95th percentiles of the projected changes, and the short black line shows the median projected change. Light colours show the unconstrained projection, and dark colours show the constrained projections. The horizontal black line spanning each of the panels indicates the actual change calculated from the withheld data from each pseudo-observational dataset, which was used for the verification of the projections. For display purposes, the pseudo-observations have been reordered in terms of future temperature change.

method projections (see Methods for further details). Practically, this exact approach might be unlikely to be directly adopted by others; however, it does appear to provide useful context, as we will go on to show, on where there may be added value in constraint information beyond that captured by the individual methods.

The constraints were performed for temperature and precipitation in the boreal summer (JJA) and boreal winter (DJF) seasons, for the three European "SREX" regions: Northern Europe (NEU), Central Europe (CEU) and Mediterranean (MED)[1]. An example of projected changes for all the pseudo-observational datasets from the different methods is shown for the Northern Europe region in Fig. 2. Whilst this plot is somewhat overwhelming in detail, it is included here to demonstrate the type of data that has been produced in this study. For each of the 125 pseudo-observational datasets, each of the methods have provided a different probabilistic constrained projection for the 2041–2060 climatology with respect to the 1995–2014 baseline period. The range of the projections is shown by the box-whisker plots, and for each method, the unconstrained projection is shown in lighter colours and the constrained projection is shown in darker colours. Also shown are the future changes over the verification period from the respective pseudo-observational dataset (black horizontal lines). It is important to note, as is evident from Fig. 2, that the different methods have quite different unconstrained projections, owing to the different underlying CMIP5-era models that are used by the different constraining methods (see Table 1 in Methods). Most notable, though, is the striking diversity across the constrained projections produced by the different methods when applied to the different pseudo-observational datasets. Some of the methods seem to show larger amplitude changes, such as Method B here, whereas for some other methods, the changes are generally more modest, such as for Method A.

Verification of out-of-sample constrained projections

We begin by analysing temperature projections, as the constraining methods were previously found to substantially influence temperature projections when applied to observations[19]. In the analysis that follows we consider several verification metrics, in particular: root-mean-square error (RMSE), Spread/Error, and continuous ranked probability score (CRPS; see Methods for further details of these metrics). Rank histograms of the projections are also provided in the Supplementary Information for further context (Supplementary Figs. 1–4).

Results for the summer surface-air temperature projections for each of the European regions are shown in Fig. 3. Across all regions there is a general reduction in the RMSE of the constrained projections compared with the unconstrained projections across the different methods (Fig. 3a–c). Importantly, even in the instances when the methods are not improving the projections, they do not substantially degrade the performance (i.e. Methods A and C in the CEU region, Fig. 3b); the only exception is Method B in the Mediterranean region (Fig. 3c). Nonetheless, the general reduction of the error in the constrained projections is an important result because these methods have been applied to the pseudo-observations in a blind setting, indicating that these methods are able to provide robust improvements.

In addition to the individual methods, we have also tested the combined impact of the constraints in terms of a multi-method projection (see Methods). The aim of analysing the multi-method projection is to gain insight into the independent sources of skill in the constraints applied by different methods, as well as examining any potential benefits of combining different methods when applying similar constraining methods to climate projections in other geographical regions. One way of broadly assessing the methods and comparing the multi-method projections is to compare the performance across all regions; to do so we rank the accuracy of the different methods and the multi-method in each region and also calculated an average rank (Supplementary Table 2). Considering the average performance across the regions in this way can provide some insight into the typical performance of the different methods and what might be anticipated when applying these methods to other geographical regions.

The general reduction in RMSE in the constrained summer temperature projections is also clear in the multi-method projection for all three European regions (i.e. Fig. 3a–c). The multi-method projection performs very similarly to the best individual constraining method in the Northern Europe and Central Europe regions and outperforms the best individual constraining method in the Mediterranean region, despite the poor performance of one of the constituent methods for the Mediterranean. The multi-method projection has the highest average rank across the three regions (where higher ranks imply more accurate projections), followed by Method D then Method B. Therefore, the multi-method projections seem to perform well across the three European regions for RMSE—indicating that, at least in terms of RMSE, the multi-method projections are are not strongly susceptible to one individual poorly performing constraining method.

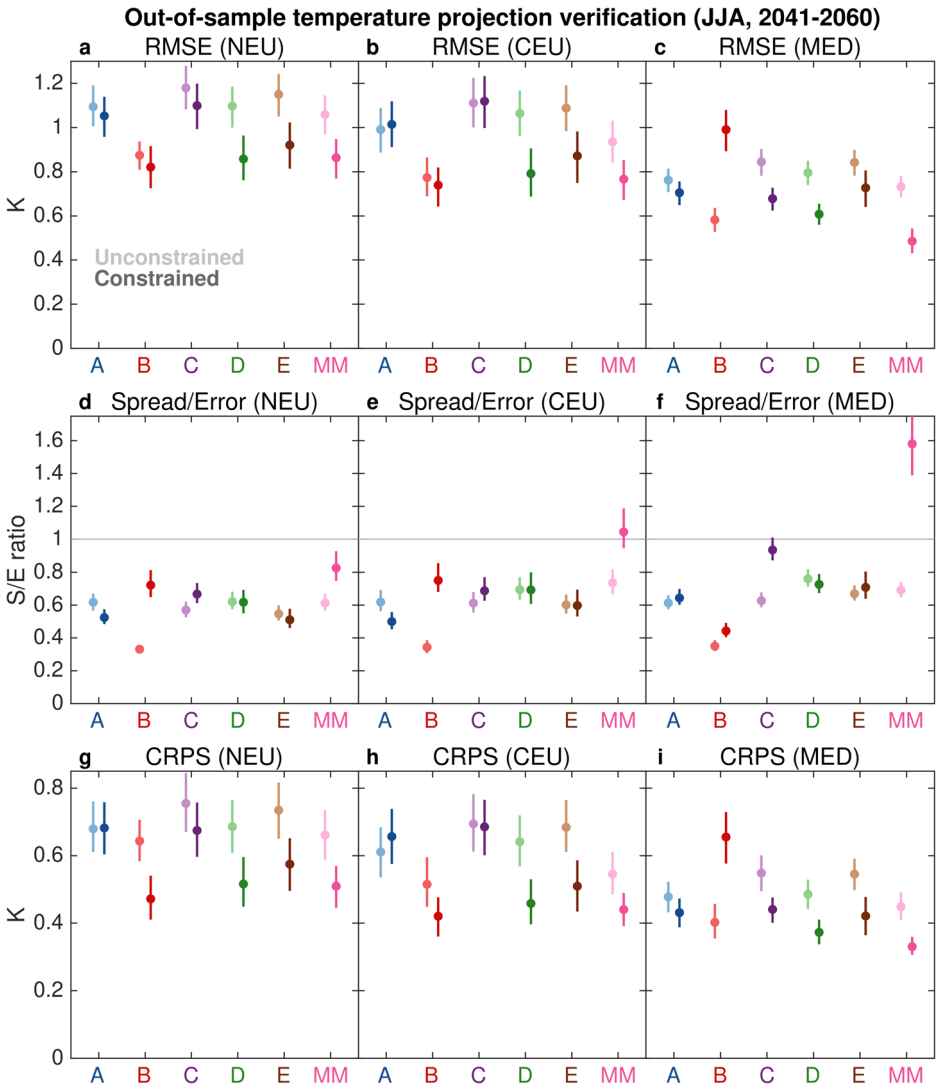

Fig. 3 | Verification of the unconstrained and constrained projections across all 125 pseudo-observations for the 2041–2060 projected summer (JJA) temperature changes in each of the European SREX regions. a–c Root-mean square error (RMSE); a lower value implies a more accurate ensemble mean projection. **d**, **e** Spread/Error ratio; values less than one mean the projections are overconfident and values greater than one mean the projections are underconfident. **g–i** Continuous ranked probability score (CRPS); a lower value demonstrates a more accurate probabilistic projection. The dots show the measured values and the lines indicate the 95% confidence intervals based on a bootstrap resampling (see Methods).

