Calibration and bias correction of climate projections for crop modelling: an idealised case study over Europe

Ed Hawkins[a,*], Thomas M. Osborne[b], Chun Kit Ho[a], Andrew J. Challinor[c]

[a]*NCAS-Climate, Department of Meteorology, University of Reading*
[b]*NCAS-Climate, Walker Institute, University of Reading*
[c]*Institute for Climate and Atmospheric Science, University of Leeds*

Abstract

Producing projections of future crop yields requires careful thought about the appropriate use of atmosphere-ocean global climate model (AOGCM) simulations. Here we describe and demonstrate multiple methods for 'calibrating' climate projections using an ensemble of AOGCM simulations in a 'perfect sibling' framework. Crucially, this type of analysis assesses the ability of each calibration methodology to produce reliable estimates of future climate, which is not possible just using historical observations. This type of approach could be more widely adopted for assessing calibration methodologies for crop modelling. The calibration methods assessed include the commonly used 'delta' (change factor) and 'nudging' (bias correction) approaches. We focus on daily maximum temperature in summer over Europe for this idealised case study, but the methods can be generalised to other variables and other regions. The calibration methods, which are relatively easy to implement given appropriate observations, produce more robust projections of future daily maximum temperatures and heat stress than using raw model output. The choice over which calibration method to use will likely depend on the situation, but change factor approaches tend to perform best in our examples. Finally, we demonstrate that the uncertainty due to the choice of calibration methodology is a significant contributor to the total uncertainty in future climate projections for impact studies. We conclude that utilising a variety of calibration methods on output from a wide range of AOGCMs is essential to produce climate data that will ensure robust and reliable crop yield projections.

Keywords: calibration, climate projections, climate model, crop model, delta method, weather generator, bias correction

1. Introduction

There is a growing need to produce crop yield projections for the next few decades to enable effective adaptation to climate variability and change. It is known from case studies of the recent past that crop yields are seen to reduce in particularly hot seasons (e.g., Battisti and Naylor, 2009), and producing estimates of the number and extent of such seasons in the future may aid crop breeding or motivate a change in the crops grown in a particular location.

Climate information for assessments of future crop yields tends to come from atmosphere-ocean global climate models (AOGCMs). These models attempt to represent the full Earth system, and simulate the future with assumed scenarios for anthropogenic emissions, producing projections of future climate (e.g., Meehl et al., 2007). However, there are a number of issues to address in using output from AOGCMs to drive crop models. Firstly, the size of the AOGCM grid cell is normally far larger than required for crop models, meaning that some form of spatial downscaling is required (e.g., Baron et al., 2005). Secondly, the reliability and realism of the daily output from AOGCMs needs to be assessed. The next set of simulations for the Coupled Model Intercomparison Project (CMIP5), which will be examined by the Intergovernmental Panel on Climate Change (IPCC), will make more daily output available at higher spatial resolution than previous assessments, allowing a more comprehensive assessment. Thirdly, no AOGCM is a perfect representation of the true climate and so some 'calibration' of the raw climate model output would appear to be appropriate, where calibration refers to any attempt to make the AOGCM output more realistic. A wide variety of approaches have been adopted to produce calibrated data for crop yield projections (see Section 2). Weather generators are one such tool; they are often designed specifically with crop modelling applications in mind (e.g. Hansen and Ines, 2005; Semenov et al., 2010; Ines et al., 2011). Although we will not consider weather generators directly in this study, the findings have implications for their design.

As an example of these issues, Fig. 1 shows the mean daily maximum temperature ($T_{\max}$) during summer (June-July-August) for the period 1970-1999 from the E-OBS v5.0 0.5° observations (top left; Haylock et al., 2008) and a range of AOGCMs over Europe. The AOGCMs have different spatial resolutions, but all have larger grid cells than the observational data available. It is immediately obvious that many features visible in the observations are not seen in

*Corresponding author
Email address: e.hawkins@reading.ac.uk (Ed Hawkins)

the AOGCMs, e.g. the cooler temperatures over the Alps. Additionally, the AOGCMs show a wide range of temperatures for the same location, differing by more than 6°C in some places, and all exhibit a bias from observations which varies spatially. A crucial point to appreciate is that even if all the AOGCMs produce the same future temperature *change* as a response to radiative forcings such as greenhouse gases, the *absolute* value of the temperatures will be very different. As most crops are known to be sensitive to absolute thresholds in temperature (e.g., Vara Prasad et al., 2000; Schlenker and Roberts, 2009), these biases are problematic, and require correcting. Additionally, the various AOGCMs produce different estimates for the future rate and magnitude of warming to increasing anthropogenic forcing.

As more daily data from both AOGCMs and observations (e.g., Caesar et al., 2006; Xie et al., 2007; Haylock et al., 2008) becomes available, an important issue is how best to combine output from the wide range of AOGCMs with observational data to produce robust future climate data relevant for crop impact assessment. The motivation for this paper is to address this key question in an idealised 'perfect sibling' framework, where a reference simulation of current climate is treated as pseudo-observations, and independent simulations with different climate models are used to try and predict the future climate of the reference simulation. A related approach was adopted by Lobell and Burke (2010) to test the ability of statistical crop models to reproduce the output from a process-based crop model. Importantly, this type of analysis allows calibration methods to be assessed into the future, something which is not possible solely using historical observations and present-day model simulations. This aspect is particularly relevant as Ho et al. (2012) demonstrated that the choice of calibration strategy can produce differences in future climate which are as large as that between future emissions scenarios, and are therefore a potentially large source of uncertainty.

Some commonly used methods for producing climate data for crop models are briefly discussed in Section 2. The calibration methodologies to be compared in this study are defined in Section 3, and they include some of the commonly used methods in making crop projections, such as the 'delta' method. The climate model datasets used, from two structurally different AOGCMs, are described in Section 4. The calibration validation results are shown in Section 5 and we discuss their implications for crop impact assessment in Section 6.

2. Existing methods for producing weather data for crop models

A full review of methods for transforming climate model output for use with crop models is beyond the scope of this study. However, we now briefly discuss six general approaches for producing weather data for crop models to evaluate future yield under long-term climatic changes:

1. Use the raw daily output from the AOGCMs: this approach has the advantage of being simple, but may suffer from the inherent biases of the AOGCMs (e.g., Fig. 1). However, for some crop models, it is possible to use the parameter calibration procedure to successfully correct for both the yield gap and climate model bias at the same time (e.g. Challinor et al., 2005b).
2. Coupled crop-climate models: this approach integrates a crop model inside a climate model, ensuring that the interactions and feedbacks between climate and crops are represented (Osborne et al., 2007, 2009). This approach will also suffer from the impact of any climate model biases.
3. Dynamical downscaling: this methodology uses a regional climate model (RCM) to downscale the output from a coarser resolution AOGCM and produce data to drive the crop model (e.g. Challinor et al., 2007). Although this approach may help reduce regional biases, it will not eliminate them because the boundary conditions on which the RCM relies will be biased, and some further calibration may be necessary (e.g. Arnell et al., 2003).
4. Use a weather generator: this description encompasses a wide range of methodologies. In general terms, this approach fits a statistical model to (daily) observations, and uses a change in mean climate (usually using monthly means) from one or more AOGCMs with the statistical model to generate daily data in the future (e.g. Mearns et al., 1997; Liu et al., 2008; Thornton et al., 2011; Ines et al., 2011; Osborne et al., 2012). This has the advantage of being able to produce a large number of possible realisations of future climate, but assumes that the statistical model can produce the correct ranges of variability. The statistical models used in weather generators are continually being developed, for example through particular attention to reproducing the observed spatial and temporal characteristics of rainfall (e.g. Hansen et al., 2006; Baigorria and Jones, 2010).
5. The 'delta' method: although there is no clear single definition for this approach, we will define it as adding a (usually monthly) mean change in climate, derived from AOGCMs, onto the daily observations themselves, without the need for a weather generator (e.g., Arnell et al., 2003; Wilby

et al., 2004; Diaz-Nieto and Wilby, 2005). This method has the advantage of using the correct distribution of daily data, but only allows a single realisation of future climate and assumes that a representative period of observations is available. Additionally, a simple additive shift will not work for precipitation; a multiplicative shift is usually adopted.

6. Simple bias correction (or 'nudging'): this approach adds the difference between AOGCM and observations in a reference period to the future AOGCM data to correct the mean bias (e.g., Huntingford et al., 2005; Ines and Hansen, 2006). However, this method uses the AOGCM distributions of daily climate, aspects of which may also need correcting, e.g. the temporal correlation or skewness.

For each approach there are a number of choices to be made in this process, e.g. which AOGCMs to use? Should the variability and/or skewness be changed as well as the mean? How should the AOGCM data be spatially downscaled? There appears to be a need for a systematic comparison of some of these different approaches in producing reliable daily climate variability for future periods. This is possible only in a 'perfect sibling' framework, where data from a set of simulations are calibrated and verified against an independent climate simulation which is treated as pseudo-observations (or 'truth'), and this is the framework we adopt.

We utilise daily output from a range of AOGCMs to analyse some different calibration approaches. Although we do not consider any weather generators directly we can draw conclusions on the relevant aspects of any weather generator which are important. Additionally, the methods we consider could also be used on output from RCMs to try and reduce remaining biases.

3. Calibration methods

The simplest way to use AOGCM output to drive crop models is to use the raw output directly. However, there are biases between the AOGCM and reality (Fig. 1) which should be corrected. In addition, the spatial scale of AOGCM output is far larger than usual crop model spatial scales. Here we consider four methods for calibrating daily AOGCM output to produce more realistic projections, as summarised by Ho et al. (2012). These methodologies have the advantage that they are independent of the shape of the distribution of climate data, and also downscale the projections to the spatial scale of the available observations. Although no downscaling occurs in this study because of the 'perfect sibling' approach, this downscaling would be achieved when considering observations

by simply using the same AOGCM data for multiple observed locations within the AOGCM grid cell.

