JOURNAL OF GEOPHYSICAL RESEARCH, VOL. ???, XXXX, DOI:10.1029/,

Detection and Prediction of mean and extreme European summer temperatures with a multi-model ensemble

H. M. Hanlon,[1] S. Morak,[2] G. C. Hegerl,[3]

H. M. Hanlon, Met Office Hadley Centre, Exeter, UK. (helen.hanlon@metoffice.gov.uk)

S. Morak, Department of Meteorology, University of Reading, Reading, UK. (S.Morak@reading.ed.ac.uk)

G. C. Hegerl, School of Geosciences, University of Edinburgh, Edinburgh, UK. (gabi.hegerl@ed.ac.uk)

[1]Met Office Hadley Centre, Exeter, UK

[2]Department of Meteorology, University of Reading, Reading, UK.

[3]School of Geosciences, University of Edinburgh, Edinburgh, UK

Abstract. We analyse observed mean to extreme summer temperature indices across Europe in order to determine whether there is evidence for a detectable climate change signal, and whether these indices show evidence for predictability.

Observations from ENSEMBLES project observational database version 6 (E-OBS) from 1960-2011 are compared with the model simulations from the global coupled climate models CanCM4, HadCM3, MIROC5 and MPI-ESM-LR, as published on the CMIP5 archive. Indices are examined that span a moderate to extreme range of the summer temperature distribution by including the summer average, the hottest 5-day average and the hottest daily maximum and daily minimum temperature during summer. The region of interest is Europe, however a number of sub-regions are also studied, which include: Western Europe, the British Isles, the Mediterranean and Central Europe.

The observed changes in the analysed indices are well represented by the multi-model mean and are within the range of the multi-model ensemble for most regions, with the exception of 1-day and 5-day average daily maximum temperature extremes across the UK. Observed changes are detectable against estimates of internal climate variability for both moderate and extreme temperature indices across all regions, in almost all cases. Exceptions are the hottest 5-day average daily maximum temperature in the UK and Central Europe, for which results are not conclusive.

An analysis of the skill in decadal hindcasts of these indices shows that there is significant prediction skill across these indices for 3 of the 4 models for some regions and some models. This skill exceeds the skill of forecasts based on observed climatology and random noise and is largely due to external forcing. However, there is some evidence that there is additional skill originating from the assimilation of observations into the initialisation in some cases.

1. Introduction

Recent years have seen two of the most devastating extreme heatwave events in Eurasia, the 2003 European Heatwave (*Schär et al.* [2004], *Fink et al.* [2004] and *Hanlon et al.* [2010]) and 2010 Russian heatwave (*Barriopedro et al.* [2011], *Dole et al.* [2011], *Rahmstorf et al.* [2011], *Otto et al.* [2012]). The European summer heatwave of 2003 exhibited anomalously hot temperatures, with the European continental mean summer average temperature exceeding the long term mean (1961-90) by 3 °C (equivalent to more than 5 standard deviations), as shown by *Schär et al.* [2004]. *Schär et al.* [2004] indicated this could be due to a shift in mean summer temperatures, combined with an increase in variability. Subsequent studies have shown there were additional meteorological factors and land surface interactions influencing the 2003 event (*Hanlon et al.* [2010], *Fischer et al.* [2007]).

These extreme heatwave events had a severe impact on society and nature, in particular the impact on human health was profound. For human health, increases in daily extreme temperatures are more damaging than changes in seasonal mean temperatures (*Diaz et al.* [2006], *Fouillet et al.* [2006], *Grize et al.* [2005] and *Pascal et al.* [2006]).

In order to determine whether the frequency and intensity of extreme events are affected by anthropogenic influences, which include increased emission of greenhouse gases, several studies have performed detection or combined detection and attribution analyses for changes in the frequency or intensity of extremes. Such analyses aim to determine the cause of an observed change in the temperature distribution. A significant change is detected if the likelihood of this change occuring, due to internal variability alone, is evaluated to be small (*Hegerl et al.* [2007], *Hegerl et al.* [2010]), while attribution analyses evaluate several potential explanations for an observed, generally detectable, change and determine the most likely explanation. Results from recent studies show evidence for human influence on the upward trend in frequency and intensity of temperature extremes (e.g., *Christidis et al.* [2012], *Morak et al.* [2011], *Morak et al.* [2013], *Zwiers et al.* [2011]), consistent with the finding that annual and summer average temperatures over many regions are influenced by greenhouse gas increases (*Christidis et al.* [2012]; *Stott et al.* [2010]).

Even changes in the probability of individual extreme events have been attributed in part to external forcing: in an attribution study of the 2003 European heatwave by *Stott et al.* [2004] it was found, with a high probability, that the risk of the event had at least doubled due to anthropogenic influences. Attribution studies have also been performed for the 2010 Russian heatwave event, however, there is seemingly conflicting conclusions over the extent to which anthropogenic factors contributed to the cause of the event in studies of *Dole et al.* [2011] and *Rahmstorf et al.* [2011]). *Otto et al.* [2012] show that the probability of such an event changed significantly due to human influences, while most of the observed extreme anomaly originated from unusual weather (as shown by *Dole et al.* [2011]), thereby explaining that the *Rahmstorf et al.* [2011] and *Dole et al.* [2011]) conclusions were not mutually exclusive.

Does the detectable influence of forcing, possibly combined with initial conditions, enable near-term prediction of changes in the intensity of extremes? For predictions of the near term future (10-20 years ahead) we look to decadal prediction models. A recent study by *Eade et al.* [2012] demonstrated skillful predictions of moderate (1 in 10) daily temperature extremes on decadal timescales using the Met Office Hadley Centre decadal prediction system (DePreSys). These are initialised decadal predictions which attempt to provide improved predictions of natural internal variability (*Smith et al.* [2010]). *Hanlon et al.* [2013] has shown there is skill in predicting the summer average and hottest 5-day average Tmax and Tmin in Europe, also with DePreSys, where this skill is mostly due to model forcing rather than initialisation of observations.