In terms of Spread/Error ratio, the unconstrained and constrained projections are all overconfident for European summer temperatures (Fig. 3d–f). It should be noted that substantial overconfidence cannot really occur in simpler "leave-one-out" testing and exemplifies the more difficult test provided by the out-of-sample approach used here. There are some clear improvements (i.e. Spread/Error ratio closer to 1) in the reliability of the projections for some of the individual methods, especially Method B in the Northern Europe and Central Europe regions and Method C in the Mediterranean region, however, the reliability worsens for Method A in the Northern Europe and Central Europe regions. Overall, the multi-method projection seems to perform well, being more reliable than any of the individual methods in the Northern Europe and Central Europe regions in terms of the Spread/Error ratio. A major exception, however, is the underconfidence of the constrained multi-method projection in the Mediterranean region (i.e. Spread/Error ratio $\gg 1$); this seems to be related to the poor performance of the Method B constraint in this region, which doesn't greatly affect the mean error of the constrained multi-method projections (i.e. Fig. 3c) but hugely inflates the spread of the multi-method projection, resulting in substantial underconfidence in the projection in terms of Spread/Error.

The overall performance of the full probabilistic projections can be assessed with the continuous ranked probability score (CRPS), shown in Fig. 3g–i. The constraining methods generally improve the accuracy of the summer temperature projections in terms of CRPS, again with the notable exception of Method B in the Mediterranean. Some of these improvements, however, seem to originate from different sources. For Method D and E, the improvements seem to be almost entirely associated with the reduction in the RMSE (i.e. Fig. 3a–c), whereas the improvements for Method B in Northern and Central Europe (Fig. 3g, h) also have a contribution from changes in the spread of the projection (i.e. Fig. 3d, e). To compare the overall performance of the different methods and the multi-method projections across all regions examined here, we again rank the accuracy in each region and calculated an average rank (see Supplementary Table 2). The multi-method projection has the highest average rank across the three regions, which suggests that, based on this out-of-sample test, that there may be more information available to constrain projections than is currently captured in any of the individual methodologies.

The CRPS of the summer precipitation projections is shown in Fig. 4a–c (see Supplementary Fig. 5 for other metrics). In contrast to the summer temperature projections, the constraining methods do not demonstrate any substantive improvements in predicting the future changes. Most of the methods actually have very similar verification measures, indicating that the constraining methods are having little impact on the accuracy of the projected changes for summer precipitation. The exception here is Method E, which exhibits a substantial degradation in performance in the Central Europe and Mediterranean regions (Fig. 4b, c), which is associated with a significant increase in the projection RMSE (Supplementary Fig. 5). The multi-method also shows no substantive improvements in RMSE, similar to the individual methods.

The verification for projections of European winter are in shown in terms of CRPS in Fig. 4d–i (see also Supplementary Figs. 6, 7). For winter temperature projections, the majority of the constraining methods show improvements for the Mediterranean region (Fig. 4f), which are broadly associated with both a decrease in the projection RMSE and an improvement in ensemble spread, with the constraining methods producing less overconfident projections (Supplementary Fig. 6). The exception is Method A, for which the constrained projections are worse than the unconstrained projection in all regions. Methods B–E show more modest improvements in general for the Northern Europe and Central Europe regions than they do for the Mediterranean (Fig. 4d, e); the most substantial instance of a method degrading the accuracy of the projections is for Method E in Northern Europe. The multi-method constrained projections show a consistent improvement (albeit more modest than in summer); ranking the accuracy of the constrained projections we find that, in terms of average rank across all three European regions (i.e. Supplementary Table 2), the multi-method constrained projections perform better than any individual method for winter temperature projections. It should be noted, however, that the multi-method is not more accurate than all individual methods in all regions, which has implications for potential users of these climate projections on regional scales (further implications are outlined in the Discussion section below).

For the winter precipitation projections, there is no substantial improvement for the constrained projections (Fig. 4g–i), similar to what we found for the summer precipitation projections. However, unlike in the summer precipitation projections, there are many instances where the constraining methods noticeably degrade the accuracy of the projections. This is particularly the case in the Northern Europe region, where even the multi-method constrained projection exhibits inferior performance to the equivalent unconstrained projection.

A noteworthy feature of the CRPS verification is the different accuracy of the unconstrained projections. This difference stems from differences in the underlying models/simulations used by each of the methods (see Methods and the example in Fig. 2). From the CRPS verification (i.e. Figs. 3, 4), higher accuracy in the constrained projections tends to be associated with higher accuracy in the underlying unconstrained projections across the different methods. To account for this, we also calculated a continuous ranked probability skill score (CRPSS; see Methods), which is defined to measure the relative improvement of the constrained projections compared to the unconstrained projections (Supplementary Fig. 8). In terms of CRPSS, substantial relative improvements are shown by the majority of the constraining methods for summer temperatures: the relative improvements of Methods B, D and E are comparable or better than the multi-method projections in the Northern Europe and Central Europe regions and Methods C, D and E show improvements comparable with the multi-method projections in the Mediterranean region. There are also clear relative improvements for several methods for winter temperature projections that are comparable to the relative improvements seen in the multi-method projections (Supplementary Fig. 8g–i). Whilst some methods demonstrate comparable relative improvements to the multi-method projections of future regional temperatures, none perform significantly better in any of the regions for either summer or winter. The improvements measured in terms of CRPSS are anti-correlated with the accuracy of the unconstrained ensemble (i.e. $CRPS_{unconst}$; see Methods and Supplementary Fig. 9), such that larger improvements are found when the unconstrained ensemble is less accurate. Therefore, the relative improvement of the constrained multi-method projections is particularly notable, as the accuracy of the unconstrained multi-method projections is relatively high compared to the individual methods (i.e. Figs. 3g–i, 4d–f).

Overall, the out-of-sample verification reveals that, whilst not universal, the constrained projections using the different methods do tend to improve the projections for European regional temperatures. The constraints exhibit the most substantial impact for summer temperature projections, though some improvements are also demonstrated for the winter temperature projections. For precipitation, the constraints do generally not improve the projections and for winter precipitation the accuracy of the projections is even degraded in many cases.

Understanding the performance of the constrained projections

To investigate sources of skill in the different methods for the different combinations of variable, season and region, we examined the correlation between the constrained projections and the future changes in the pseudo-observations (Supplementary Fig. 10). Each correlation was calculated between 125 pairs of values for the constrained ensemble mean projected change and the actual change in the model realisation used for the pseudo-observations. For summer temperatures—and also winter temperatures for most methods—the constraining methods have projections that are positively correlated with the future changes in the pseudo-observations, demonstrating that the methods are able to capture some of the variation in future changes across the different pseudo-observations. For precipitation, the correlations are generally lower and even negative for many methods/