We will only consider daily maximum temperature ($T_{\max}$) in summer over Europe for this case study. To perform a calibration we require daily temperature timeseries from an AOGCM and observations for a particular location for the same reference period, which we denote by $T_{\mathrm{REF}}(t)$ and $O_{\mathrm{REF}}(t)$ respectively. We also need output from the AOGCM for some future period of the same length as the reference period, $T_{\mathrm{RAW}}(t)$. The question remains about how to best combine these three sources of information into the most robust projections of the unknown future observations ($\widehat{O}_{\mathrm{FUT}}$) to use as input for crop models. We consider two general approaches (Fig. 2), namely bias correction and change factor, with and without including corrections for the variability as well as the mean climate. Ho et al. (2012) demonstrate that these various approaches can give differences in future calibrated climates which are as large as differences between emission scenarios.

3.1. Bias Correction

The bias correction (BC) methodology (Fig. 2a) corrects the projected raw daily AOGCM output using the differences in the mean and variability between observations and the AOGCM in a particular reference period (Huntingford et al., 2005; Ines and Hansen, 2006). In the simplest case, where the variability in observations and AOGCM is assumed to be the same, the daily data is simply shifted by the mean bias in the reference period,

$$T_{\mathrm{SH}}(t) = T_{\mathrm{RAW}}(t) + \left(\overline{O_{\mathrm{REF}}} - \overline{T_{\mathrm{REF}}}\right), \tag{1}$$

where the time mean is denoted by the bar above a symbol. However, in a more general case when the variability is corrected also (Ho et al., 2012; also see Appendix A),

$$T_{\mathrm{BC}}(t) = \overline{O_{\mathrm{REF}}} + \frac{\sigma_{O,\mathrm{REF}}}{\sigma_{T,\mathrm{REF}}}\left(T_{\mathrm{RAW}}(t) - \overline{T_{\mathrm{REF}}}\right), \tag{2}$$

and $\sigma_{T,\mathrm{REF}}$ and $\sigma_{O,\mathrm{REF}}$ represent the standard deviation of the daily AOGCM output and observations in the reference period respectively.

3.2. Change Factor

The change factor (CF) methodology (Fig. 2b) instead utilises the *observed* daily variability and changes the mean and daily variance as simulated by the AOGCM (e.g. Arnell et al., 2003; Gosling et al., 2009). In the simplest case this

is the ‘delta method’, where the daily variability is assumed to have the same magnitude in the future and reference periods, and the corrected daily data is,

$$T_{\mathrm{DEL}}(t) = O_{\mathrm{REF}}(t) + \left(\overline{T_{\mathrm{RAW}}} - \overline{T_{\mathrm{REF}}}\right). \tag{3}$$

However, the more general form, considering changes in variance, is (Ho et al., 2012; also see Appendix A),

$$T_{\mathrm{CF}}(t) = \overline{T_{\mathrm{RAW}}} + \frac{\sigma_{T,\mathrm{RAW}}}{\sigma_{T,\mathrm{REF}}}\left(O_{\mathrm{REF}}(t) - \overline{T_{\mathrm{REF}}}\right), \tag{4}$$

and $\sigma_{T,\mathrm{RAW}}$ represents the standard deviation of the daily raw model output for the future period.

3.3. Choice of methods and caveats

The future evolution of $T_{\max}$ derived from the various methods will be different. Later we will compare results for $\widehat{O}_{\mathrm{FUT}}$ using the five different methods: T_{RAW}, T_{SH}, T_{BC}, T_{DEL} and T_{CF}.

Both BC and CF methods transform the mean and daily variance but CF starts with the variability from observations and BC starts with the AOGCM variability. There is no obvious *a priori* reason to prefer one of these two approaches over the other. The DEL and SH methods do not correct or change the variability - this may be advantageous if the AOGCM does not predict the correct change in variability. Also note that the time mean of T_{SH} is the same as T_{DEL},

$$\overline{T_{\mathrm{SH}}} = \overline{T_{\mathrm{DEL}}} = \overline{T_{\mathrm{RAW}}} - \overline{T_{\mathrm{REF}}} + \overline{O_{\mathrm{REF}}}. \tag{5}$$

However, it is vital to appreciate that considering the variability may also be important for climate impacts such as crops (e.g., Semenov and Porter, 1995; Mearns et al., 1996, 1997; Porter and Semenov, 2005), heat-related mortality (Gosling et al., 2009) or runoff (e.g. Arnell et al., 2003). Crop damage occurs when temperatures cross certain thresholds (e.g., Vara Prasad et al., 2000; Schlenker and Roberts, 2009), and the frequency of such temperatures depends critically on the projected daily variability as well as the mean. An AOGCM projection with too high or too low daily variability will produce different numbers of days over a set threshold, even if the mean is the correct. All the above methodologies will produce different realisations of the variability, but note that the expected variance (V) for T_{CF} is the same as T_{BC},

$$V(T_{\mathrm{CF}}) = V(T_{\mathrm{BC}}) = \frac{\sigma^2_{O,\mathrm{REF}}\,\sigma^2_{T,\mathrm{RAW}}}{\sigma^2_{T,\mathrm{REF}}}. \tag{6}$$

In deriving the equations above it was assumed that the trends in the climate variables are far smaller than the variability. This assumption will often hold for regional daily data but necessitates using periods without significant trends. Here we will use 30 years of data, choosing 1970-99 as the reference period and 2030-59 as the future period. We also note that variables with a strong seasonal cycle may need to perform calibrations using individual months rather than seasons.

It was also assumed above that only the mean and variance of the daily distributions required correcting for BC and CF methods. However, it is possible that, for particular location, the *shape* of the daily distributions of observations and model output is not the same (BC), or the shape changes in the future (CF). For example, they might have a different skewness. In all that follows here we assume the distributions do have the same shape, but further corrections can be made to produce adjusted calibrations if the distributions are significantly different (Ho, 2010). This could be particularly important for the DEL and BC methods as they utilise model output daily variability rather than observed daily variability. It may also be more important for precipitation, which is not considered here.

4. Climate model data

4.1. The QUMP ensemble

To test various calibration methodologies in producing climate projections relevant to crop modelling, we mainly utilise data from a QUMP (Quantifying Uncertainty in Model Predictions) ensemble of AOGCM simulations (Murphy et al., 2004; Collins et al., 2006; Collins et al., 2011).

The QUMP ensemble used here consists of 16 different versions of the HadCM3 AOGCM (Gordon et al., 2000). Each member of the ensemble differs only in the chosen values of particular uncertain atmospheric parameters which govern physical processes which are not fully resolved in the model, e.g. certain cloud parameters. This approach produces what is termed a 'perturbed physics' ensemble (see Collins et al., 2006 for more details). Each QUMP ensemble member has an identical model structure with an atmospheric resolution of $2.5^\circ \times 3.75^\circ$.

In this analysis, we use daily $T_{\max}$ from each ensemble member from 1970-2059. Historical radiative forcings were used before year 2000, and the SRES A1B emissions scenario was followed after year 2000. This ensemble is ideal for this type of study as all the members produce a reasonable climate for the present day (globally) but have different variability characteristics and regional biases. Also, all the data is on the same spatial grid and the different members

have a wide range of climate sensitivities and produce different future climates (Collins et al., 2006; Collins et al., 2011).

4.2. IPSL CM4 data

Ultimately, the methods described here will be applied to the historical observations to produce calibrated projections of future climate. However, as climate models have biases it is important to perform this calibration methodology assessment with AOGCMs which are as different as possible, to examine how well the methodologies deal with large biases. Although the QUMP ensemble members have slightly different climates, they are all produced by an AOGCM with the same structure, which may limit the differences between the members (e.g. Masson and Knutti, 2011).

To overcome this we also utilise daily $T_{\max}$ from a simulation with the IPSL CM4 AOGCM (Marti et al., 2010) for the same time period and emissions scenario as used for QUMP. In addition, we also use the output from a different emissions scenario (SRES A2) in Section 6.2. IPSL CM4 has a very similar atmospheric resolution to HadCM3 but is structurally very different and has a different mean climate (Fig. 1). In addition, the temporal and spatial variability characteristics are different to the QUMP ensemble - in particular, the daily variance is much smaller for most locations over Europe (not shown). Therefore performing a QUMP–IPSL comparison is more akin to comparing with real observations than comparisons within the QUMP ensemble alone.

5. Evaluation of calibration methods

To demonstrate the use of the calibration methodologies, we use the perfect sibling framework, i.e. use reference period data from one AOGCM simulation as pseudo-observations, and attempt to predict the future evolution of that simulation using other independent simulations. This process is often a useful first step in examining methodologies for improving projections, and could be more widely adopted as a validation test for impact analyses (e.g., Räisänen and Palmer, 2001; de Ela et al., 2002; Lobell and Burke, 2010).

5.1. Cross-validation using QUMP

5.1.1. Future changes in mean summer daily $T_{\max}$

We first select the daily summer (JJA) $T_{\max}$ data from 1970-1999 from a particular QUMP member (#4) to act as pseudo-observations ('truth' or reference simulation), and compare with the same data from two other QUMP members (#8, #13). QUMP8 and QUMP4 produce a rather similar mean JJA $T_{\max}$ in the

period chosen, but QUMP13 is cooler (left column of Fig. 3). These particular ensemble members are selected to demonstrate two different situations where the AOGCM agrees well, or not, with the reference simulation. All three of these QUMP ensemble members disagree on the mean daily JJA $T_{\max}$ for the future period 2030-2059, denoted as the 'raw' or uncalibrated data (second column of Fig. 3).

As an assessment of the skill of the methodologies, we calculate the RMSE difference between the spatial patterns of $T_{\max}$ from the reference simulation (QUMP4) and the other two QUMP members, where the mean is calculated over the region displayed in Fig. 3. These RMSE differences between the fields, denoted by E (measured in K) are shown for each time period and calibration methodology.

The DEL & SH, BC and CF calibrated projections of $T_{\max}$, using QUMP4 as the reference simulation are also shown in Fig. 3. Note that the data for 2030-2059 from QUMP4 is *not* used in the calibration, and is solely used for verification. Also, the DEL and SH methods give the same answer by construction (Eqn. 5).