In this paper we determine if external forcing has significantly changed the intensity of summer mean and extreme temperatures, and if such a change leads to predictable changes in the near-term. We expand on the work performed in *Morak et al.* [2013], a single model detection study, which found detectable changes in the *frequency* of hot daytime and nighttime temperatures during summer on the global scale but also for smaller regions such as Europe (*Morak et al.* [2011]) and, for the number of warm nights, for Central Europe (*Morak et al.* [2011]). In this study we perform a multi-model detection analysis using indices for the *intensity* of summer extreme temperatures across Europe and for smaller European regions. Alongside this detection analysis we will also consider the skill in prediction of these summer heatwave indices with a number of CMIP5 decadal prediction models by expanding the work undertaken by *Hanlon et al.* [2013]. This will include a comparison of decadal prediction skill to that obtained with the CMIP5 historical simulations to determine whether there is added skill in the decadal predictions due to initialisation of these models with observed values.

Section 2 of this paper introduces observations and models used in the study, with methods for both detection and prediction introduced in Section 3. Section 4 shows results which are discussed in Section 5.

2. Data

This study uses gridded observed and model simulated data sets of mean daily minimum and maximum temperature. The analysis period is 1961-2005. Seasons of interest are the summer half-year April-September and the summer season June-August. The regions considered include: Europe (EU) (35 - 65°N latitude,12°W - 40°E longitude), along with sub-regions Western Europe (WEU, 34-61°N latitude, 12°W - 26°E longitude)), UK and Ireland (UK, 50 - 60°N latitude, 12°W - 2°E longitude), Mediterranean (MED, 35 - 50°N latitude, 12°W - 40°E longitude) and Central Europe (CEU, 42 - 55°N latitude, 2°W - 20°E longitude). For a graphical representation of the spatial extent of these regions see Figure 1.

2.1. Observations

The observed data originates from the ENSEMBLES project observational database (E-OBS), which is a high resolution (0.5° latitude by 0.5° longitude grid) gridded dataset of observations (see *Haylock et al.* [2008] for more details). The data set is based on observations of individual stations which have been interpolated on a regular grid. The data density is high with only small amounts of missing data earlier on in the record. In this study we use the data sets of daily minimum and maximum temperature from E-OBS version 6, which spans the period 1950-2011.

2.2. Models

All model simulated data sets of daily minimum and maximum temperature were retrieved from the CMIP5 archive (*Taylor et al.* [2012]). This work uses data from the historical simulations as well as from the decadal predictions. The models chosen for the analysis were CanCM4, HadCM3, MIROC5 and MPI-ESM-LR, as these models provided daily minimum and maximum surface temperature data from the historical and decadal simulations in time for our analysis. The use of four models provides multi-model information that is much more robust than the use of single models which is often applied in detection studies for extremes (*Morak et al.* [2013] and *Christidis et al.* [2005]). For model description see Table 1.

The forcing of the historical runs includes anthropogenic forcing, such as the observed concentrations of green-house gases, aerosols, generally direct as well as indirect, and natural forcing such as the recorded changes in volcanic aerosol or changes in solar activity for the 20th century. The historical simulations span the period 1850 to 2005 and consist of 27 simulations from across the four global coupled climate models. The 27 single model runs are distributed as follows: CanCM4 (10 ensemble members), HadCM3 (10 ensemble members), MIROC5 (4 ensemble members) and MPI-ESM-LR (3 ensemble members).

The decadal simulations consist of a set of runs, each 10 years in length starting at 5-year intervals, which are forced in the same way as the historical runs, but initialised from observations (*Meehl et al.* [2009]). The start times are 1 January 1961, 1966, 1971, 1976, 1981,1986, 1991, 1996, 2001 and 2006. For each model there is an ensemble of decadal simulations CanCM4 (10 ensemble members), HadCM3 (10 ensemble members), MIROC5 (6 ensemble members) and MPI-ESM-LR (10 ensemble members).

3. Methodology

3.1. Indices Computation and Processing

The following six indices have been computed and analysed throughout this study:

- Summer average minimum temperature: The mean average daily minimum temperature computed over the summer season June-August.
- Summer average maximum temperature: The mean average daily maximum temperature computed over the summer season June-August.
- Max1-day Tmin: The highest daily minimum temperature that occurred between 1st of April and 30th of September.
- Max1-day Tmax: The highest daily maximum temperature that occurred between 1st of April and 30th of September.
- Max5-day Tmin: The highest 5-day rolling mean average daily minimum temperature that occurred between 1st of April and 30th of September.
- Max5-day Tmax: The highest 5-day rolling mean average daily maximum temperature that occurred between 1st of April and 30th of September.

The indices were computed for the observations and the model runs (both historical and decadal simulations) on their respective grids, which were then re-gridded using nearest neighbour interpolation to the grid of HadCM3, which is the coarsest grid of all data sets (3.75° longitude x 2.5° latitude). Following this, the model data sets were masked in time and space in order to match the observations. Next the spatial average of the indices was computed for the regions of interest, as both skill score analysis and detection analysis is performed on time series of regional means. The anomalies of the resulting time series were calculated relative to the entire period (1961-2005) for the detection analysis. In contrast, for the skill analysis, a bias correction was applied to absolute values (see Section 3.3). Finally, the 5-year average of each time series was computed in order to reduce the effect of inter-annual variability. The multi-model mean time series was computed by averaging over all multi-model ensemble members.