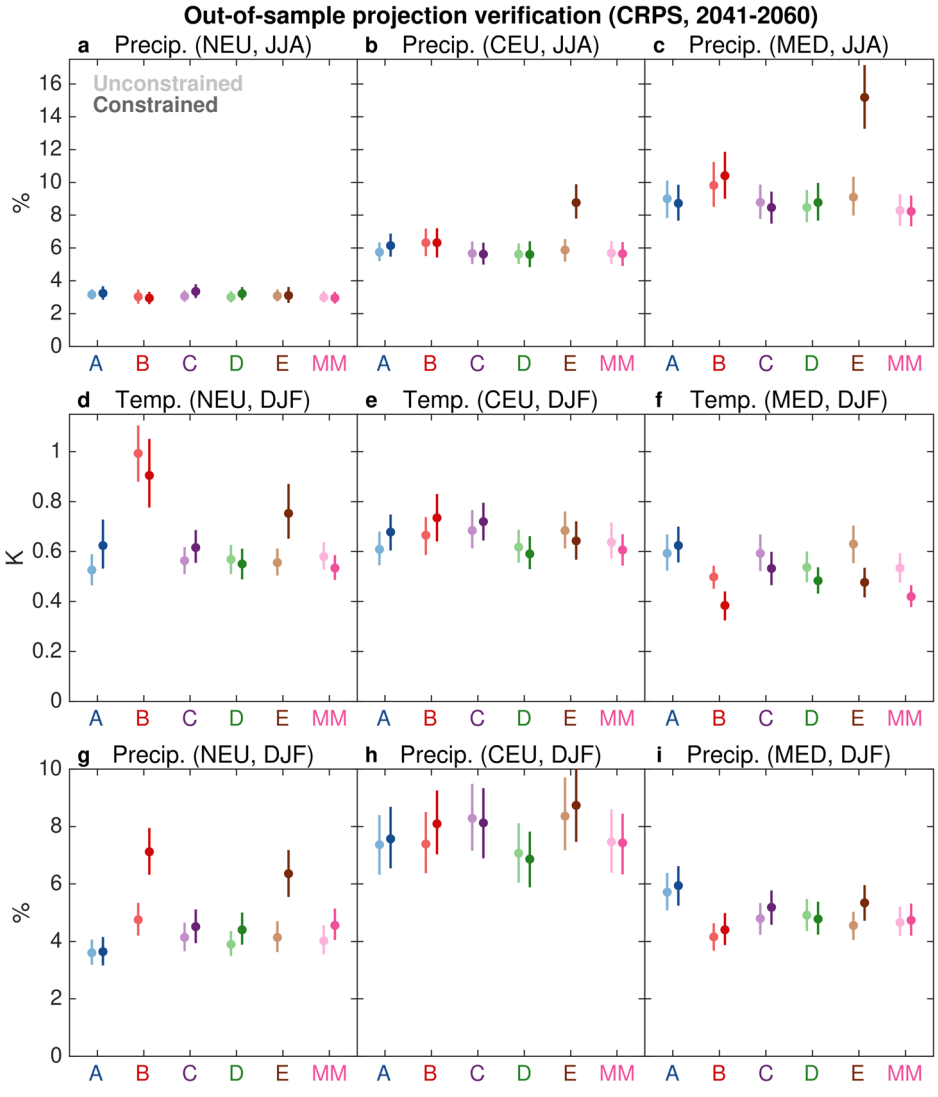

Fig. 4 | Verification of the unconstrained and constrained projections across all 125 pseudo-observations, in terms of continuous ranked probability score (CRPS), for the 2041–2060 projected changes in each of the European SREX regions. **a**–**c** Summer precipitation changes; **d**, **e** winter temperature changes; **g**–**i** winter precipitation changes. The dots show the measured values and the lines indicate the 95% confidence intervals based on a bootstrap resampling (see Methods).

regions, consistent with the lack of improvement found for the constrained projections (i.e. Fig. 4). The ensemble means of the constrained projections are generally positively correlated to one another for summer and winter temperatures (shown in Supplementary Fig. 11). However, the correlation coefficients are less than $r \approx 0.7$ in most cases, indicating that there is substantial independence between the different methods (as also evident in Fig. 2). In addition we examined how this is linked to global changes by calculating correlation between the ensemble mean of the constrained projections and the global mean surface temperature (GMST) change in the pseudo-observations (Supplementary Fig. 12). The correlations show that the constrained changes are clearly related to the change in GMST in some instances. For the summer temperatures, there is the strongest link, but interestingly the constrained changes are not so strong for the other variables/seasons, with the overall behaviour being very similar to the correlations between the projected and actual regional changes (i.e. Supplementary Fig. 10). This indicates that the methods are often picking up patterns of regional changes that are closely linked to global changes, albeit implicitly for the methods that do not include global data (see Methods).

The independence of the constraining methods and their positive correlations with the future change in the pseudo-observations (e.g. Supplementary Fig. 10) gives a clue as to why the multi-method projections perform relatively well: the different methods capture different signals over the common reference period (i.e. 1860–2014) from which they provide useful constraints, which, when combined provide a projection competitive with the best individual method. It is notable that the multi-method projections of regional summer temperatures not only have relatively high correlations with the future change in the pseudo-observations—at least as high as all the individual methods across the three European regions—but the multi-method projections also produce the most reliable probabilistic projections in terms of Spread/Error ratio for two of the three European regions. An exception being the Mediterranean region, where one poorly performing method severely degrades the Spread/Error (i.e. Fig. 3d–f). Combining multiple predictions has also been found to produce improved reliability in the analysis of seasonal forecasts, in which combining multiple models generally performs better than any individual model[30,31]. The improvement in reliability for the multi-method summer temperature projections in the Northern and Central European regions likely occurs because the individual constraining methods are overconfident and lacking in spread (e.g. Fig. 3), therefore, combining the projections inflates the spread and the reliability of the multi-method projection is improved.

The impact of the multi-method approach is clear when examining the ability of the constrained projections to capture the more unlikely outcomes in the pseudo-observations. We calculated how often the actual changes in the pseudo-observations fell below the first percentile and above the 99th percentile of the different constrained projections (Supplementary Figs. 13–16). The methods do demonstrate some ability to improve the projections of these outlier cases, however, there are still many more cases when the change in the pseudo-observations falls outside the 1–99% range for all the methods. This is

especially the case for temperature outcomes above the 99th percentile of the projections, likely due to the higher climate sensitivity in the CMIP6 models used as the pseudo-observations (i.e. Supplementary Fig. 12). For the multi-method projections; however, changes outside the 1–99% range are all reduced to reliable levels in almost all cases (i.e. $\mathcal{O}(1\%)$). For the summer temperatures above the 99th percentile of the projections, the improvement stems from the constraints provided by Methods B and E. These methods differ from the other methods in that they scale the projected signals (rather than by weighting individual members; see Methods), demonstrating one utility of the scaling methods. Overall, the combination of the different overconfident projections results in a multi-method ensemble projection that is remarkably reliable when it comes to projecting the more unlikely outcomes, which is even evident by visual inspection of the raw constrained multi-method projections (i.e. Fig. 2).

The CMIP6 models used as pseudo-observations in these out-of-sample tests have been shown to generally exhibit higher levels of climate sensitivity to CO2 increases than the CMIP5-era models used for the constrained projections here[32,33] and project higher levels of warming over Europe during the 21st century[28]. To examine how the results depend on the equilibrium climate sensitivity of the models used as pseudo-observations, we split the 125 pseudo-observations into subsets that were within the CMIP5 range ($n = 38$) and above the CMIP5 range ($n = 87$), using model-specific data equilibrium climate sensitivity values from[32]. The distributions of the changes in the pseudo-observations and the subsets of different climate sensitivities are shown in Supplementary Figs. 17, 18, and the verification statistics for the high and low-sensitivity subsets are shown in Supplementary Figs. 19–24.

The accuracy of the projections is generally better (i.e. lower RMSE) for the lower sensitivity subset compared to the higher sensitivity subset, but the qualitative improvement seen across all models (i.e. Figs. 3, 4) is evident in both subsets (Supplementary Figs. 19, 20). The relative improvement (in terms of CRPSS), however, is much larger in the higher climate sensitivity subset and is particularly clear for the temperature projections (Supplementary Figs. 21, 22). This indicates that the methods are able to detect these stronger signals over the common reference period (1860–2014), and the constraints have a larger impact on these high-sensitivity models. In contrast, for the lower sensitivity models, the pseudo-observations are already much closer to the uncalibrated ensembles—somewhat by definition since they are models that have similar characteristics—and as a result, there are lower errors in the calibrated projections.

The overconfidence seen in the temperature projections (i.e. Fig. 3d–f and Supplementary Fig. 6) is primarily a feature of the high-sensitivity models (Supplementary Figs. 23, 24), which are substantially different from the CMIP5 models that make up the unconstrained projections in the majority of the projections analysed here. This is perhaps unsurprising given that the methods which use a weighting approach (see Table 1, Methods) rely on an underlying ensemble that does not have as many warm outcomes as in the high climate sensitivity models (i.e. Supplementary Figs. 17, 18). The differences in the responses to the forcing scenarios used in the unconstrained projections (i.e. historical/RCP8.5 from CMIP5) compared to the pseudo-observations (i.e. historical/RCP8.5 from CMIP6) likely generate the different responses and contribute to the apparent overconfidence[34]. These differences in the underlying projections/pseudo-observations indicate that this out-of-sample verification is a stiff test of the constraining methods. However, a large majority of the pseudo-observational (CMIP6) future changes are within the range of the CMIP5 future changes (Supplementary Figs. 17, 18, 25). It would therefore not be impossible for constraining methods based on CMIP5 weighting to capture the future changes, however, the signals in the higher sensitivity models over the observational period clearly have features that are not well captured in the underlying CMIP5 models and/or are not accounted for in the methods examined.