In all the cases shown the RMS error, E, between the calibrated projections and the reference simulation has decreased significantly from the uncalibrated case. In fact, more than 95% of all 240 possible combinations of using different QUMP members as the reference simulation and uncalibrated output produce more accurate projections when calibrated for all the methods (Fig. 4). Overall, DEL & SH are slightly more accurate for this particular domain, with $E = 0.89 \pm 0.25$K. CF and BC have $E = 1.01 \pm 0.35$K and $E = 1.18 \pm 0.50$K respectively. For raw projections, $E = 3.26 \pm 1.60$K.

Fig. 4 also indicates that models with a smaller bias when compared to the reference simulation (as indicated by E_{REF}) show a smaller bias when calibrated using all the methodologies, although there is some scatter. This implies that the development of better models will lead to an improvement in calibrated projections, but for models with a small bias, calibration can occasionally make the projections worse.

5.1.2. Future changes in the number of hot days

The above analysis is repeated for the number of summer days exceeding 30°C (Fig. 5). This particular choice of threshold is motivated by Schlenker and Roberts (2009), who found that maize yields dropped markedly if exposed to temperatures above around 30°C. Again, the calibrated projections perform significantly better than the raw projections (Fig. 5; compare the E values, now in units of number of days).

As found for the mean of $T_{\max}$, nearly all (93%) of all possible QUMP combinations produce more accurate projections using the calibration methods than RAW (not shown). For this metric, DEL and CF methods perform best, with $E = 3.4 \pm 1.3$ and 3.5 ± 1.3 days respectively. BC and SH have errors of $E = 4.1 \pm 1.5$ and $E = 5.7 \pm 2.6$ days respectively.

Overall, these results with the QUMP ensemble suggest the use of the DEL and CF methods is best. The use of SH would not be recommended, although it vastly outperforms RAW. However, this particular analysis has been restricted to a single AOGCM which could bias the results.

5.2. Cross-validation using IPSL data

Similar analyses can be performed using the IPSL data, which comes from a very different AOGCM, and is a lot cooler than the QUMP members over Europe (Fig. 3). However, example cross-calibrations of the IPSL data demonstrate that it can still be calibrated reasonably well to the QUMP4 reference simulation (bottom rows of Figs. 3 and 5).

When considering the IPSL data itself as the reference simulation, and using all 16 QUMP members, all of SH, DEL and CF methods produce an error of $E = 1.0 \pm 0.3$K, and BC an error of $E = 1.9 \pm 0.6$K. For the number of days over 30°C, CF produces the lowest error of $E = 3.8 \pm 1.4$ days, with the other methods producing $E = 4.3 \pm 1.4$ (DEL), 8.1 ± 1.2 (BC), and 9.1 ± 2.2 (SH) days respectively.

This test, although restricted to a single other AOGCM, suggests that CF and DEL methods are again the most reliable, with SH and BC methods performing less well. Again, all the methods are superior to RAW. These examples are simple demonstrations that calibrating AOGCM (or RCM) output could produce more robust projections of climate variables of interest to crop modellers.

5.3. Cross-validation for an individual location

Next we consider calibrated projections for an example location in more detail. Specifically, we select a grid point in south-west France (45°N, 0°E) where a large fraction of French maize is grown (Monfreda et al., 2008). Similar results are found for other locations (not shown).

Using each QUMP member in turn as the reference simulation, and calibrating the independent QUMP simulations, produces a range of results for mean summer $T_{\max}$ for this particular grid point (Fig. 6). The various calibration methods (colours) produce a smaller mean absolute error than using raw output (grey). When considering the number of hot days (Fig. 7), again the calibration methods produce smaller mean absolute errors than RAW (not shown).

The bottom rows of Figs. 6, 7 show the results using all QUMP members, calibrated using the IPSL CM4 AOGCM data as the reference simulation. In this example, the calibration methods get close to the reference simulation even though the IPSL CM4 model is far cooler than the QUMP members over this region. However, the SH method produces too many hot days, probably because it does not correct the differences in variability. A wider cross-AOGCM analysis would be required to examine this finding further.

5.4. Probabilistic measures of change

Note that projections of crop yields are necessarily probabilistic. It is important for predictions to be 'reliable' - meaning that the predicted *probabilities* are correct (e.g. Räisänen and Palmer, 2001).

For example, assuming Gaussian distributions, the standard deviation of the QUMP results represents the 68% confidence intervals, and thus it would be expected that the 'truth' would fall within 1σ uncertainties for around 11 of the 16 QUMP sets, and this is seen (Fig. 6). Further probabilistic measures, such as the normalised errors (Table 1), demonstrate that the calibration methods are producing fairly reliable statistics of the expected change in mean summer T_{max}.

The same probabilistic analysis can be performed on the number of summer days exceeding 30°C (Fig. 7). The performance of the calibration methods would not perhaps be expected to be quite as good, as the number of days over a threshold is far more variable than the mean. In this case, the CF & DEL methods do not produce enough hot days on average and are overconfident (Table 1). In other words, the calibration produces results that are too similar, probably because the *shape* of the distribution of daily temperatures may also need correcting (Ho, 2010). BC and SH methods are the most reliable. This highlights the need to rigorously assess the assumptions underlying the calibration approach in a variety of ways.

5.5. Application to a heat stress parameterisation

We now consider how well the calibration methodologies perform in predicting the heat stress index from a crop model parameterisation - we choose an example of French maize in the GLAM crop model (see Osborne et al., 2012). In this parameterisation the heat stress index for a particular day is,

$$\text{daily heat stress index} = \begin{cases} 1 & \text{if } T_{\mathrm{max}} < T_{\mathrm{crit}} \\ 1 - \frac{T_{\mathrm{max}} - T_{\mathrm{crit}}}{T_{\mathrm{zero}} - T_{\mathrm{crit}}} & \text{if } T_{\mathrm{zero}} > T_{\mathrm{max}} \geq T_{\mathrm{crit}} \\ 0 & \text{if } T_{\mathrm{max}} \geq T_{\mathrm{zero}} \end{cases} \tag{7}$$

Table 1: Probabilistic measures of the reliability of calibrated T_{max} projections for a grid point in south-west France (45°N, 0°E), using the QUMP ensemble. The normalised errors (the error divided by the uncertainty) for a Gaussian distribution should have a mean close to zero and standard deviation (spread) of close to unity. A spread greater than 1 indicates over-confidence.

Method	Mean T_{max}		Number of days over 30°C		Heat stress index	
	Mean error	Spread	Mean error	Spread	Mean error	Spread
DEL	-0.03	1.16	-0.93	1.76	0.68	3.04
SH	-0.03	1.16	0.00	1.09	0.33	1.31
CF	0.05	1.07	-1.21	1.87	-0.24	1.39
BC	-0.05	1.18	-0.24	1.00	0.17	1.11

All quantities are normalised and unitless.

For the illustrative examples here we average this daily heat stress index, which is the reduction factor for harvest index, during July when the maize crop is expected to flower (USDA, 2012), using T_{crit} = 37°C and T_{zero} = 45°C. Note that 1 is no effect on yield in this scheme, opposite to a similar scheme used in Teixeira et al. (2012).

Fig. 8 illustrates the uncertainties in the heat stress index produced when using the raw and calibrated QUMP ensembles. Estimates of the crop yield relevant heat stress from the calibrated temperatures have generally around half the uncertainty when compared to the raw ensemble, and the biases are not significant (not shown). However, the DEL approach appears to be far too confident in its projections (Table 1).

5.6. Projected changes in daily variability

It has been shown above that considering changes in daily variability, as well as the mean, are important. But, are the projected future changes in variability robust? Fig. 9 illustrates the mean projected change in summer daily T_{max}variability for the QUMP ensemble. For many central European areas there is a projected increase in daily variability of several percent. Other areas show little or no change and a few areas show a decrease in daily variability. However, not all the ensemble members agree. The diagonal hatching indicates areas where 12 or more members agree on the sign of the change; the cross-hatched areas indicate agreement of 15 or 16 members.

This uncertainty in future changes in daily variability makes it necessary to examine the robustness of any projected change produced by the AOGCMs, and

to have a good physical understanding for why the projected change may occur. A wider cross-AOGCM study shows similar results to the QUMP ensemble, but demonstrates that the physical reasons disagree amongst the AOGCMs (Fischer and Schär, 2009).

6. Discussion and summary

We now discuss these findings, and their relevance for the design of crop yield projections using the daily output from AOGCMs.

6.1. Intercomparison of calibration methods

It is seen that using raw daily AOGCM data is to be avoided, and calibration methods which use the observed daily variability tend to perform better than those which use model variability. These results therefore favour the change factor ('delta') methods over the bias correction ('nudging') methods. This is probably because the *shapes* of the $T_{\max}$ distributions and temporal correlations of $T_{\max}$ differ more across climate models than they change with time in a single climate model (not shown). The analysis here has not considered changes or differences in the shape of the daily distributions, and this may affect the results, especially for the BC methods, which rely on the shape of the AOGCM output distribution rather than the observations. However, it must also be noted that the DEL & CF results tend to be overconfident in their projections of the number of hot days and the impact of high temperature stress on harvest index, and thus not 'reliable', which is not a desirable feature (Räisänen and Palmer, 2001).

The AOGCM simulations being performed for the 5th Coupled Model Intercomparison Project (CMIP5) will include more daily output and allow a more comprehensive across-AOGCM test of these calibration methodologies.

6.2. Sources of uncertainty

There is a 'cascade' of uncertainty in producing crop yield projections. This cascade ranges from uncertainty in future emissions of greenhouse gases, through a range of AOGCM responses to specified emissions, combined with the natural, internal variability of the climate. In addition, as demonstrated here, there is uncertainty in the choice of calibration method in producing climate data for the impact model, and finally, there is uncertainty in the impact model itself.

For policy relevant advice it is vital that studies on the impacts of climate change consider all the important uncertainties. If not, there is a significant risk of underestimating the total uncertainty in impacts predictions. A key issue is

then quantifying which type (or types) of uncertainty are the most important for a particular impact projection.