3.2. Detection Analysis

The detection analysis aims to determine whether an observed change can be explained solely due to internal variability or whether a combination of external forcing and variability explains this change. In a methodology, introduced by *Hasselmann* [1993] with further improvements by *Allen and Tett* [1999]) and *Allen and Stott* [2003], the relationship between observations and model simulated indices is expressed as:

$$Y = \alpha(X - \nu_1) + \nu_0 \tag{1}$$

where Y stands for the time series of the observations (here, one of the time series of regionally averaged indices over Europe), α for the scaling factor, X represents the multi-model mean time series for the corresponding index, ν_1 stands for a realisation of the model internal variability and ν_0 for a realisation of the observed variability.

Using this method, we obtain scaling factors α, which are the factors by which the fingerprints (here we have used ‘non-optimised’ fingerprints) are to be scaled in order to best match the observations. Much of the detection and attribution literature uses a metric that improves the signal to noise ratio (see discussion of optimised fingerprints in *Hegerl et al.* [2007]), this has not been done here as previous work showed that the improvement for detection of changes in temperature extremes is limited *Morak et al.* [2013]. The scaling factors have been determined by a total least squares fit (*Allen and Stott* [2003]) of the 5-year average time series of the modelled index, in the form of anomalies from the

1961-2005 climatology, to that of observations. The uncertainty in α has been computed by adding an appropriate estimate of noise onto both the fingerprint and the observations and repeating the scaling factor calculations. The noise estimate that is added to the fingerprint is divided by the ensemble size in order to account for the reduction in noise due to averaging across the ensemble (see *Allen and Stott* [2003]).

The samples of internal variability (noise) are obtained from the model simulated variability of each individual model run after subtracting the multi-model mean change. The variance around a sample mean from a small ensemble of n simulations leads to a low bias in variance, which we have corrected for by multiplying the variance by a factor of $\sqrt{\frac{n}{n-1}}$ (see *Von Storch and Zwiers* [2000]), where n is the total number of historical simulations (27). Thus we arrive at 27 realizations of internal climate variability that have a similar space-time autocorrelation structure as the variability simulated within the individual climate models. Using these samples, which estimate the internal variability, the uncertainty is calculated, along with the fifth and ninety-fifth percentile of the scaling factors.

Finally, the regression residual has been compared with the noise samples used in the analysis. The detection result is only considered to be robust if the residual variability in the observations after subtracting the fitted signal ν_0 is within the distribution (we chose the central 80th percentile) of the model internal variability. Where the scaling factor calculated from the analysis is significantly different from 0, the fingerprint is detected, and where it is consistent with 1, given its uncertainty, this indicates that the multi-model mean is statistically consistent with the observations.

3.3. Prediction Analysis

The indices detailed in Section 3.1 are also computed for the decadal hindcasts, which are model simulations initialised with observations with start dates between 1961 and 2001 (inclusive), as described in Section 2.2. For this part of the analysis we use the absolute values of the indices rather than anomalies from climatology. Hence these indices exhibit some considerable biases when compared to the observations. To account for this, the mean bias between the modelled index (x) and the observed index (y) averaged over 1961-2000 is computed and removed for each member (m) of each separate model ensemble, by:

$$x_{i,t,m=m*} = x_{i,t,m=m*} - \frac{1}{10}\sum_{i=0}^{9}\frac{\sum_{m \notin m*} x_{i,t=0,m}}{n-1} + \frac{1}{40}\sum_{yr=1961}^{2000} y_{yr} \tag{2}$$

Where m is a set of all ensemble members, $m*$ is each individual ensemble member, ($m* = 1$ to n) where n is the number of members. i corresponds to each of the ten year runs started every five years starting with 1961 and t is the lead-time for each run e.g.: $i = 0$, $t = 0$ relates to summer 1961, $i = 0$, $t = 3$ is summer 1964 from the first run started in 1961 then $i = 1$, $t = 0$ is summer 1966, the first summer in the run started in 1965 and so on. The mean modelled index is taken as the averaged over the index computed at leadtime zero ($t = 0$, the index value for the first summer) over all ensemble members but the member being corrected (this is sometimes described as "leave one out"), then following that averaged over each run i. following which the ensemble mean of this is taken. However, in order to avoid over-correcting, the member being corrected is left out of this average when the ensemble average of the mean modelled index is calculated. Hence the correction applied across each ten year run remains constant across different leadtimes within that run and as such does not account for drift in the model at later lead-times. To perform the correction the mean modelled index is subtracted from the modelled index for each member individually, then then mean observed index (averaged over all years between 1961-2000) is added on. The historical runs have been bias corrected with exactly the same method prior to the skill score analysis. The model drift has not been corrected due to the limited sample size. Ideally, the correction should be calculated with data outside the time period of the sample being tested to allow for the correction to be applied to the future model data which could then be used to make a prediction (see e.g., *Hanlon et al.* [2013]). However, due to limited sample size an out-of-sample correction procedure was not possible with this set of models.

After this is calculated for each model separately, the multimodel mean is taken as the mean average of the ensemble mean of each set of model simulations after bias correction. No model weighting is used in the computation of the multi-model mean, but since HadCM3 and MPI-ESM-LR consist of a larger number of simulations, they may indirectly be weighted slightly higher and contribute more to the multi-model mean.