Observationally constrained projections of European climate

Our analysis to this point shows that the constraining methods provide useful changes to future projections when applied to pseudo-observations, particularly for summer temperatures over Europe. Therefore, assuming that similar improvements may be possible when applied to the real-world climate system, we have some justification for applying the constraints to the real observations for European summer temperatures, as in ref. 19.

Constrained projections of the summer temperature change for 2041–2060 for each of the European regions are shown in Fig. 5. Based on our prior analysis, we have reason to expect the multi-method projection that has been constrained with observations to be at least as accurate as any of the individual methods. The observationally constrained multi-method projection predicts a median warming of less than 2K for the Northern Europe region, but there is substantial uncertainty, with changes of less than 1K and upto 3K within the 5–95% range of outcomes. For the Central Europe and Mediterranean regions, the observationally constrained multi-method projections predict a median warming of over 2K and interestingly, the projections have a markedly narrower spread compared with the respective unconstrained projections for Northern Europe. Whilst the difference between the constrained and unconstrained projections when the methods are applied to real observations is fairly modest, it is important to note that the out-of-sample tests reveal that the difference produced by the constraining methods could potentially have been much larger (e.g. Fig. 2). That the influence of the constraining methods is more muted reveals that the signals in the observational data are more consistent with the unconstrained (CMIP5-era) ensembles than many of the pseudo-observational (i.e. CMIP6) datasets.

Discussion

In this study, we have demonstrated an assessment of different methods of constraining future regional climate projections, using an out-of-sample testing framework. The target of this study was climate projections for European regions but the methods themselves and the results presented here are likely to be relevant for other regional climate projections. Overall, the constraining methods demonstrate some clear improvements over the unconstrained projections of European temperatures when applied to 125 pseudo-observational datasets. The summer season demonstrates the greatest improvements, but there are also improvements to the winter temperature projections. For precipitation projections, however, there is little evidence that any of the constraining methods provide any substantial and consistent improvement. Constrained regional precipitation projections in the extratropics should, therefore, be treated with caution and could be more likely to have a lower projection accuracy than unconstrained projections. None of this had been clearly demonstrated prior to this study, therefore these are important results for the use and development of observational constraints to future regional climate projections.

The superior performance of the constraining methods when applied to regional temperature compared with regional precipitation, particularly for summer temperatures, is something that might have been anticipated but was not clear prior to this out-of-sample study. The reason for the efficacy of the methods when applied to temperature projections compared with precipitation projections is likely linked in part to the signal-to-noise ratio of the climate change signal in the observational period. For summer temperatures, the signal-to-noise ratios are relatively large over Europe, and by some estimates, the climate change signal has 'emerged' from the noise over the observational period (e.g. refs. 35,36). For winter temperatures, the larger internal variability (that is itself likely underestimated in the CMIP6 simulations that are used here as pseudo-observations[37]) reduces the signal-to-noise ratio over the observational period, though there are evidently still substantial signals to constrain upon and the constraining methods do broadly provide an improvement to the projections, particularly the multi-method projection (i.e. Fig. 4d–f and Supplementary Table 2). There are generally much lower signal-to-noise ratios for regional precipitation over the observational period and the methods struggle to identify behaviour on which to accurately constrain future projections. This particularly interesting because in observations there is a drying trend[38] that we might expect to provide a useful constraint, however the methods tested here do not demonstrate they are able to effectively capture this apparent

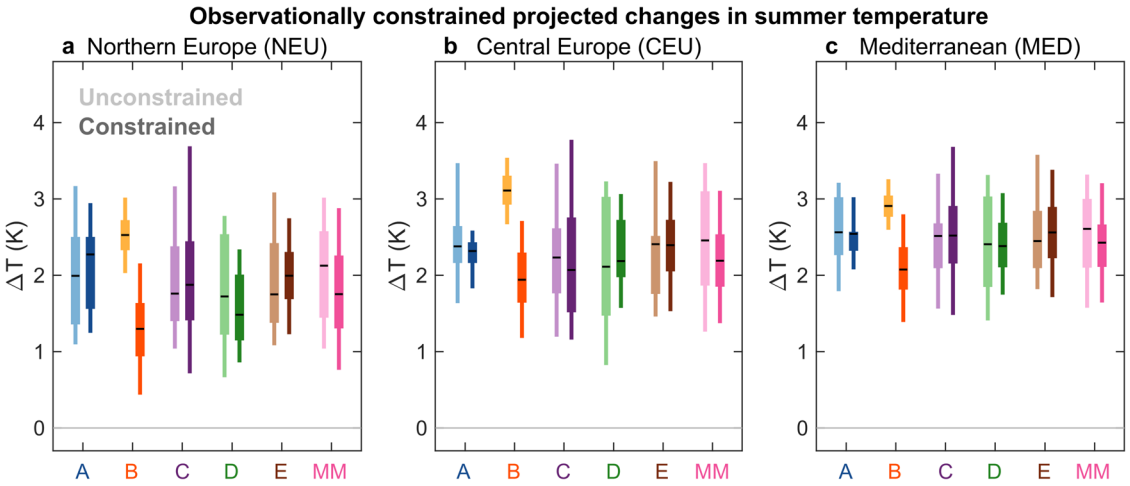

Fig. 5 | Projected summer temperature changes for each method constrained using real observational products, along with the multi-method projections. The changes are for the mean 2041–2060 temperature relative to the 1995–2014 mean. The box and whiskers show the 5th, 25th, 75th, and 95th percentiles of the projected changes, and the short black line shows the median projected change for **a** Northern Europe, **b** Central Europe and **c** Mediterranean regions.

signal; this warrants investigation in future studies. It is worth noting that there may be benefits in future studies applying the methods over the earlier historical period to predict the later part of the observational record, albeit with some limitations due to the small sample size, and that models use observations for tuning during their development. The results from the constraining methods examined in this study provide a useful benchmark for future developments in the constraining methodologies.

Our results do not mean that constraining precipitation projections is not possible, of course, and improved methods and models applied to a different observational period may lead to more useful constraints for future projections. A reason why the methods might struggle to constrain precipitation is that none of the methods directly constrain large-scale circulation changes. In the extratropics, and Europe in particular, changes in large-scale circulation strongly control precipitation variability and changes on decadal timescales (e.g. refs. 39,40). For example, future Mediterranean drying has been attributed to forced changes in large-scale circulation in several modelling studies[41–43]. However, none of the constraining methods tested in this study are targeted at directly constraining circulation changes, and some omit dynamical variables entirely (see Methods). Therefore, it is clear that any precipitation changes that are driven by circulation changes would not be well detected by the methods used in this study. Exploring possible observational constraints on such dynamical changes is an important topic of future research and the out-of-sample framework in this study could be used in future studies to assess if other constraining methods can provide robust improvements.

Whilst this study allows us to compare how the different methods perform relative to one another, an important result is that, in terms of average performance evaluated across all three European regions, the constrained multi-method temperature projections are broadly competitive with any individual method (i.e. Figs. 3, 4d–f, Supplementary Fig. 6 and Supplementary Table 2). The multi-method projections seem to produce relatively accurate and reliable projections for two main reasons. Firstly, the projections identify different aspects of the behaviour over the common reference period, and these combine to provide more accurate projections, whilst compensating errors largely cancel out. Secondly, most of the individual projections are overconfident and combining the different ensembles can increase the spread of the projection and reduce the reliance on any single overconfident projection; this is perhaps most notable when assessing the more unlikely outcomes, which rarely fall outside the 1–99% range of the multi-method projections. These improvements in reliability in the constrained multi-method projections are analogous to the improvements seen in multi-model initialised forecasts on shorter timescales (e.g. refs. 30,31). The use of all the methods analysed here to produce a multi-method projection is not straightforward for individual climate applications, however, approaches utilising several constraining methods could be used for operational purposes without too much difficulty. Moreover, the relative success of the multi-method approach, in terms of accuracy and reliability for many of the constrained temperature projections analysed here, clearly demonstrates that improvements to the individual constraining methods are achievable.