For short-term projections, the natural, internal variability of climate tends to dominate the total uncertainty in temperature and precipitation as the climate does not evolve smoothly over time (e.g., Hawkins and Sutton, 2009; 2011). For example, decades which exhibit a cooling trend are expected regularly over Europe, even in a warming climate (Hawkins, 2011).

For lead times longer than about a decade, the uncertainty due to different AOGCM responses to prescribed emissions becomes more important. Towards the end of the 21st century, uncertainty due to different emissions scenarios becomes dominant for temperature for most regions, but remains small for precipitation (Hawkins and Sutton, 2009; 2011). In addition, the various calibration methods provide quite different answers (also see Ho et al., 2012).

In Fig. 10 we attempt to quantify the different sources of uncertainty in producing calibrated projections of $T_{\max}$ in the IPSL AOGCM. We show that the uncertainty due to the choice of calibration is of comparable importance to the spread in calibrated model responses using the QUMP ensemble, and larger than the uncertainty due to the choice of emissions scenario (for the period chosen, 2030-59). This demonstrates the crucial need to consider the uncertainty in calibration choice in crop impact assessments.

Finally, comparisons of the relative importance of impacts model and climate model uncertainty have shown that climate uncertainty can often dominate. This is true of Amazon rainforest tree type fractions (Poulter et al., 2010) and groundnut yield in Western India (Challinor et al., 2009). For other impacts, the AOGCM uncertainty is one of the most important contributors, but not dominant (e.g. Wilby et al., 2009; Buisson et al., 2010).

In summary, uncertainty in the climate response to emissions is likely to be the dominant source of uncertainty in projections of future crop yield. This necessitates the use of multi-AOGCMs, combined with multi-calibration approaches, to drive a range of crop models.

6.3. The importance of observations and other complications

The calibration methods described here require daily observations of $T_{\max}$, or any other climate variable of interest, over a relatively long period ($\approx$ 30 years) to enable a more precise estimation of the calibration method parameters. This could limit the regions to which these methods can be applied, but note the growing availability of daily observational datasets: e.g. HadGHCND (global, Caesar et al., 2006), E-OBS (Europe, Haylock et al., 2008) and APHRODITE (Asia, Xie et al., 2007).

One potential complication with these calibration methods is when other variables are considered. There is no reason why similar techniques cannot be applied to other climate variables required for crop modelling, such as mean temperature or solar radiation. Further work is required to explore the potential for extending the methodologies to perform simultaneous calibration to ensure covariance statistics are maintained. In addition, calibrating precipitation is more complicated, due to its positive definite nature.

Other complications in using these calibration methodologies could include: (i) defining reference and future periods with relatively small trends relative to the variability, and (ii) different shapes of distributions of the climate variable in the observations when compared to the AOGCMs (Ho, 2010). However, note that Huth et al. (2001) compared different downscaling methods with station data across Europe and concluded that none reproduced all the characteristics of the observations successfully.

Crop yield observations are also important in narrowing uncertainty in future yield projections. For example, the use of stronger observational constraints on crop responses to elevated CO_2 in Challinor et al. (2009) resulted in a significantly lower crop model uncertainty compared to Challinor et al. (2005a).

6.4. Key messages

This idealised study has examined calibration methods to produce AOGCM projections of climate variables relevant for crop modelling. The main findings are as follows:

1. Assessing the ability of calibration methods to produce realistic estimates of future climate (or equivalently, crop yields) is essential. The perfect model (or sibling) framework allows such an assessment and could be adopted more widely (e.g. Lobell and Burke, 2010), enabling a ranking of the various techniques available.
2. Change factor ('delta' type) approaches tend to be more robust than bias correction ('nudging') methods in the results presented here. Both methodologies tend to outperform using raw climate model output. However, this conclusion needs to be explored in a wider analysis of different regions, climate variables and AOGCM output.
3. The uncertainty due to the choice of calibration methodology is a significant contributor to total uncertainty in future crop yields.

In summary, the need to produce reliable and robust probabilistic projections of future crop yield necessitates the use of a wide range of AOGCMs, combined

with a variety of calibration approaches, whilst considering different crop modelling strategies (e.g., Palmer et al., 2005). This process will increase the burden of any yield assessment, and will likely also increase the uncertainty. This type of analysis is presently being extended to consider the use of historical observations for calibration and to examine the resulting uncertainty in crop yield projections.

Appendix A. Derivation of bias correction and change factor transfer functions

We now briefly illustrate the derivation of Eqns. 2, 4. More details are given in Ho (2010).

Take two distributions with the same shape but different means (M_1, M_2) and variances (σ_1^2, σ_2^2), for example, two normal distributions of random variables. It is clear that to map a point (X) on the two distributions,

$$\frac{X_1 - M_1}{\sigma_1} = \frac{X_2 - M_2}{\sigma_2}. \tag{A.1}$$

This is effectively equating the normalised difference between the point X and the mean in each distribution. Re-arranging, this gives,

$$X_2 = M_2 + \frac{\sigma_2}{\sigma_1}\left(X_1 - M_1\right), \tag{A.2}$$

which can be used to relate two distributions, X_1 and X_2, assuming their shape is the same. Note that these equations hold for non-normal distributions also.

Bias correction

In the BC methods, we use Eqn. A.2 to represent the relationship between present day simulations (T_{REF}, distribution 1) and present-day observations (O_{REF}, distribution 2), providing estimates of $M_{1,2}$ and $\sigma_{1,2}$. We then assume that the same relationship holds when relating future simulations (T_{RAW}) to future observations (O_{FUT}). Hence,

$$\widehat{O}_{\mathrm{FUT}} = \overline{O_{\mathrm{REF}}} + \frac{\sigma_{O,\mathrm{REF}}}{\sigma_{T,\mathrm{REF}}}\left(T_{\mathrm{RAW}} - \overline{T_{\mathrm{REF}}}\right), \tag{A.3}$$

equivalent to Eqn. 2.

Change Factor

In the CF methods, we use Eqn. A.2 to represent the relationship between present day simulations (T_{REF}, distribution 1) and future simulations (T_{RAW}, distribution 2), providing estimates of $M_{1,2}$ and $\sigma_{1,2}$. We then assume that the same relationship holds when relating present observations (O_{REF}) to future observations (O_{FUT}). Hence,

$$\widehat{O}_{\mathrm{FUT}} = \overline{T_{\mathrm{RAW}}} + \frac{\sigma_{T,\mathrm{RAW}}}{\sigma_{T,\mathrm{REF}}}\left(O_{\mathrm{REF}} - \overline{T_{\mathrm{REF}}}\right), \tag{A.4}$$

equivalent to Eqn. 4.

Acknowledgements

We thank David Stephenson and three anonymous reviewers for their valuable comments which helped improve the analysis, and the UK Met Office QUMP and IPSL modelling groups for making their daily data available to the community. The research leading to this paper has been supported by NCAS-Climate (EH, TMO, CKH), the NERC EQUIP project (EH, TMO, AJC) and by a NERC grant (CKH). AJC gratefully acknowledges the support of the CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS).

References

Arnell, N.W., Hudson, D.A., Jones, R.G., 2003. Climate change scenarios from a regional climate model: Estimating change in runoff in southern Africa. J. Geophys. Res. 108, 4519–4535.

Baigorria, G.A., Jones, J.W., 2010. GiST: A stochastic model for generating spatially and temporally correlated daily rainfall data. J. Climate 23, 5990–6008.

Baron, C., Sultan, B., Balme, M., Sarr, B., Traore, S., Lebel, T., Janicot, S., Dingkuhn, M., 2005. From GCM grid cell to agricultural plot: scale issues affecting modelling of climate impact. Phil. Trans. B. 360, 2095–2108.

Battisti, D.S., Naylor, R.L., 2009. Historical warnings of future food insecurity with unprecedented seasonal heat. Science 323, 240–244.

Buisson, L., Thuiller, W., Casajus, N., Lek, S., Grenouillet, G., 2010. Uncertainty in ensemble forecasting of species distribution. Glob. Change Bio. 16, 1145–1157.

Caesar, J., Alexander, L., Vose, R., 2006. Large-scale changes in observed daily maximum and minimum temperatures: Creation and analysis of a new gridded data set. J. Geophys. Res. 111, D05101.

Challinor, A., Wheeler, T., Slingo, J., Hemming, D., 2005a. Quantification of physical and biological uncertainty in the simulation of the yield of a tropical crop using present-day and doubled CO_2 climates. Phil. Trans. B. 360, 2085–2094.

Challinor, A.J., Slingo, J.M., Wheeler, T.R., Doblas-Reyes, F.J., 2005b. Probabilistic simulations of crop yield over western india using the DEMETER seasonal hindcast ensembles. Tellus A 57, 498–512.

Challinor, A.J., Wheeler, T., Hemming, D., Upadhyaya, H.D., 2009. Ensemble yield simulations: crop and climate uncertainties, sensitivity to temperature and genotypic adaptation to climate change. Climate Research 38, 117–127.

Challinor, A.J., Wheeler, T.R., Craufurd, P.Q., Ferro, C.A.T., Stephenson, D.B., 2007. Adaptation of crops to climate change through genotypic responses to mean and extreme temperatures. Agr. Eco. Environ. 119, 190 – 204.

Collins, M., Booth, B., Harris, G., Murphy, J., Sexton, D., Webb, M., 2006. Towards quantifying uncertainty in transient climate change. Climate Dynamics 27, 127–147.

Collins, M., Booth, B.B.B., Bhaskaran, B., Harris, G.R., Murphy, J.M., Sexton, D.M.H., Webb, M.J., 2011. Climate model errors, feedbacks and forcings: a comparison of perturbed physics and multi-model ensembles. Clim. Dyn. 36, 1737–1766.

Diaz-Nieto, J., Wilby, R.L., 2005. A comparison of statistical downscaling and climate change factor methods: impacts on low flows in the River Thames, United Kingdom. Climatic Change 69, 245–268.

de Ela, R., Laprise, R., Denis, B., 2002. Forecasting skill limits of nested, limited-area models: A perfect-model approach. Monthly Weather Review 130, 2006–2023.