When considering how useful or significant a forecast is, it needs to be compared against alternative information which could be used to make a prediction, otherwise referred to as a reference forecast. Where a modelled forecast is closer to the observation than the alternative method of prediction (eg. observed climatology) the model is described as being more skillful than the alternative. Following *Hanlon et al.* [2013] we use the Mean Square Skill Score (MSSS) (see *Murphy* [1988], *Goddard et al.* [2012]) to estimate how accurately the model hindcasts recreate the corresponding observed values, compared to E-OBS observational climatology. It compares the mean square errors between each bias corrected forecast with the observations.

$$MSE(\mathbf{x}_t, y_t) = \frac{1}{10}\sum_{i=0}^{9}(\mathbf{x}_{i,t} - y_{i,t})^2 \tag{3}$$

$$MSSS(\mathbf{x}_t, y_t, r) = 1 - \frac{MSE(\mathbf{x}_t, y_t)}{MSE(r, y_t)} \tag{4}$$

Where MSE denotes the mean square error calculated across all ten year runs at individual lead-times($\mathbf{x}_t$) compared to corresponding observations ($\mathbf{y}_t$), $\mathbf{x}_{i,t} = \sum_{m=0}^{n} x_{i,t,m}$ is the ith ensemble mean decadal forecast at leadtime t, y_i is the ith observed value corresponding to the same year as leadtime t, r is the corresponding reference forecast. As the decadal simulations $\mathbf{x}_t$ consist of 10-year long runs started every 5 years, there are 10 decadal forecasts spanning the period 1961-2005 for each member of each model individually, which can be used for this calculation.

A skillful prediction is considered to be a forecast that is closer to the observed value than our reference forecast r. Here, the reference forecast r, observed climatology, is calculated by taking the mean average of the index considered over the observed values between 1961-2000 (as outlined in Section 3.1).

This skill score analysis is repeated using the ensemble mean of the historical runs as the reference forecast (as in equation 5), where the years selected from the historical runs are the same as those simulated by the decadal runs. The historical runs used here are the same ones used for the detection analysis (Section 3.2). This determines whether the initialised decadal runs (x) are more skillful than the unassimilated historical runs (h). The reason for computing the difference in skill with this method, as opposed to a simpler method such as subtracting the MSSS for the historical simulation from the

MSSS for the initialized forecasts, is that the method used here removes the dependence on the skill of the comparison to observed climatology. Instead, the mean squared errors for the two sets of modelled results are compared directly; using the MSSS exactly as it was designed to compare the skill of two forecasting methods. The difference in skill between the two ensembles shows how much more skill the ensemble has that assimilates observations over the unassimilated runs which have no initial knowledge of the observed state of the climate.

$$MSSS(\mathbf{x}_t, \mathbf{h}_t, y_t) = 1 - \frac{MSE(\mathbf{x}_t, y_t)}{MSE(\mathbf{h}_t, y_t)} \tag{5}$$

where $MSE(\mathbf{h}_t, y_t) = \frac{1}{10}\sum_{i=0}^{9}(\mathbf{h}_{i,t} - y_{i,t})^2$ and $\mathbf{h}_{i,t} = \sum_{m=0}^{n_{hist}} h_{i,t,m}$ is the ensemble mean historical simulation corresponding to times for the ith forecast at leadtime t and n_{hist} is the number of historical ensemble members.

The MSSSs (Equations 4 and 5) are calculated for 5-year and 10-year averages of the annual indices because *Hanlon et al.* [2013] showed that skill is larger for these than for annual indices, for which the skill was not significant due to a larger influence of weather noise compared to possibly predictable interdecadal variability and role of forcing.

The MSSS is computed from the ensemble average of the regionally averaged index at each leadtime for a particular run. Sampling uncertainty arises from the limited ensemble size, which is estimated using bootstrapping with-replacement across each ensemble (see *Efron and Tibshirani* [1993, Chapter 6]). For each realisation, all members of the ensemble are drawn at random with replacement, from the entire ensemble. Then the same MSSS computations are performed on the bootstrapped sample as applied to the ensemble average. This generates a thousand realisations of the MSSS and the 10-90%

range from these provide the uncertainty on the MSSS. If the score is significantly above zero then the forecast has more skill in predicting the index than the reference forecast, for example, the in-sample observed climatology or the uninitialised historical simulations.

An additional method of estimating uncertainty is to compare a random forecast, which should have no significant skill, to the observed climatology. A random forecast is generated assuming a normal distribution for each decadal hindcast index (annual, 5-year average and 10-year average) and member. The mean and standard deviation for the normal distribution is estimated from each member of decadal hindcasts separately and used to normalise the random forecast. 1000 random forecast realisations are generated and a distribution of MSSSs is computed from these. The 90th percentile of this distribution is taken as a cut off point, below which the MSSSs for the decadal hindcasts are considered not significantly better than random noise.

4. Results

The time series of the indices of mean and extreme summer temperatures show clear increases in the magnitude of hot extremes during summer for most regions. These increases are notable since the early 1980s, which follows a period of negligible or even negative changes (refer to Figure 2 to see this in the time series for the Europe region). This change can be seen in the moderate extremes (summer average minimum and maximum temperature), as well as in the 1-day and 5-day extremes. The observed change is well represented by the multi-model mean of the historical and decadal simulations, mostly lying within the range of the individual ensemble members. Both initialized and non-initialized forecasts also show visible small decreases in averaged temperature following the volcanic eruptions of 1982 and 1991, while the observations appear to show a less clear drop in temperature as expected from a single realization of observed climate that is more influenced by weather noise than the ensemble average forecast. The magnitude of the observed changes for Europe is about 1.5°C in 25 years, and even larger in some sub-regions. The Western Europe region (see Supplementary Figure fs01), as well as the Mediterranean region (see Supplementary Figure fs02), show very similar changes to those seen for the European region. Even the Central European (see Supplementary Figure fs03) and UK (see Supplementary Figure fs04) regions, which are generally quite noisy, show this steady increase since the 1980s for most indices, with the exception of the time series of the Max1-day Tmax and the Max5-day Tmax across the UK region (Supplementary Figure fs04), in which a trend in the observations is less clear. The UK also features a particularly cold period in the 1960s, which seemed to have the strongest effect on the Max1-day Tmax and the Max5-day Tmax index.