In this comparison of different regional projection methodologies, we made use of newer generation climate projections which provided a tougher test of constraints based on earlier generation simulations (both because the out-of-sample simulations included future changes outside the earlier ranges and because these introduced historical forcing differences that are likely to be akin to constraining simulations based on real-world data). However, some caution is needed in applying these methods to the real world, as we are not able to control for potential common model biases in both the constrained projections and the CMIP6 simulations used for the out-of-sample tests. For example, almost all current simulations fail to capture the lack of recent warming in the east Pacific[44], which may be linked to the suppression of stronger climate feedbacks that could re-emerge in future climate change (e.g. refs. 45,46). The potential impact of biases such as these are a challenge for all climate projections (constrained or otherwise), and represent a context which one must bear in mind when producing future climate projections that inform decision making. Nevertheless, the assessment of methods to constrain climate projections, taken here, represents a step forward in understanding the potential merit (or otherwise) in an imperfect model evaluation, which exposes the various methods to a more difficult test of their potential skill. Findings, such as where these methods don't add skill and identifying where combining underlying information captured in different methods may add skill, represent an important step forward in how we develop, test and make use of these constraining methods.

The focus of this study has been on examining the impact of these different constraining methods, but a question remains over how these results might influence the application of constraints by practitioners. Given that many CMIP6 models tend to show more warming than previous model generations, the so-called 'hot model' problem[47], there is growing consensus that it is necessary to apply observational constraints to make regional climate projections. The best approach for practitioners will likely depend on their specific application and focus. For example, those interested in summer temperature projections for Central Europe (CEU) might choose to apply one of the best-performing individual methods, B or D perhaps, if they are primarily interested in the general accuracy of the probabilistic projection and the relative improvement of the method (i.e. measures of

RMSE, CRPS and CRPSS). However, these methods individually demonstrate clear overconfidence in terms of their overall distributions (i.e. Spread/Error ratio, Fig. 3e) and in terms of capturing extreme outcomes within the projection (i.e. outside the 1–99% range, Supplementary Fig. 13b, e), therefore a practitioner primarily concerned about such outcomes might opt to combine methods within a multi-method projection to reduce the overconfidence in the projections (e.g. Fig. 3e). Whilst this example demonstrates how some of the choices of methodology could depend on the specific application and focus, our study highlights that for regional temperature projections over Europe there is likely benefit in applying observational constraints. However, we find that there is little benefit in constraining European precipitation projections using these methods. This study can also provide some guidance to practitioners applying observational constraints to other geographical regions. The results here indicate that, in the absence of a specific comprehensive out-of-sample study, where possible combining multiple constraining methods may be a more reasonable approach than selecting an individual method.

There is growing demand from governments of different countries and regions for their own customised climate projection products to aid adaptation and improve resiliency to future climate outcomes, for example, the 'UK Climate Projections 2018' (UKCP18)[48] and 'Swiss Climate Change Scenarios 2018' (CH2018)[49]. Our study demonstrates that such regional constrained projections are likely to add value over using unconstrained climate model output. One takeaway message from our results is that incorporating information from a range of sources—as demonstrated by the performance of the multi-method projections here—is important when developing constraints and offers the potential to further improve individual constraining methods.

Methods

Overview of constraining methods

Five different constraining methods were assessed in the study. Each of the methods was assigned a letter between A–E for the purpose of the analysis. Table 1 shows a summary of the different methods and highlights some important features and assumptions that the methods rely upon. Further details of the specific methods are provided below and a further comparison of some of the methods characteristics is shown in Supplementary Table 1.

It is important to note that the Methods use different underlying ensembles to constrain. Some of these differences are unavoidable consequences of the different methodological approaches. For example, ASK needs single-forcing experiments to identify patterns, that are available for only a small subset of CMIP5. Whilst smaller ensembles are less likely to adversely impact ASK (which scales ensemble mean fingerprints to be consistent with observed changes), it was felt that limiting other methods to this same subset would limit their predictive capabilities. So, whilst some efforts were made within the participating groups to standardise unconstrained projections, evident differences remain. The impact of different underlying ensembles was explored in ref. 19. Further details are provided below, and the list of underlying models is included in the Supplementary Data File.

Method A (REA)

The reliability ensemble averaging (REA[10]) method applies a weighting scheme based on model performance and independence. For the model performance component, REA applies weights on a variable-by-variable basis, expressed in terms of the bias of the model control run with respect to observations (specific models and number of ensemble members are listed in Supplementary Data). Model convergence is treated as an indication of projection confidence, effectively downweighting outliers in projection space[9,50]. The value of the weights increases as the model biases and distances. REA's goal is to minimise the contribution of simulations that either perform poorly in the representation of present-day climate or provide outlier future simulations with respect to the other models in the ensemble in order to reduce the uncertainty and increase the skill of climate projections.

Table 1 | A summary of the different methods used to constrain future projection in this study, including some important features and assumptions that the methods rely upon

Method	Underlying model data	Observational constraints	Treatment of model dependencies	Key Assumptions
(A) REA	CMIP5 MME	Weighted based on the historical performance of the target variable	Weighted based on the distance to the MME mean	Future model performance can be inferred from historical performance on a variable-by-variable basis; the ensemble is truth-centred
(B) CALL	Single model large-ensemble	Ensemble projection distribution is scaled to optimise fit to observations	-	The relationship between the past evolution of the ensemble dataset and the observations contains meaningful information for the future evolution
(C) ClimWIP	CMIP5 MME	Weighted based on the historical performance of seven diagnostics	Weighted based on historical independence from other models	Future model performance can be inferred from historical performance; interdependence of models can be inferred from model outputs
(D) KCC	CMIP5 MME	Constrained based on the historical warming trend	-	Real-world response to forcings is statistically indistinguishable from model responses; response to anthropogenic forcing is smooth over time
(E) ASK	CMIP5 MME	Scaled based on observed time-space change over the historical period	-	Space-time pattern of climate response to forcing is governed by known physics and is correctly represented in models, whereas the amplitude is governed by uncertain feedbacks and is, hence, estimated from observations

The different models and ensemble members underlying each method is provided in Supplementary Data.

Method B (CALL)

The CALL method involves the 'CALibration of Large-ensembles' using available observational data following the approach described in ref. 51 and which was included in the multi-method comparison study of ref. 19. The CALL method uses an approach that considers the observational changes and variability of the target variable time-series (computed over the region of interest) and uses this information to calibrate a single model large ensemble projection, in this case data from the CESM-LENS[52]. The time-series from the large ensemble and the observations are first separated into dynamical and residual components[39,53]; the dynamical and residual time-series are then separately calibrated towards the dynamical and residual observational target time-series using homogenous Gaussian regression (a simplified version of Ensemble Model Output Statistics[54]) in which the calibration finds an optimally scaled ensemble mean anomaly and ensemble variance; finally, the calibrated dynamical and residual ensembles are combined to produce an observationally constrained ensemble projection. This approach corresponds to the 'HGR-decomp' method described and analysed by ref. 51. Since this method relies on a single model large-ensemble, it's unconstrained projections can be quite different from those in the other methods for some regions/variables (e.g. Fig. 2).

Method C (ClimWIP)

ClimWIP assigns weights to models in the ensemble to produce weighted, reliable percentiles. Each weight is designed to quantify a model's performance in a given region and for a given target variable as well as its independence from the other model's in the ensemble based on the model's output[7,8,13]. Earlier versions of ClimWIP have been applied to regional projections of different variables such as sea ice[8], temperature[11,55], ozone[56] and precipitation[57] based on different generations of CMIP as well as large ensembles[8,12]. Recently, this approach has also been applied for the first time on a global scale[19], for the weighting of downstream regional climate projections[58], and to the prediction horizon spanning decadal predictions and climate projections[59]. Here, we use a version of ClimWIP that is largely consistent with[11] and also included in the multi-method comparison presented in ref. 19 (specific models and number of ensemble members are listed in Supplementary Data). The original setup was updated to include temperature trend as a predictor for model performance based on the results from ref. 13 and to use larger scale metrics to establish model dependence as suggested by ref. 12. More details can be found in the corresponding publications, here we limit the further description of ClimWIP to several properties important in the context of this study:

1. As a weighting method, ClimWIP can not shift the distribution outside of its original, unweighted range, limiting the possible reduction of the error for CMIP6 pseudo-observations lying outside the full CMIP5 range.
2. The independence part of the weighting is not mainly intended to optimise performance nor to reduce the uncertainty but to account for structurally similar models (i.e. models with a large overlap in their source code) but it is still included here as it accounts important aspect of multi-model projections and is an integral part of ClimWIP.
3. A perfect model test comparable to the one performed here has already been carried out for the case of global mean temperature change by ref. 13, who found an increase in the CRPS by 10–20%. In their assessment, ref. 13 excludes models that are related by a direct line of development between CMIP5 and CMIP6 in order to maximise the 'out-of-sampleness' of the pseudo-observations. This was not done for this study for simplicity and to stay consistent with the other methods. Sensitivity experiments using the setup from ref. 13 show that not accounting for closely related models might bias the CRPS high by a few percentage points in the median while, in turn, also increasing the risk for negative skill changes, so that the overall effect can be considered small.