Fischer, E., Schär, C., 2009. Future changes in daily summer temperature variability: driving processes and role for temperature extremes. Climate Dynamics 33, 917–935.

Gordon, C., Cooper, C., Senior, C.A., Banks, H., Gregory, J.M., Johns, T.C., Mitchell, J.F.B., Wood, R.A., 2000. The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. Climate Dyn. 16, 147–168.

Gosling, S., McGregor, G., Lowe, J., 2009. Climate change and heat-related mortality in six cities part 2: climate model evaluation and projected impacts from changes in the mean and variability of temperature with climate change. Int. J. of Biomet. 53, 31–51.

Hansen, J., Challinor, A.J., Ines, A., Wheeler, T.R., Moron, V., 2006. Translating climate forecasts into agricultural terms: advances and challenges. Climate Research 33, 27–41.

Hansen, J., Ines, A., 2005. Stochastic disaggregation of monthly rainfall data for crop simulation studies. Agr. For. Met. 131, 233–246.

Hawkins, E., 2011. Our evolving climate: communicating the effects of internal variability. Weather 66, 175–179.

Hawkins, E., Sutton, R., 2009. The potential to narrow uncertainty in regional climate predictions. Bull. Amer. Met. Soc. 90, 1095–1107.

Hawkins, E., Sutton, R., 2011. The potential to narrow uncertainty in projections of regional precipitation change. Clim. Dyn. 37, 407–418.

Haylock, M.R., Hofstra, N., Klein Tank, A.M.G., Klok, E.J., Jones, P.D., New, M., 2008. A European daily high-resolution gridded dataset of surface temperature and precipitation. J. Geophys. Res. 113, D20119.

Ho, C.K., 2010. Projecting extreme heat-related mortality in Europe under climate change. Ph.D. thesis. University of Exeter, UK.

Ho, C.K., Stephenson, D.B., Collins, M., Ferro, C.A.T., Brown, S.J., 2012. Calibration strategies: a source of additional uncertainty in climate change projections. Bull. Amer. Met. Soc. 93, 21–26.

Huntingford, C., Lambert, H.F., Gash, J.H.C., Taylor, C.M., Challinor, A.J., 2005. Aspects of climate change prediction relevant to crop productivity. Phil. Trans. B. 360, 1999–2009.

Huth, R., Kysely, J., Dubrovsky, M., 2001. Time structure of observed, GCM-simulated, downscaled, and stochastically generated daily temperature series. J. Climate 14, 4047–4061.

Ines, A.V.M., Hansen, J.W., 2006. Bias correction of daily GCM rainfall for crop simulation studies. Agr. For. Met. 138, 44 – 53.

Ines, A.V.M., Hansen, J.W., Robertson, A.W., 2011. Enhancing the utility of daily GCM rainfall for crop yield prediction. Int. J. Clim. in press.

Liu, L., Fritz, S., van Wesenbeeck, C.F.A., Fuchs, M., You, L., Obersteiner, M., Yang, H., 2008. A spatially explicit assessment of current and future hotspots of hunger in sub-saharan africa in the context of global change. Glob. Planet. Change 64, 222–235.

Lobell, D.B., Burke, M.B., 2010. On the use of statistical models to predict crop yield responses to climate change. Agricultural and Forest Meteorology 150, 1443 – 1452.

Marti, O., Braconnot, P., Dufresne, J.L., Bellier, J., Benshila, R., Bony, S., Brockmann, P., Cad-

ule, P., Caubel, A., Codron, F., de Noblet, N., Denvil, S., Fairhead, L., Fichefet, T., Foujols, M.A., Friedlingstein, P., Goosse, H., Grandpeix, J.Y., Guilyardi, E., Hourdin, F., Idelkadi, A., Kageyama, M., Krinner, G., Lévy, C., Madec, G., Mignot, J., Musat, I., Swingedouw, D., Talandier, C., 2010. Key features of the IPSL ocean atmosphere model and its sensitivity to atmospheric resolution. Climate Dynamics 34, 1–26.

Masson, D., Knutti, R., 2011. Climate model genealogy. Geophys. Res. Lett. 38, L08703.

Mearns, L.O., Rosenweig, C., Goldberg, R., 1997. Mean and variance change in climate scenarios: methods, agricultural applications, and measures of uncertainty. Clim. Change 35, 367–396.

Mearns, L.O., Rosenzweig, C., Goldberg, R., 1996. The effect of changes in daily and interannual climatic variability on CERES-Wheat: A sensitivity study. Climatic Change 32, 257–292.

Meehl, G.A., Stocker, T.F., Collins, W., Friedlingstein, P., Gaye, A.T., Gregory, J.M., Kitoh, A., Knutti, R., Murphy, J.M., Noda, A., Raper, S.C.B., Watterson, I.G., Weaver, A.J., Zhao, Z.C., 2007. Global climate projections. In: Climate Change 2007: The Physical Science Basis. Cambridge University Press, Cambridge, UK.

Monfreda, C., Ramankutty, N., Foley, J.A., 2008. Farming the planet: 2. geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. Glob. Biogeochem. Cyc. 22, GB1022.

Murphy, J.M., Sexton, D.M.H., Barnett, D.N., Jones, G.S., Webb, M.J., Collins, M., Stainforth, D.A., 2004. Quantification of modelling uncertainties in a large ensemble of climate change simulations. Nature 430, 768–772.

Osborne, T., Slingo, J., Lawrence, D., Wheeler, T., 2009. Examining the interaction of growing crops with local climate using a coupled crop-climate model. J. Climate 22, 1393–1411.

Osborne, T.M., Lawrence, D.M., Challinor, A.J., Slingo, J.M., Wheeler, T.R., 2007. Development and assessment of a coupled crop-climate model. Glob. Change Bio. 13, 169–183.

Osborne, T.M., Rose, G., Wheeler, T., 2012. Variation in the global-scale impacts of climate change on crop productivity due to climate model uncertainty and adaptation. Agr. For. Met. , this issue.

Palmer, T., Doblas-Reyes, F., Hagedorn, R., Weisheimer, A., 2005. Probabilistic prediction of climate using multi-model ensembles: from basics to applications. Phil. Trans. B. 360, 1991–1998.

Porter, J.R., Semenov, M.A., 2005. Crop responses to climatic variation. Phil. Trans. B. 360, 2021–2035.

Poulter, B., Hattermann, F., Hawkins, E., Zaehle, S., Sitch, S., Restrepo-Coupe, N., Heyder, U., Cramer, W., 2010. Robust dynamics of Amazon dieback to climate change with perturbed ecosystem model parameters. Glob. Change Bio. 16, 2476–2495.

Räisänen, J., Palmer, T.N., 2001. A probability and decision-model analysis of a multimodel ensemble of climate change simulations. Journal of Climate 14, 3212–3226.

Schlenker, W., Roberts, M.J., 2009. Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. PNAS 106, 15594–15598.

Semenov, M.A., Donatelli, M., Stratonovitch, P., Chatzidaki, E., Baruth, B., 2010. ELPIS: a dataset of local-scale daily climate scenarios for Europe. Climate Research 44, 3–15.

Semenov, M.A., Porter, J.R., 1995. Climatic variability and the modelling of crop yields. Agr. For. Met. 73, 265 – 283.

Teixeira, E.I., Fischera, G., van Velthuizena, H., Walter, C., Ewert, F., 2012. Global hot-spots of heat stress on agricultural crops due to climate change. Agr. For. Met. this issue, in press.

Thornton, P.K., Jones, P.G., Ericksen, P.J., Challinor, A.J., 2011. Agriculture and food systems in sub-saharan Africa in a 4°C+ world. Phil Trans 369, 117–136.

USDA, 2012. Global crop production analysis , http://www.pecad.fas.usda.gov.

Vara Prasad, P.V., Craufurd, P.Q., Summerfield, R.J., Wheeler, T.R., 2000. Effects of short episodes of heat stress on flower production and fruit-set of groundnut (Arachis hypogaea L.). J. Exp. Botany 51, 777–784.

Wilby, R.L., Charles, S.P., Zorita, E., Timbal, B., Whetton, P., Mearns, L.O., 2004. Guidelines for use of climate scenarios developed from statistical downscaling methods. Supporting Material. Intergovernmental Panel on Climate Change.

Wilby, R.L., Tronim, J., Biot, Y., Tedd, L., Hewitson, B.C., Smith, D.M., Sutton, R.T., 2009. A review of climate risk information for adaptation and development planning. Int. J. Clim. 29, 1193–1215.

Xie, P., Yatagai, A., Chen, M., Hayasaka, T., Fukushima, Y., Liu, C., Yang, S., 2007. A gauge-based analysis of daily precipitation over East Asia. J. Hydromet. 8, 607–627.