The results of the detection analysis of all indices show that, which the exception of the changes in Max5-day Tmax across the UK and Central European region, all changes have been found to be significantly different from changes expected solely due to internal variability. Scaling factors are generally around magnitude 1 or larger, indicating that the observed change is well captured in the models or slightly underestimated (see red dots in Figure 3; see also Figure2). The best guess scaling factors of the UK region are found to be large for most indices, consistent with a trend that is possibly inflated due to the cold conditions in the initial period of the record analysed, but with large uncertainty ranges. For all regions and all indices considered the multi-model mean is consistent with the observations given uncertainty, which is illustrated by the uncertainty bar encompassing '1'. Figure 3 also shows that the uncertainty in scaling factors is larger for indices of the daily maximum temperature (right panel) than for indices of daily minimum temperature (left panel). The variance of the regression residual of the observations is found to be of comparable size to the one of the model internal variability, therefore the detection results can be considered robust. We also find that the uncertainty in scaling factors increases only slightly when analysing daily extremes rather than seasonal mean temperatures. This is consistent with *Hegerl et al.* [2004] who showed that daily extremes are almost as detectable as seasonal means over global land areas.

We have repeated the detection analysis with annual data (not shown) which shows very similar results to those obtained by analysing the indices smoothed by 5 years. The only exceptions were that in constrast to the analysis based on 5-yr averaged data, no detectable change was found in the 1day maximum indices across the UK and Central Europe. In conclusion, extremes of daily, 5-day and summer mean temperature show detectable changes across Europe in almost all subregions considered, with the exception of 5-day extremes of maximum temperature over the UK. This adds to a growing body of evidence that changes in the intensity and frequency of temperature extremes are detectable relative to climate variability. In some cases, these changes have been attributed to anthropogenic forcing (e.g., *Morak et al.* [2011]; *Christidis et al.* [2012]). The use of multi-model data as done here makes this result more robust to model uncertainty.

This detectable response to external forcing also leads to skill in near-term predictions through recreating reasonable trends in these indices. This skill due to forcing has a predictive capability which is useful to quantify (*Lee et al.* [2006]). MSSSs displayed in Figure 4 show how well the models forecast these extreme temperature indices on a decadal timescale. The different extremes studied can be affected by different physical processes, so we consider the skill of each index individually. Here, skill is defined as the absolute value of the modelled index being closer to the corresponding observation than the observed climatology (here calculated as the mean average observed value of this index calculated for 1961-2000). However, the same methodology could be used to test other benchmarks such as persistence (the index observed in the previous year) or a statistical model for example extrapolating observations. Since the study by *Hanlon et al.* [2013] showed the forecast skill, for similar indices, with the DePreSys forecasting system exceeds not only that of using climatology but also persistence, we do not further investigate persistence here.

Summer average Tmin is found to be significantly more skillfully predicted than climatology and random noise for HadCM3, MIROC5 and MPI-ESM-LR, across all regions (Figure 4, top right) and for all forecast periods considered. In contrast, CanCM4 shows very poor skill for this index across all regions considered here. Similar to the summer average Tmin, the summer average Tmax is more skillful than climatology and random noise across all time averages and regions for the MPI-ESM-LR, also for HadCM3 (except UK 6-9 year average) (Figure 4, top left). MIROC5 does not show consistent skill across leadtime averages, however, the decadal averages show skill in all regions but CEU (not shown). CanCM4 again shows no skill beyond climatology (see discussion below). As models do not show agreement for this index across regions/time averages the skill of the multi-model average also varies. Further investigation could enquire as to whether excluding models with lower skill would allow for more skilful multi-model predictions than that obtained when all are included. EU is predicted skillfully at all leadtimes. Over the UK the predictions are only skillful for the average of the first 5 years and MED is skillful for the last five years (6-9 years) of the forecast, and the decadal average (0-9 years). The reason for this is that the index computed with the decadal simulations is not fitting the observations as well in the UK as it does for the other more regions. As such, the decadal trend produced is not as close to that observed and affects how skilful the prediction is. This can be seen the time series for the UK region, shown in Supplementary Figure fs04, and echoes what was concluded in *Hanlon et al.* [2013] for the HadCM3 (DePreSys) model.

Closer investigation of the low skill scores obtained for CanCM4 reveals that this appears due to the model resisting bias correction. Specifically, some of the indices calculated with the CanCM4 decadal simulations display larger inter-annual variance than the observed index. As the bias correction applied has only corrected for the bias in the mean index over time, not the inter-annual variability, some significant bias remains. Since even small remaining biases influence the Mean Square error highly, this has a large negative impact on the skill of the CanCM4 model; and also on the skill of the multi-model averaged index. Methods for correcting the variance were explored, however a way of correcting the variance effectively across all indices could not be determined. Hence no correction to the variance was performed in order to prevent overcorrecting the index.

MPI-ESM-LR, HadCM3 and MIROC5 show skill beyond observed climatology and random noise for all time averages and regions except the UK for the Max5-day and Max1-day Tmin and Tmax indices (Figure 4, middle and bottom panels respectively). This is reflected by the multi-model average, which is generally skillful in these regions for the Tmax extremes but not in all cases and least often for the Tmin extremes. The forecast for the UK generally shows no skill beyond observed climatology and random noise except for the decadal average Max5-day/Max1-day Tmin (MIROC5 and MPI-ESM-LR) and the CanCM4 decadal average Max5-day/Max1-day Tmax.