Method D (KCC)

The Kriging for Climate Change (KCC) method has been previously applied to global mean warming[14] and regional warming[60,61] and works in three steps. Here, we apply the same version that was used by the refs. 11,55. First, the forced response of each CMIP5 model is estimated over the whole 1860–2100 period (after concatenation of historical simulations with corresponding 21st-century projections). In order to also get attribution statements, the responses to ALL (all forcings) and NAT (natural forcings only) are estimated separately (specific models and number of ensemble members are listed in Supplementary Data). Second, the sample of the forced responses from available climate models is used as a prior of the forced response within each pseudo-observations, assuming that 'models are statistically indistinguishable from the truth'. Third, pseudo-observations are used to derive a posterior distribution of the past and future forced response given the pseudo-observations. This Bayesian method can be summarised using the following equation:

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{\epsilon},$$

where $\boldsymbol{y}$ is the time-series of pseudo-observations (a vector), $\boldsymbol{x}$ is the time-series of the forced response (a vector), $\boldsymbol{H}$ is an observational operator (matrix), $\boldsymbol{\epsilon}$ is the random noise associated with internal variability (a vector), and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_y)$, where $\mathcal{N}$ stands for the multivariate Gaussian distribution. Climate models are used to construct a prior on $\boldsymbol{x} : \Pi(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$. Then the posterior distribution given pseudo-observations $\boldsymbol{y}$ can be derived as $p(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$.

$\boldsymbol{\mu}_x$ and $\boldsymbol{\Sigma}_x$ are estimated as the sample mean and covariance of the forced responses from the CMIP5 models. We use a mix of auto-regressive processes of order 1 (AR1) to model internal variability within $\boldsymbol{\Sigma}_y$. The intrinsic variance is derived from pseudo-observations after subtracting the CMIP5 multi-model mean estimate of the forced response. The full documentation of the method is detailed by ref. 61. Unlike some of the other methods, KCC was originally designed to assess the forced response only. Here, the uncertainty related to internal variability is included to ensure consistency with other approaches. To do this, internal variability is estimated from pseudo-observations, after subtracting the estimated forced response. Then, random drawings of internal variability are added to the projected forced values to get projection ranges, including full uncertainty.

Method E (ASK)

The 'Allen–Stott–Kettleborough' ASK method[15–17] applies a detection and attribution approach to disentangling the impact of external forcing(s) and internal variability on observed trends. The method assumes that the true observed climate response is a simple linear combination of one or more individual forcing fingerprints. Implementations of these methods have been found to produce overconfident constraints when signal-to-noise ratios are low in the observed trends[62]. This particular implementation of ASK largely follows its application in recent studies within a regional European context[19,63]. The model fingerprints are comprised of the annual multi-model mean time-series (1950–2014), spatially averaged over different European regions (NEU, CEU and MED), and conjoined following normalisation by a measure of the region's internal variability. A total-least-squares (TLS) regression is used to estimate the scaling factor(s) required to scale the model fingerprints to match the amplitude of the observed response, and accounts for uncertainty in both the observations and modelled response to each of the forcings due to internal variability. Confidence intervals were estimated by adding equivalent samples from the piControl simulations to both the noise-reduced fingerprints and pseudo-observations, and then recomputing the TLS regression (10,000 times) in order to build a distribution of scaling factors, from which the 5th-95th percentile range is computed.

For each of the 125 pseudo-observational datasets of temperature, a two-signal TLS regression using the all-forcing historical and single-forcing (GHG-only) historical multi-model mean fingerprints (from 25 ensemble members, 9 different models) yielded the GHG scaling factor

(and uncertainty range) used to constrain the future temperature projection (specific models and number of ensemble members are listed in Supplementary Data). This method was used as single forced CMIP5 projections are not available, which is a known problem when using CMIP5 simulations to constrain future projections using this type of detection and attribution approach[64]. Due to the lower signal-to-noise in European regional precipitation trends (compared to temperature), for each of the pseudo-observational datasets of precipitation, a one-signal TLS regression using the all-forcing historical multi-model mean fingerprint (from 81 ensemble members, 34 different models) yielded the all-forcing scaling factor that was applied as a constraint on the future precipitation projection. For each of the pseudo-observations, if the scaling factor resulting from the TLS regression was determined to be significantly negative (including the full 5–95% uncertainty range), suggesting an unphysical scaling, the resulting constraint was rejected and not supplied for further analysis.

Multi-method projection

The multi-method projection (MMP) is a simple linear combination of the probabilistic projections from all five methods (i.e. Methods A–E), all equally weighted. To do so, 5000 equally likely outcomes were randomly resampled from the each of the methods (for some of the methods many of these are duplicates) and combined to produce an MMP ensemble of 25,000, the large number number of samples are used so that we can fairly combine the different methods. The results are not sensitive to the exact number of samples that make up the MMP. This method was used for both the constrained and unconstrained projections. The unconstrained MMP is not in itself particularly meaningful as it includes several duplicated CMIP5 ensembles in some cases because these are used in multiple methods; nonetheless, the unconstrained MMP is included in some of the analysis to act as a baseline from which to examine the relative performance of the constrained multi-method projection.

Pseudo-observational datasets

The pseudo-observational datasets were taken from 125 available ensemble members from the Coupled Model Intercomparison Project 6 (CMIP6) archive that had data for all variables required for the historical (1860–2014) and *ssp585* (2015–2099) experiments. The datasets were downloaded and regridded to a common 2.5° × 2.5° grid. Each ensemble member were then anonymised by stripping most of the metadata from the netCDF files and labelled with a random number from 1 to 125. Data from 1860 to 2014 in these pseudo-observational datasets were then all uploaded to the online public Zenodo archive[25] and made available to the different groups. The future portion of the pseudo-observations was held back and revealed only when the projections from all groups/methods had been produced, to ensure an out-of-sample test. The specific CMIP6 model and member that has been randomly assigned as each pseudo-observation are listed in Supplementary Data.

Verification metrics

The strength of the out-of-sample projections is that the probabilistic projections and the actual change in the pseudo-observations can be used to establish the accuracy and reliability of the constraining methods. Here we calculate the following verification metrics:

- Root-mean-square error (RMSE), which measures the accuracy of the ensemble mean projection with respect to the actual future change in the pseudo-observations and is calculated as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\overline{f_i} - o_i)^2},$$

  where $\overline{f_i}$ and o_i are the ensemble mean projected change and observed change, respectively, for pseudo-observation i.

- Spread/Error, which is the ratio of the spread of projected changes (measured by the standard deviation) and the ensemble mean error (e.g. ref. 65); a ratio of less than one indicates an overconfident prediction, whereas a ratio of greater than one indicates an underconfident prediction. Specifically, it is calculated as:

$$\mathrm{Spread/Error} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} s_i^2}}{\mathrm{RMSE}},$$

  where s_i^2 is the sample variance of the projected ensemble change (i.e. Fig. 1b).