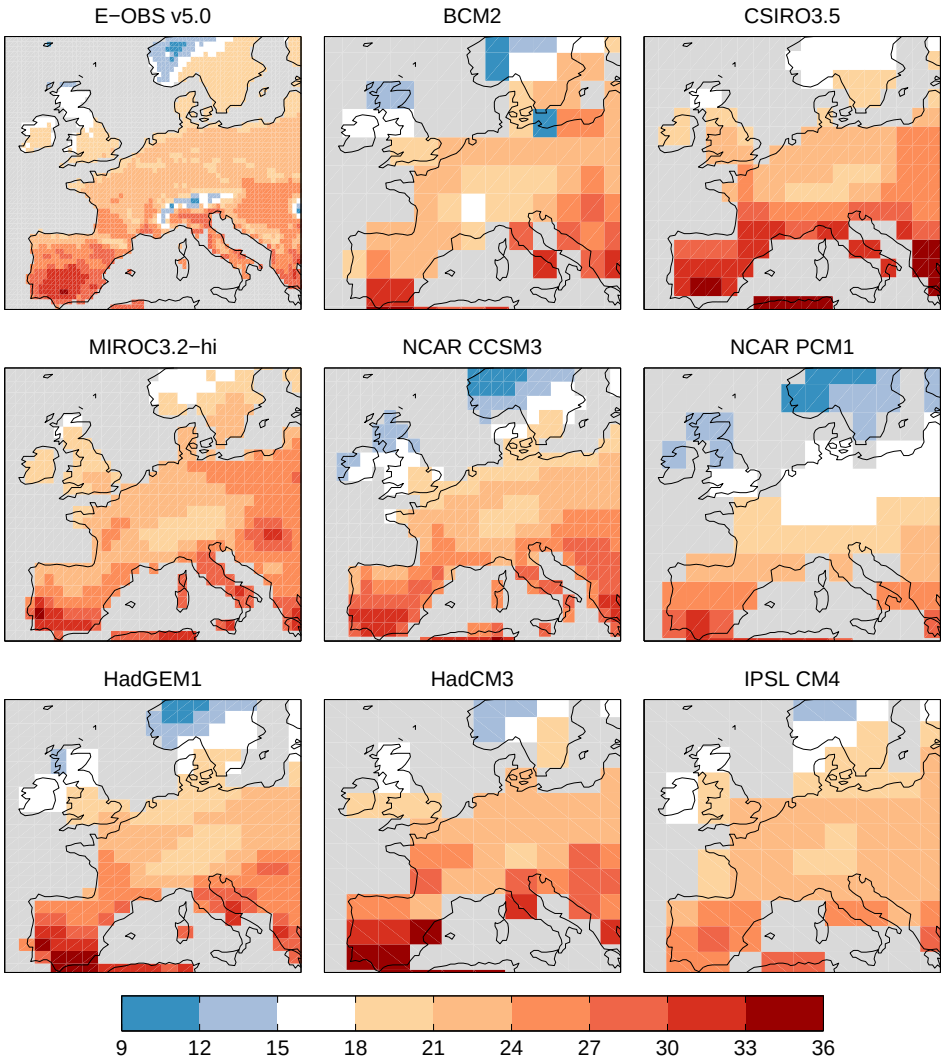
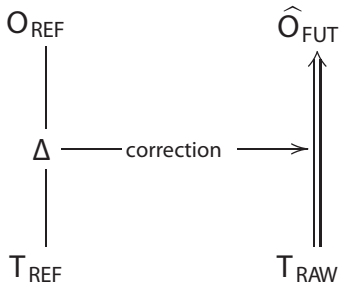


Figure 1: Mean summer (JJA) $T_{\max}$ for the reference period 1970-1999 from observations (E-OBS v5.0 0.5°, Haylock et al., 2008) and a range of AOGCMs in the CMIP3 database as labelled. For the AOGCMs, only grid cells with a land portion of larger than 40% are shown. The units are °C.

(a) Bias correction (SH and BC)



(b) Change factor (DEL and CF)

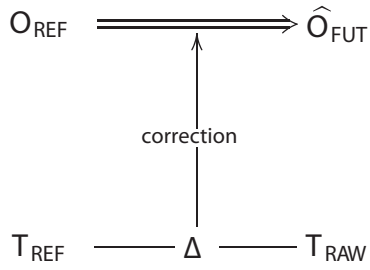


Figure 2: Schematic of the two general types of calibration. (a) Bias correction uses raw model output and corrects it using the differences (Δ) between reference data from the model and observations. If no correction is used then this is the RAW method. (b) Change factor uses present day observations, corrected using the differences between present and future model data. The corrections considered here include changes in only the mean (SH and DEL) or mean and variance together (BC and CF).

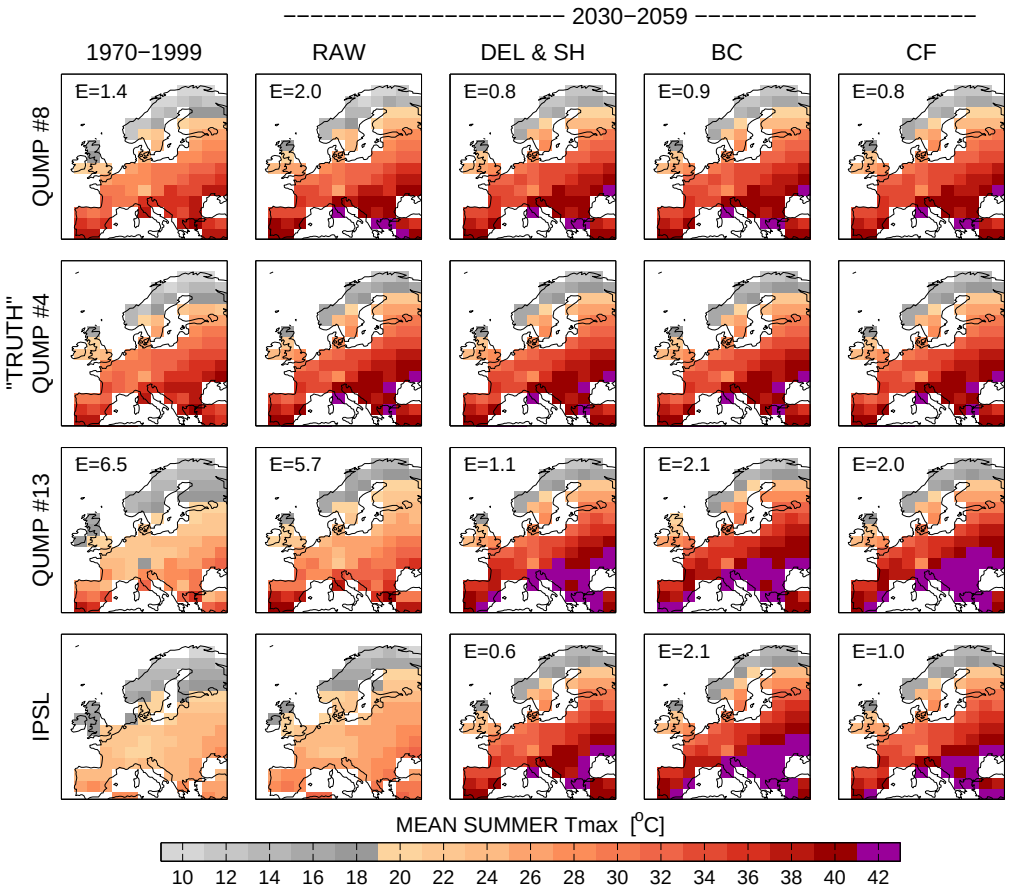


Figure 3: Demonstrating the calibration methodologies using a range of AOGCM simulations for mean summer $T_{\max}$. QUMP4 is selected to act as 'truth' for verification against the calibrated projections using other QUMP members (#8, #13) and the IPSL data. The RMS error for the region shown is given as the E value. Columns (from left to right) represent T_{REF}, T_{RAW}, T_{DEL} & T_{SH}, T_{BC}, and T_{CF}.

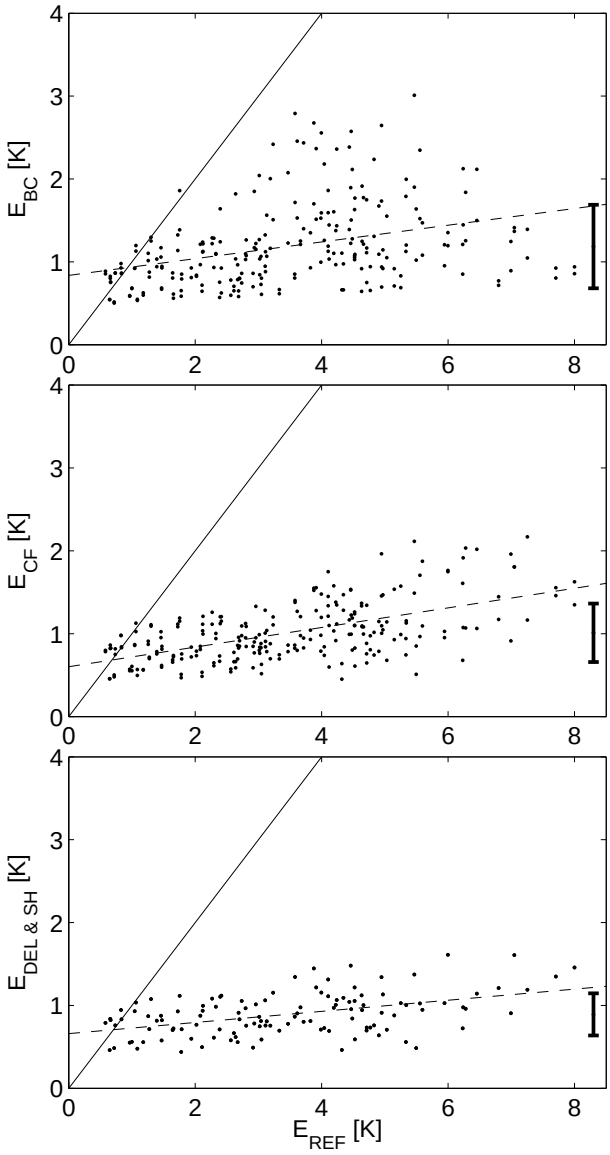


Figure 4: The relationship between model bias and calibration ability for mean JJA $T_{\max}$. Shown is the RMS error (E) for each calibration method (rows) for each of the 240 combinations of QUMP ensemble pairs (dots), as a function of the RMS error (E_{REF}) in the reference period, which is a measure of the model bias. The solid line shows equal errors in calibrated and reference periods. The dashed line is the regression of the QUMP ensemble members; the slopes of which are positive suggesting that the smaller the model bias, the smaller the error in calibrated mean temperatures. 95% of the dots lie to the right of the solid lines demonstrating that calibration has improved the projection.