The majority of models and the multi-model average indices do not show any improvement of skill of the initialised decadal runs over the historical runs which do not assimilate observations (Figure 5). There are exception to this, especially for the MPI-ESM-LR, whose decadal runs are more skillful than the historical runs for most indices, consistent with findings of skill in annual data (*Matei et al.* [2003]). As these runs were also skillful beyond climatology (Figure 4), the initialisation is improving the prediction in this case. Other cases which hint at some improvement due to initialisation include: HadCM3 Europe average extreme indices, HadCM3 Europe 5-year average summer average Tmax, HadCM3 Mediterranean summer average and Max5-day Tmin, MIROC5 UK Max5-day extremes and MIROC5 UK decadal average Max1-day extremes. However, since not all models show this improvement by initialisation the multi-model mean does not either, in general. Where the skill seen in Figure 4 is not added to by the initialisation, the alternative source of skill is due to the model forcing, recreating the observed trend in temperatures over time. This could originate both from the model correctly simulating long-term warming trends, or from correctly simulating circulation changes. As most of the robust skill originates from forcing, this suggests a large role for long-term warming.

5. Discussion and Conclusion

This work shows evidence of an increase of the magnitude in both moderate and 1- or 5-day temperature extremes during summer over the analysis period 1961-2005. This observed increase is well represented by the multi-model mean and the observed variability is within the ensemble range. Changes in most indices are found to be detectable across

Europe and most of its sub-regions. Only changes in the average 5-day maximum temperature across the UK and Central Europe region are not significant. This suggests that the forced response should have predictive skill for the near-term, for example, following the ASK method (*Allen et al.* [2000]; *Stott and Kettleborough* [2002]), although in the present case it is based on the total response rather than greenhouse gas only response.

Analysis of the decadal simulations has confirmed this potential for skill: predictions from 3 out of the 4 models tested are closer to observations than predictions made using observed climatology and random noise for summer average maximum and minimum temperatures and for 5 and 10-year averaged indices of daily and 5-day extremes, again with the exception of daily extremes in the UK. There is also significantly increased skill in the initialised simulations relative to the non-initialised simulations, in some models, for some indices. However, the majority of the skill is due to the model representation of the external forcing allowing the model to recreate the observed trend, consistent with the detection results. The MPI-ESM-LR seems to be the most skillful for our regions, with additional skill coming from the initialisation of this model. The other models do not consistently show that the skill of predictions increases due to the initialisation compared to the historically forced simulations. Also, poor skill in some prediction systems for these European summer temperature indices leads to reduced skill in the multi-model mean prediction.

Across the regions, most models show decadal skill for the regions consisting largely of mainland Europe, while the UK region is the least skillful region, likely due to greater variability in this smaller region, which has also impacted detection results by increasing uncertainties (Figure 3). The varying amounts of skill obtained for the different indices across different models and regions highlights the need to take care when using model forecasts to make predictions of changes in extremes. Different models include different physics and have different forecasting abilities so it is important to measure the skill of each prediction system for each case individually before using it to make a prediction. This point is particularly important when using global models. Further downscaling/impact modelling may be employed to get relevant information on smaller spatial scales, particularly for variables with high spatial variability such as precipitation. Even where downscaling methods are used, analysis of the skill of global models over large regional scales, is useful to determine if any driver model for downscaling captures changes reasonably well, since it can inform the choice of global model which would be best to drive these downscaling/impact models.

Acknowledgments. HMH was supported by the UK Natural Environment Research Council through the EQUIP project (grant NE/H003525/1) and all authors HMH and SM were supported by a NERC CMIP5 project.

We acknowledge the E-OBS dataset from the EU-FP6 project ENSEMBLES (http://ensembles-eu.metoffice.com) and the data providers in the ECA&D project (*Haylock et al.* [2008]) and also Edinburgh Compute and Data Facility (ECDF) for providing computer resources. We would also like to acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modelling groups (Canadian Centre for Climate Modelling and Analysis, Met Office Hadley Centre, Max Planck Institute and Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies) for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. We also thank Simon Tett for discussion of the error estimates, and the anonymous reviewers for their constructive and helpful suggestions.

References

Allen, M.R., P.A. Stott, J.F.B Mitchell, R. Schnur and T.L. Delworth. Quantifying the uncertainty in forecasts of anthropogenic climate change *Nature*, 407:617–620, 2000.

Allen, M.R. and P.A. Stott. Estimating signal amplitudes in optimal fingerprinting, Part I: theory. *Climate Dynamics*, 21(5):477–491, 2003.

Allen, M.R. and S.F.B. Tett. Checking for model consistency in optimal fingerprinting. *Climate Dynamics*, 15(6):419–434, 1999.

Barriopedro, D., E.M. Fischer, J. Luterbacher, R.M. Trigo and R. Garca-Herrera. The hot summer of 2010: redrawing the temperature record map of Europe. *Science*, 332:220–224, 2011.

Collins, M., S.F.B Tett, and C. Cooper. The internal climate variability of HadCM3, a version of the Hadley centre coupled model without flux adjustments. *Climate Dynamics*, 17(1):61–81, 2001.

Christidis, N., P.A. Stott, S. Brown, G. C. Hegerl and J. Caesar Detection of changes in temperature extremes during the 20th century. *Geophysical Research Letters*, 32, L20716, 2005.

Christidis, N., P.A. Stott, G.S. Jones, H. Shiogama, T. Nozawa and J. Luterbacher. Human activity and anomalously warm seasons in Europe *International Journal of Climatology*, 32(2):225–239, 2012.

Díaz, J., C. Linares and A. Tobías. Impact of extreme temperatures on daily mortality in Madrid (Spain) among the 45-64 age-group *International journal of biometeorology*, 50(6):342–348, 2006.

Dole, R., M. Hoerling, J. Perlwitz, J. Eischeid, P. Pegion, Tao Zhang, Xiao-Wei Quan, Taiyi Xu, and D. Murray. Was there a basis for anticipating the 2010 Russian heat wave? *Geophys. Res. Lett.*, 38(6), 03 2011.