- Continuous ranked probability score (CRPS), which measures the difference between the cumulative density function of a probabilistic prediction and the verifying observation (e.g. refs. 66,67). The CRPS is lower when the probabilistic projections are more accurate and is commonly used in weather forecast evaluation (e.g. ref. 68); the CRPS can be considered a generalised version of the RMSE metric applied to full probabilistic predictions. The CRPS for the projection of each pseudo-observation i is calculated as:

$$\mathrm{CRPS}_i = \int_{-\infty}^{\infty} [P_f(x_i) - P_o(x_i)]^2 dx,$$

  where $P_f(x_i)$ is the cumulative density function of the projected change in quantity x_i and $P_o(x_i)$ is the corresponding distribution for the pseudo-observation, denoted by the Heaviside function:

$$P_o(x_i) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

  The CRPS, which has the same units as the variable in question (i.e. x), is calculated by averaging over all pseudo-observations:

$$\mathrm{CRPS} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{CRPS}_i.$$

- Continuous ranked probability skill score (CRPSS): In addition to CRPS, we also examine a related skill score, CRPSS, which we define here as a measure of the relative skill of the constrained probabilistic projections with respect to the unconstrained probabilistic projections. The CRPSS is calculated as follows:

$$\mathrm{CRPSS} = 1 - \frac{\mathrm{CRPS}_{\mathrm{const}}}{\mathrm{CRPS}_{\mathrm{unconst}}}.$$

  When the constrained projections are more accurate than the unconstrained projections, the (unit-less) CRPSS is positive, whereas when the constrained projection is less accurate than the unconstrained projections, the CRPSS is negative. The CRPSS is a useful measure for comparing relative improvements across the methods. However, CRPSS is sensitive to the level of accuracy of the unconstrained projections (i.e. $\mathrm{CRPS}_{\mathrm{unconst}}$) such that less accurate unconstrained projections tend to be associated with higher CRPSS (Supplementary Fig. 9).

Confidence intervals

The 95% confidence intervals shown on the verification plots were calculated using a random Monte Carlo bootstrap resampling of the projection-verification pairs, with replacement, to match the number used to calculate the actual verification metric (i.e. $n = 125$ for Figs. 3, 4). This random resampling was repeated 1000 times to generate the 95% confidence

intervals shown on the verification plots. Estimates of p values of the differences between all pairs of points for each of the verification verification metrics are not shown for practical reasons; however, approximate p values of the differences can be inferred by eye based on the level of overlap of the individual arms of the confidence intervals (e.g. refs. 69,70). Where the proportion of overlap (measured in terms of the average length of the different confidence interval arms) is equal to 1 indicates $p \approx 0.2$ and equal to 1/2 indicates $p \approx 0.05$. Confidence intervals with zero overlap but just touching have $p \approx 0.01$[71].

Data availability

The model datasets used in this simulation are mostly available online from the CMIP5 and CMIP6 archives. The anonymised pseudo-observational datasets have been uploaded to the Zenodo archive[25]. The constrained projections made by each of the methods have also been uploaded to the Zenodo archive[72].

Code availability

The codes and methodologies underlying the constraints applied and analysed in this paper have been documented in previous manuscripts (see Methods for specific references).

Received: 4 January 2023; Accepted: 15 April 2024;
Published online: 26 April 2024

References

1. Field, C. B., Barros, V., Stocker, T. F. & Dahe, Q. *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change* (Cambridge Univ. Press, 2012).
2. Knutti, R. & Sedláček, J. Robustness and uncertainties in the new cmip5 climate model projections. *Nat. Clim. Change* **3**, 369–373 (2013).
3. Stocker, T. *Climate change 2013: The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge Univ. Press, 2014).
4. Lehner, F. et al. Partitioning climate projection uncertainty with multiple large ensembles and cmip5/6. *Earth Syst. Dyn.* **11**, 491–508 (2020).
5. Masson-Delmotte, V. et al. Climate change 2021: the physical science basis. *Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change* 2 (2021).
6. Knutti, R., Masson, D. & Gettelman, A. Climate model genealogy: generation cmip5 and how we got there. *Geophys. Res. Lett.* **40**, 1194–1199 (2013).
7. Sanderson, B. M., Knutti, R. & Caldwell, P. A representative democracy to reduce interdependency in a multimodel ensemble. *J. Clim.* **28**, 5171–5194 (2015).
8. Knutti, R. et al. A climate model projection weighting scheme accounting for performance and interdependence. *Geophys. Res. Lett.* **44**, 1909–1918 (2017).
9. Giorgi, F. & Mearns, L. O. Calculation of average, uncertainty range, and reliability of regional climate changes from aogcm simulations via the "reliability ensemble averaging"(rea) method. *J. Clim.* **15**, 1141–1158 (2002).
10. Giorgi, F. & Mearns, L. O. Probability of regional climate change based on the Reliability Ensemble Averaging (REA) method. *Geophys. Res. Lett.* **30**, 1629 (2003).
11. Brunner, L., Lorenz, R., Zumwald, M. & Knutti, R. Quantifying uncertainty in European climate projections using combined performance-independence weighting. *Environ. Res. Lett.* **14**, 124010 (2019).
12. Merrifield, A. L., Brunner, L., Lorenz, R., Medhaug, I. & Knutti, R. An investigation of weighting schemes suitable for incorporating large ensembles into multi-model ensembles. *Earth Syst. Dyn.* **11**, 807–834 (2020).
13. Brunner, L. et al. Reduced global warming from cmip6 projections when weighting models by performance and independence. *Earth Syst. Dyn.* **11**, 995–1012 (2020).
14. Ribes, A., Qasmi, S. & Gillett, N. P. Making climate projections conditional on historical observations. *Sci. Adv.* **7**, eabc0671 (2021).
15. Allen, M. R. & Stott, P. A. Estimating signal amplitudes in optimal fingerprinting, part i: theory. *Clim. Dyn.* **21**, 477–491 (2003).
16. Stott, P. A., Kettleborough, J. A. & Allen, M. R. Uncertainty in continental-scale temperature predictions. *Geophys. Res. Lett.* **33**, L02708 (2006).
17. Kettleborough, J., Booth, B., Stott, P. & Allen, M. Estimates of uncertainty in predictions of global mean surface temperature. *J. Clim.* **20**, 843–855 (2007).
18. Hewitt, C. D. & Lowe, J. A. Toward a european climate prediction system. *Bull. Am. Meteorol. Soc.* **99**, 1997–2001 (2018).
19. Brunner, L. et al. Comparing methods to constrain future european climate projections using a consistent framework. *J. Clim.* **33**, 8671–8692 (2020).
20. Hall, A., Cox, P., Huntingford, C. & Klein, S. Progressing emergent constraints on future climate change. *Nat. Clim. Change* **9**, 269–278 (2019).
21. Sanderson, B. M. On the estimation of systematic error in regression-based predictions of climate sensitivity. *Clim. Change* **118**, 757–770 (2013).
22. Caldwell, P. M., Zelinka, M. D. & Klein, S. A. Evaluating emergent constraints on equilibrium climate sensitivity. *J. Clim.* **31**, 3921–3942 (2018).
23. Schlund, M., Lauer, A., Gentine, P., Sherwood, S. C. & Eyring, V. Emergent constraints on equilibrium climate sensitivity in cmip5: do they hold for cmip6? *Earth Syst. Dyn.* **11**, 1233–1258 (2020).
24. Simpson, I. R. et al. Emergent constraints on the large-scale atmospheric circulation and regional hydroclimate: do they still work in cmip6 and how much can they actually constrain the future? *J. Clim.* **34**, 6355–6377 (2021).
25. O'Reilly, C. H. Pseudo-observational datasets for testing projection calibration methods (EUCP WP2), https://doi.org/10.5281/zenodo.3892252 (2020).
26. Smith, C. J. et al. Effective radiative forcing and adjustments in cmip6 models. *Atmos. Chem. Phys.* **20**, 9591–9618 (2020).
27. Eyring, V. et al. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geosci. Model Dev.* **9**, 1937–1958 (2016).
28. Palmer, T. E., Booth, B. & McSweeney, C. F. How does the cmip6 ensemble change the picture for european climate projections? *Environ. Res. Lett.* **16**, 094042 (2021).
29. Schurer, A. et al. Estimating the transient climate response from observed warming. *J. Clim.* **31**, 8645–8663 (2018).
30. Hagedorn, R., Doblas-Reyes, F. J. & Palmer, T. N. The rationale behind the success of multi-model ensembles in seasonal forecasting—i. basic concept. *Tellus A* **57**, 219–233 (2005).
31. Weigel, A. P., Liniger, M. & Appenzeller, C. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Q. J. R. Meteorol. Soc.* **134**, 241–260 (2008).
32. Zelinka, M. D. et al. Causes of higher climate sensitivity in cmip6 models. *Geophys. Res. Lett.* **47**, e2019GL085782 (2020).
33. Meehl, G. A. et al. Context for interpreting equilibrium climate sensitivity and transient climate response from the cmip6 earth system models. *Sci. Adv.* **6**, eaba1981 (2020).
34. Fyfe, J. C., Kharin, V. V., Santer, B. D., Cole, J. N. & Gillett, N. P. Significant impact of forcing uncertainty in a large ensemble of climate model simulations. *Proc. Natl Acad. Sci. USA* **118**, e2016549118 (2021).