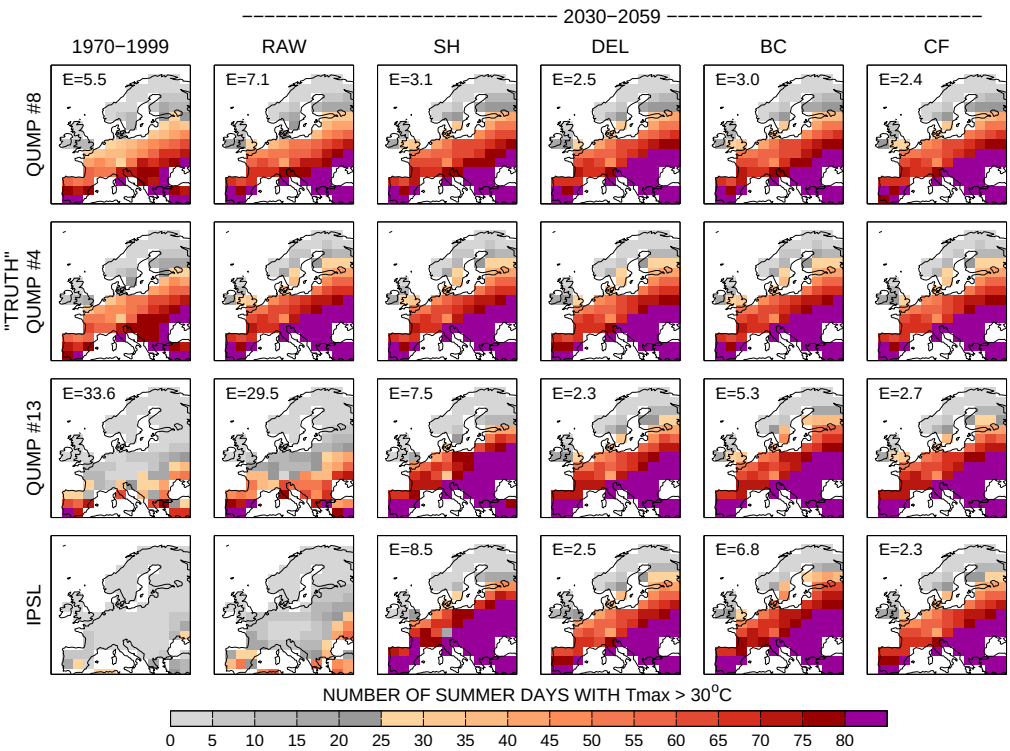


Figure 5: Demonstrating the calibration methodology using a range of AOGCM simulations for the number of summer days with $T_{\max} > 30°$C. QUMP4 is selected to act as 'truth' for verification against the calibrated projections using other QUMP members (#8, #13) and the IPSL data. The RMS error for the region shown is given as the E value. Columns (from left to right) represent T_{REF}, T_{RAW}, T_{SH}, T_{DEL}, T_{BC}, and T_{CF}.

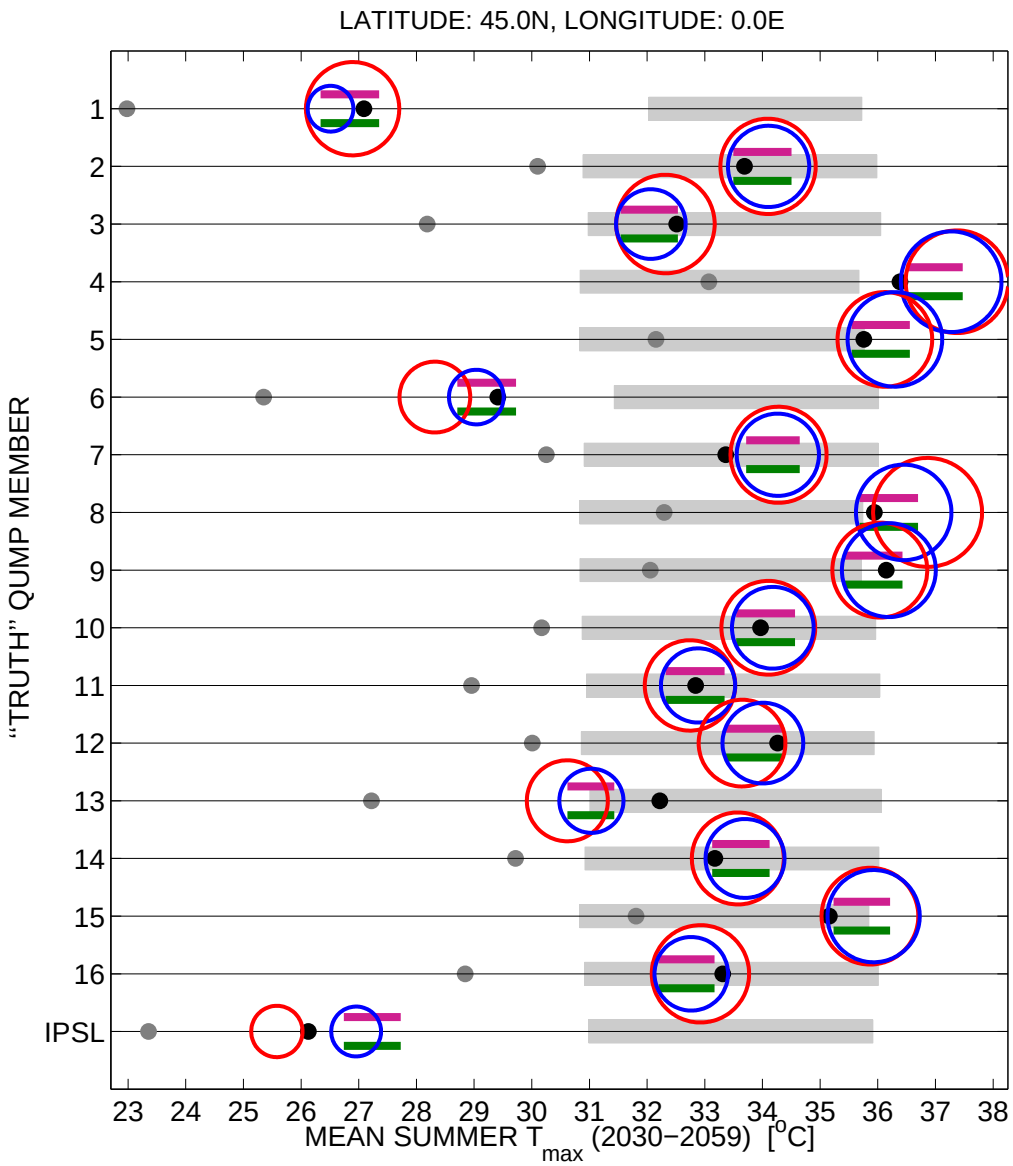


Figure 6: Demonstrating the calibration methodology for a single grid point. Rows 1-16: Different projections of mean summer $T_{\max}$, calibrated using selected QUMP member. Bottom row: Different projections of mean summer $T_{\max}$, calibrated using data from IPSL CM4. Light grey bars: Raw output of the independent QUMP simulations. Circles represent projections using BC (red) and CF (blue) methodologies. Green and Purple bars represent SH and DEL methodologies respectively. All show mean $\pm 1\sigma$. Dark grey and black dots represent 'truth' from reference and future periods respectively. The black dot therefore represents the target for calibration.

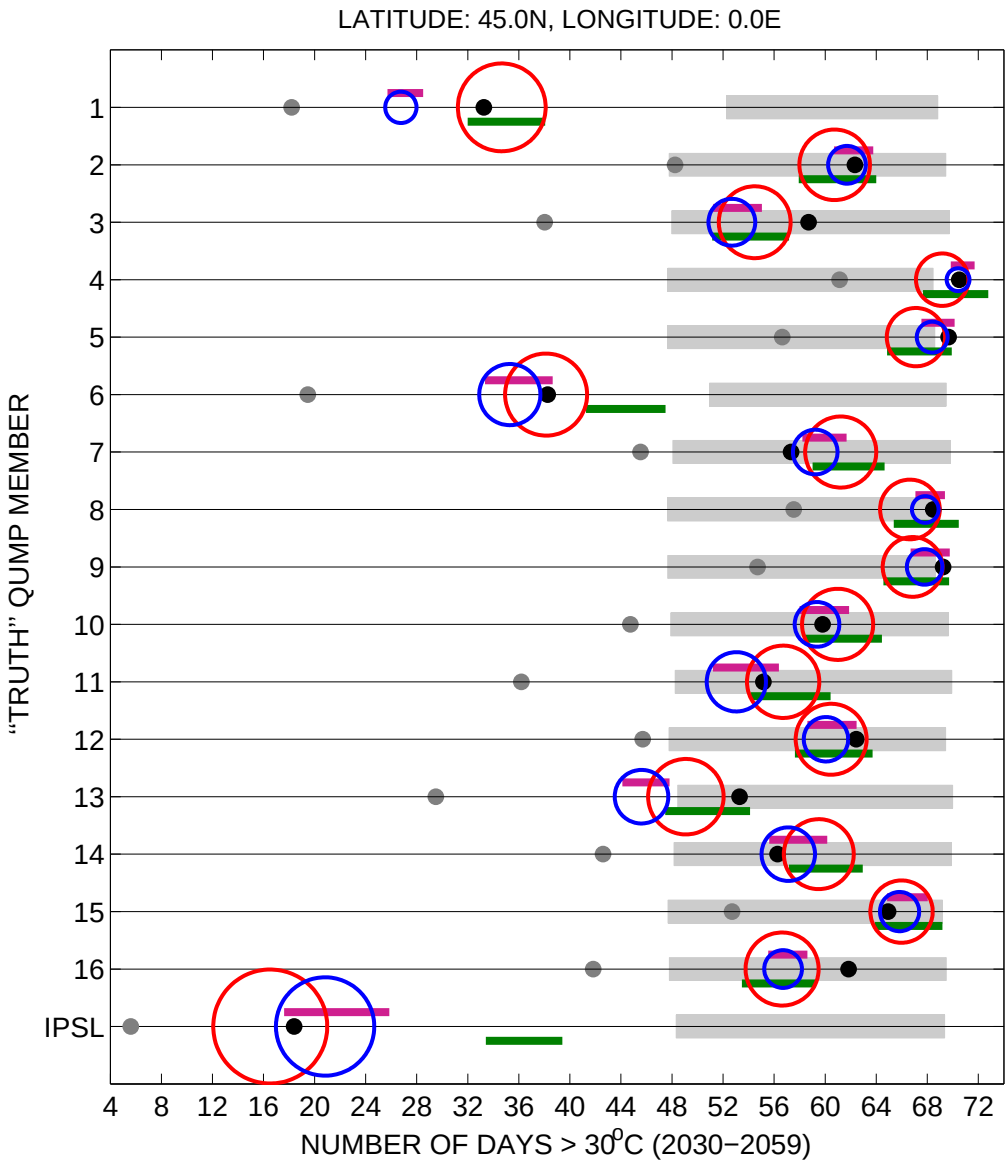


Figure 7: As Fig. 6 for the number of days over 30°C.