Eade, R., E. Hamilton, D.M. Smith, R.J. Graham and A.A. Scaife. Forecasting the number of extreme daily events out to a decade ahead *Journal of Geophysical Research*, 117(D21110), 2012.

Efron, B. and R J Tibshirani. An introduction to the bootstrap *Chapman And Hall*, 1993.

Ferranti, L. and P. Viterbo. The European summer of 2003: sensitivity to soil water initial conditions *ECMWF Technical Memorandum*, 438:1–29, 2006.

Fink, A.H., T. Brücher, A. Krüger, G.C.Leckebusch, J.G. Pinto, and U. Ulbrich. The 2003 European summer heatwaves and drought - synoptic diagnosis and impacts. *Weather*, 59(8):209–216, 2004.

Fischer, E.M., S.I. Seneviratne, D. Lüthi and C. Schär. Contribution of land-atmosphere coupling to recent European summer heat waves *Geophysical Research Letters*, 34, 2007.

Fouillet, A., G. Rey, F. Laurent, G. Pavillon, S. Bellec, C. Guihenneuc-Jouyaux, J. Clavel, E. Jougla, D. Hémon. Excess mortality related to the August 2003 heat wave in France

International archives of occupational and environmental health , 80(1):16–24, 2006.

Goddard, L., A. Kumar, A. Solomon, D. Smith, G. Boer, P. Gonzalez, C. Deser, S. Mason, B. Kirtman, R. Msadek, R. Sutton, E. Hawkins, T. Fricker, S. Kharin, W. Merryfield, G. Hegerl, C. Ferro, D. Stephenson, G.A. Meehl, T. Stockdale, R. Burgman, A. Greene, Y. Kushnir, M. Newman, J. Carton, I. Fukumori, D. Vimont and T. Delworth. A verification framework for interannual-to-decadal prediction experiments *Climate Dynamics*, 40, 245-272, 2012.

Grize, L., A. Hussa, O. Thommena, C. Schindlera, C. Braun-Fahrländera. Heat wave 2003 and mortality in Switzerland *Swiss Medical Weekly*, 135:200–205, 2005.

Fischer, E. M., S.I. Seneviratne, P.L. Vidale, D. Lüthi and C. Schär. Soil moisture-atmosphere interactions during the 2003 European summer heat wave. *Journal of Climate*, 20:50815099, 2007.

Hamilton, E., R. Eade, R.J. Graham, A.A. Scaife D.M. Smith A. Maidens and C. MacLachlan. Forecasting the number of extreme daily events on seasonal timescales *Journal of Geophysical Research*, 117, 2012.

Hanlon, H., G.C. Hegerl, S.F.B. Tett and D. M. Smith. Can a decadal forecasting system predict temperature extreme indices? submitted to *Journal of Climate*, 26: 3728–3744, 2013.

Hanlon, H. An investigation of causes of the 2003 heatwave in Europe using an atmospheric climate model *University of Oxford*, DPhil Thesis, 2010.

Hasselmann, K. Optimal fingerprints for the detection of time-dependent climate change. *Journal of Climate*, 6(10):1957–1971, 1993.

Hasumi, H. and S. Emori Coupled GCM (MIROC)Description. *Center for Climate System Research, University of Tokyo.*

Haylock, M.R., N. Hofstra, A.M.G. Klein Tank, E.J. Klok, P.D. Jones, and M. New. A European daily high-resolution gridded data set of surface temperature and precipitation for 1950 *J. Geophys. Res.*, 113(D20):D20119, 10 2008.

Hegerl, G.C., F. Zwiers, S. Kharin and P. Stott Detectability of anthropogenic changes in temperature and precipitation extremes. *J. Climate*, 17: 3683-3700, 2004.

Hegerl, G.C., O. Hoegh-Guldber, G. Casassa, M.P. Hoerling, R.S. Kovats, C. Parmesan, D.W. Pierce, and Stott P.A. Good practice guidance paper on detection and attribution related to anthropogenic climate change. *IPCC Working Group I Tech. Support Unit*, Univ. of Bern, Bern, 2010.

Hegerl, G.C., F.W. Zwiers, P. Braconnot, N.P. Gillett, Y. Luo, J.A. Marengo Orsini, N. Nicholls, J.E. Penner, and P.A. Stott. Understanding and Attributing Climate Change. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. *Cambridge University Press*, Cambridge, United Kingdom and New York, 2007.

Jungclaus, J.H. et al. Characteristics of the ocean simulations in MPIOM, the ocean component of the MPI-Earth System Model. *Journal of Advances in Modeling Earth Systems*, DOI: 10.1002/jame.20023, http://onlinelibrary.wiley.com/doi/10.1002/jame.20023/abstract

Karoly, D.J. and P.A. Stott. Anthropogenic warming of central England temperature. *Atmospheric Science Letters*, 7(4):81–85, 2006.

Lee, T.C.K, F.W. Zwiers, X. Zhang and M Tsao. Evidence of decadal climate prediction skill resulting from changes in anthropogenic forcing *Journal of climate*,19(20):5305–5318, 2006.

Marsland, S.J., H. Haak, J.H. Jungclaus, M. Latif and F. Roske. The Max-Planck-Institute global ocean/sea ice model with orthogonal curvilinear coordinates. *Ocean Modelling*, 5(2):91-127, 2003.

Matei, D., H. Pohlmann, J. Jungclaus, W. Müller, H. Haak and J. Marotzke. Two tales of initializing decadal climate prediction experiments with the ECHAM5/MPI-OM model *Journal of Climate*, 25(24):8502–8523, 2012.