35. Hawkins, E. & Sutton, R. Time of emergence of climate signals. *Geophys. Res. Lett.* **39**, L01702 (2012).
36. Lehner, F., Deser, C. & Terray, L. Toward a new estimate of "time of emergence" of anthropogenic warming: Insights from dynamical adjustment and a large initial-condition model ensemble. *J. Clim.* **30**, 7739–7756 (2017).
37. O'Reilly, C. H. et al. Projections of northern hemisphere extratropical climate underestimate internal variability and associated uncertainty. *Commun. Earth Environ.* **2**, 1–9 (2021).
38. European Environment Agency. Europe's changing climate hazards — an index-based interactive eea report (2021).
39. O'Reilly, C. H., Woollings, T. & Zanna, L. The dynamical influence of the atlantic multidecadal oscillation on continental climate. *J. Clim.* **30**, 7213–7230 (2017).
40. Deser, C., Hurrell, J. W. & Phillips, A. S. The role of the north atlantic oscillation in european climate projections. *Clim. Dyn.* **49**, 3141–3157 (2017).
41. Seager, R. et al. Causes of increasing aridification of the mediterranean region in response to rising greenhouse gases. *J. Clim.* **27**, 4655–4676 (2014).
42. Seager, R. et al. Climate variability and change of mediterranean-type climates. *J. Clim.* **32**, 2887–2915 (2019).
43. Brogli, R., Sørland, S. L., Kröner, N. & Schär, C. Causes of future mediterranean precipitation decline depend on the season. *Environ. Res. Lett.* **14**, 114017 (2019).
44. Seager, R. et al. Strengthening tropical pacific zonal sea surface temperature gradient consistent with rising greenhouse gases. *Nat. Clim. Change* **9**, 517–522 (2019).
45. Andrews, T. et al. Accounting for changing temperature patterns increases historical estimates of climate sensitivity. *Geophys. Res. Lett.* **45**, 8490–8499 (2018).
46. Sherwood, S. C. et al. An assessment of earth's climate sensitivity using multiple lines of evidence. *Rev. Geophys.* **58**, e2019RG000678 (2020).
47. Hausfather, Z., Marvel, K., Schmidt, G. A., Nielsen-Gammon, J. W. & Zelinka, M. Climate simulations: recognize the 'hot model' problem. *Nature* **605**, 26–29 (2022).
48. Lowe, J. A. et al. Ukcp18 science overview report. *Met Office Hadley Centre: Exeter, UK* (2018).
49. Fischer, A. et al. Climate Scenarios for Switzerland CH2018 – Approach and Implications. *Clim. Serv.* **21**, 100288 (2019).
50. Tegegne, G., Kim, Y.-O. & Lee, J.-K. Spatiotemporal reliability ensemble averaging of multimodel simulations. *Geophys. Res. Lett.* **46**, 12321–12330 (2019).
51. O'Reilly, C. H., Befort, D. J. & Weisheimer, A. Calibrating large-ensemble european climate projections using observational data. *Earth Syst. Dyn.* **11**, 1033–1049 (2020).
52. Kay, J. 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. *Bull. Am. Meteorol. Soc.* **96**, 1333–1349 (2015).
53. Deser, C., Terray, L. & Phillips, A. S. Forced and internal components of winter air temperature trends over north america during the past 50 years: mechanisms and implications. *J. Clim.* **29**, 2237–2258 (2016).
54. Gneiting, T., Raftery, A. E., Westveld III, A. H. & Goldman, T. Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Mon. Weather Rev.* **133**, 1098–1118 (2005).
55. Lorenz, R. et al. Prospects and caveats of weighting climate models for summer maximum temperature projections over North America. *J. Geophys. Res. Atmos.* **123**, 4509–4526 (2018).
56. Amos, M. et al. Projecting ozone hole recovery using an ensemble of chemistry–climate models weighted by model performance and independence. *Atmos. Chem. Phys.* **20**, 9961–9977 (2020).
57. Gründemann, G. J., van de Giesen, N., Brunner, L. & van der Ent, R. Rarest rainfall events will see the greatest relative increase in magnitude under future climate change. *Commun. Earth Environ.* **3**, 235 (2022).
58. Sperna Weiland, F. C. et al. Estimating regionalized hydrological impacts of climate change over Europe by performance-based weighting of CORDEX projections. *Front. Water* **3**, 713537 (2021).
59. Befort, D. J. et al. Combination of decadal predictions and climate projections in time: challenges and potential solutions. *Geophys. Res. Lett.* **49**, e2022GL098568 (2022).
60. Ribes, A. et al. An updated assessment of past and future warming over france based on a regional observational constraint. *Earth Syst. Dyn.* **13**, 1397–1415 (2022).
61. Qasmi, S. & Ribes, A. Reducing uncertainty in local temperature projections. *Sci. Adv.* **8**, eabo6872 (2022).
62. DelSole, T., Trenary, L., Yan, X. & Tippett, M. K. Confidence intervals in optimal fingerprinting. *Clim. Dyn.* **52**, 4111–4126 (2019).
63. Hegerl, G. C. et al. Toward consistent observational constraints in climate predictions and projections. *Front. Clim.* **3**, 43 (2021).
64. Shiogama, H. et al. Predicting future uncertainty constraints on global warming projections. *Sci. Rep.* **6**, 1–7 (2016).
65. Fortin, V., Abaza, M., Anctil, F. & Turcotte, R. Why should ensemble spread match the rmse of the ensemble mean? *J. Hydrometeorol.* **15**, 1708–1713 (2014).
66. Hersbach, H. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast.* **15**, 559–570 (2000).
67. Jolliffe, I. T. & Stephenson, D. B. *Forecast Verification: A Practitioner's Guide in Atmospheric Science* (John Wiley & Sons, 2012).
68. Leutbecher, M. & Haiden, T. Understanding changes of the continuous ranked probability score using a homogeneous gaussian approximation. *Q. J. R. Meteorol. Soc.* **147**, 425–442 (2021).
69. Cumming, G. & Finch, S. Inference by eye: confidence intervals and how to read pictures of data. *Am. Psychol.* **60**, 170 (2005).
70. Belia, S., Fidler, F., Williams, J. & Cumming, G. Researchers misunderstand confidence intervals and standard error bars. *Psychol. Methods* **10**, 389 (2005).
71. Cumming, G. Inference by eye: reading the overlap of independent confidence intervals. *Stat. Med.* **28**, 205–220 (2009).
72. O'Reilly, C. Calibrated and uncalibrated projection data from the paper "Assessing observational constraints on future European climate in an out-of-sample framework". https://doi.org/10.5281/zenodo.10931996 (2024).

Acknowledgements

This study was supported by the EUCP project funded by the European Union under Horizon 2020 (Grant Agreement 776613). C.H.O. was supported by a Royal Society University Research Fellowship.

Author contributions

C.H.O. performed the analysis, conceived of the study and lead the writing of the manuscript. C.H.O. and D.J.B. prepared the anonymised pseudo-observational datasets and made them available online. C.H.O., L.B., S.Q., A.P.B., A.P.S. and A.R. applied the different constraining methods. All authors contributed to the analysis of the results and the writing of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41612-024-00648-8.

Correspondence and requests for materials should be addressed to Christopher H. O'Reilly.