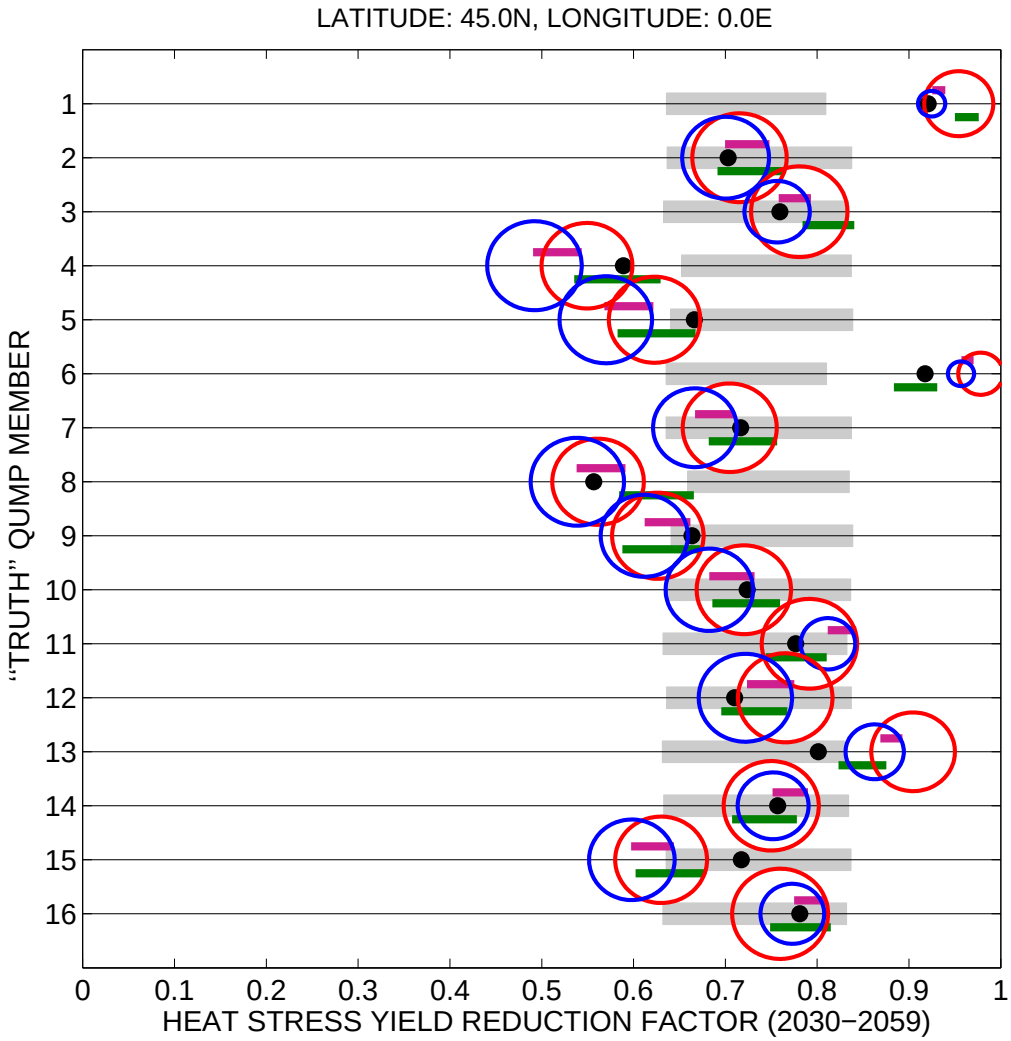


Figure 8: As Fig. 6 for the heat stress index (Eqn. 7)

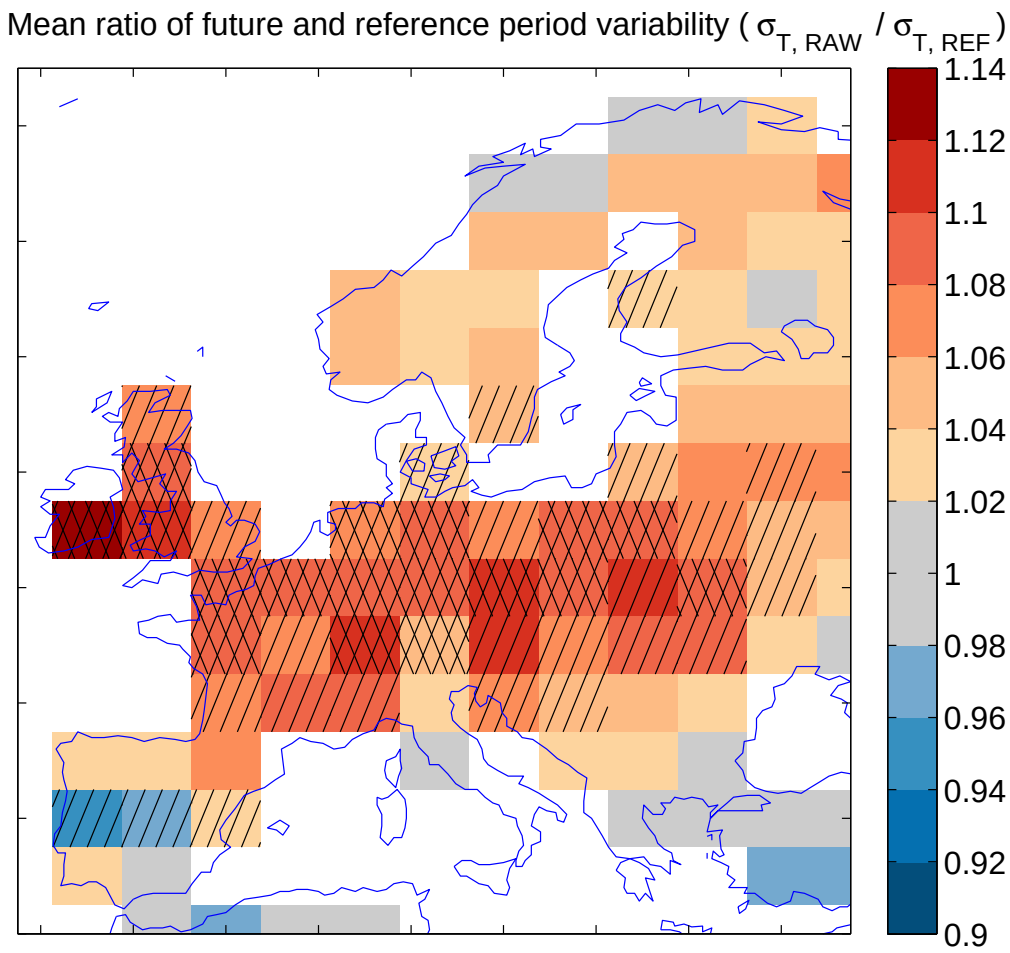


Figure 9: The mean fractional change in $T_{\max}$ daily variability for the 16 members of the QUMP ensemble. Red areas denote an increase in variability, and blue a decrease. The diagonal hatched areas denote regions where 12 or more ensemble members agree on the sign of the change, and the cross-hatched areas denote 15 or more members.

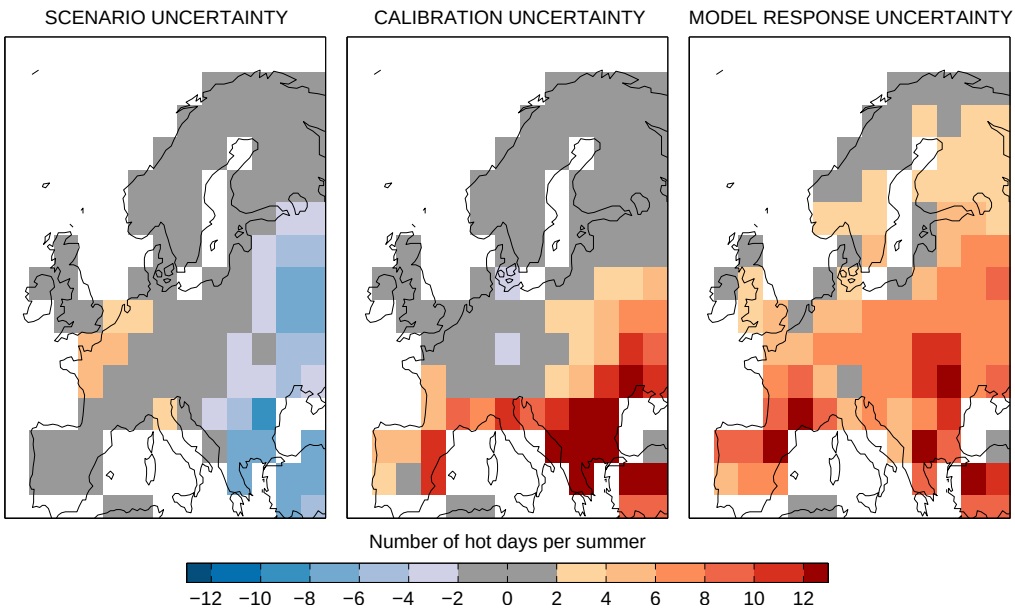


Figure 10: Uncertainties in the number of summer days in 2030-2059 with $T_{\max} > 30°C$. (left) Difference between the raw output of the IPSL AOGCM from two different emissions scenarios (SRES A1B and A2). (middle) Difference between the mean BC and CF calibrated projections using the QUMP ensemble to predict the IPSL AOGCM data. (right) 2× the standard deviation in the BC calibrated QUMP ensemble, predicting the IPSL AOGCM data.