Meehl, G.A., L. Goddard, J. Murphy, R. J. Stouffer, G. Boer, G. Danabasoglu, K. Dixon, M. A. Giorgetta, A. M. Greene, E. Hawkins, G. Hegerl, D. Karoly, N. Keenlyside, M. Kimoto, B. Kirtman, A. Navarra, R. Pulwarty, D. Smith, D. Stammer, and T. Stockdale. Decadal Prediction Can It Be Skillful? *Bull. Amer. Meteor. Soc.*, 90: 1467–1485, 2009.

Merryfield, W.J., W.S. Lee, G.J. Boer, V.V. Kharin, J.F. Scinocca, G.M. Flato, R.S. Ajayamohan, J.C.Fyfe, Y. Tang, S. Polavarapu. The Canadian Seasonal to Interannual Prediction System. Part I: Models and Initialization. *Monthly Weather Review*, e-View doi: http://dx.doi.org/10.1175/MWR-D-12-00216., 2013.

Morak, S., G.C. Hegerl and J. Kenyon. Detectable regional changes in the number of warm nights. *Geophysical Research Letters*, 38(17), 2011.

Morak, S., G.C. Hegerl and N. Christidis. Detectable changes in the frequency of temperature extremes *Journal of Climate.*, 26: 1561–1574, 2013.

Murphy, A.H. Skill scores based on the mean square error and their relationships to the correlation coefficient *Monthly Weather Review*, 16:2417–2424, 1988.

Otto, F.E.L., N. Massey, G.J. van Oldenborgh, R.G. Jones, and M.R. Allen. Reconciling two approaches to attribution of the 2010 Russian heat wave. *Geophysical Research Letters*, 39(4), 02 2012.

Pascal, M., K. Laaidi, M. Ledrans, E. Baffert, C. Caserio-Schönemann, A. Le Tertre, J. Manach, S. Medina, J. Rudant and P. Empereur-Bissonnet. France's heat health watch warning system *International Journal of Biometeorology*, 50(3):144–153, 2006.

Raddatz, T.J., C.H. Reick, W. Knorr, J. Kattge, E. Roeckner, R. Schnur, K.G. Schnitzler, P. Wetzel, and J. Jungclaus. Will the tropical land biosphere dominate the climate–carbon cycle feedback during the twenty-first century? *Climate Dynamics*, 29(6):565–574, 2007.

Rahmstorf, S. and D. Coumou. Increase of extreme events in a warming world. *Proceedings of the National Academy of Sciences*, 10 2011.

Schär, C., P.L. Vidale, D. Lüthi, C. Frei, C. Häberli, M. Liniger and C. Appenzeller. The role of increasing temperature variability for European summer heat waves. *Nature*, 427(6972):332–336, 2004.

Seneviratne, S., T. Corti, E.L. Davin, M. Hirschi, E.B. Jaeger, I. Lehner, B. Orlowsky and A.J. Teuling. Investigating soil moisture-climate interactions in a changing climate: A review *Earth-Science Reviews*, 99:125–161, 2010.

Seneviratne, S., D. Lüthi, M. Litschi and C. Schär. Land-Atmosphere Coupling and climate change in Europe *Nature*, 443:205–209, 2006.

Smith, D.M., S. Cusack, A.W. Colman, C.K. Folland, G.R. Harris, and J.M. Murphy. Improved surface temperature prediction for the coming decade from a global climate model. *Science*, 317(5839):796–799, 08 2007.

Smith, D.M., R. Eade, N.J. Dunstone, D. Fereday, J.M. Murphy, H. Pohlmann, and A.A. Scaife. Skilful multi-year predictions of atlantic hurricane frequency. *Nature Geosci*, 3(12):846–849, 12 2010.

Stott,P.A., and J.A. Kettleborough. Origins and estimates of uncertainty in predictions of twenty-first century temperature rise *Nature*, 416:723–726, 2002.

Stott, P.A., D.A. Stone, and M.R. Allen. Human contribution to the European heatwave of 2003. *Nature*, 432(2):610–613, 2004.

Stott, P.A., N.P. Gillett, G.C. Hegerl, D.J. Karoly, D.A. Stone, X. Zhang and F. Zwiers Detection and attribution of climate change: a regional perspective *WIRES*, 1:191–211, 2010.

Taylor, K.E., R.J. Stouffer and G.A. Meehl. An Overview of CMIP5 and the experiment design *Bull. Amer. Meteor. Soc.*, 93:485–498, 2012.

Vautard, R., P. Yiou, F. D'Andrea, N. de Noblet, N. Viovy, C. Cassou, J. Polcher, P. Ciais, M. Kageyama and Y. Fan. Summertime European heat and drought waves induced by wintertime Mediterranean rainfall deficit *Geophysical Research Letters*, 34, 2007.

Von Storch, H., and F.W. Zwiers Statistical Analysis in Climate Research *Cambridge University Press*, 2000.

Watanabe, M., T. Suzuki, R. O'ishi, Y. Komuro, S. Watanabe, S. Emori, T. Takemura, M. Chikira, T. Ogura, M. Sekiguchi, K. Takata, D. Yamazaki, T. Yokohata, T. Nozawa, H. Hasumi, H. Tatebe, and M. Kimoto. Improved climate simulation by MIROC5: Mean states, variability, and climate sensitivity. *Journal of Climate*, 23(23):6312–6335, 2010.

Zwiers, F.W., X. Zhang and Y. Feng. Anthropogenic influence on long return period daily temperature extremes at regional scales. *Journal of Climate*, 24(3):881–892, 2011.

Von Salzen, K., J.F. Scinocca, N.A. McFarlane, J. Li, J.N.S. Cole, D. Plummer, D. Verseghy, M.C. Reader, X. Ma, M. Lazare and L. Solhiem. The Canadian fourth generation atmospheric global climate model(CanAM4). Part 1: representation of physical processes. *Atmosphere Ocean*, 51(1):104–125, 2013.

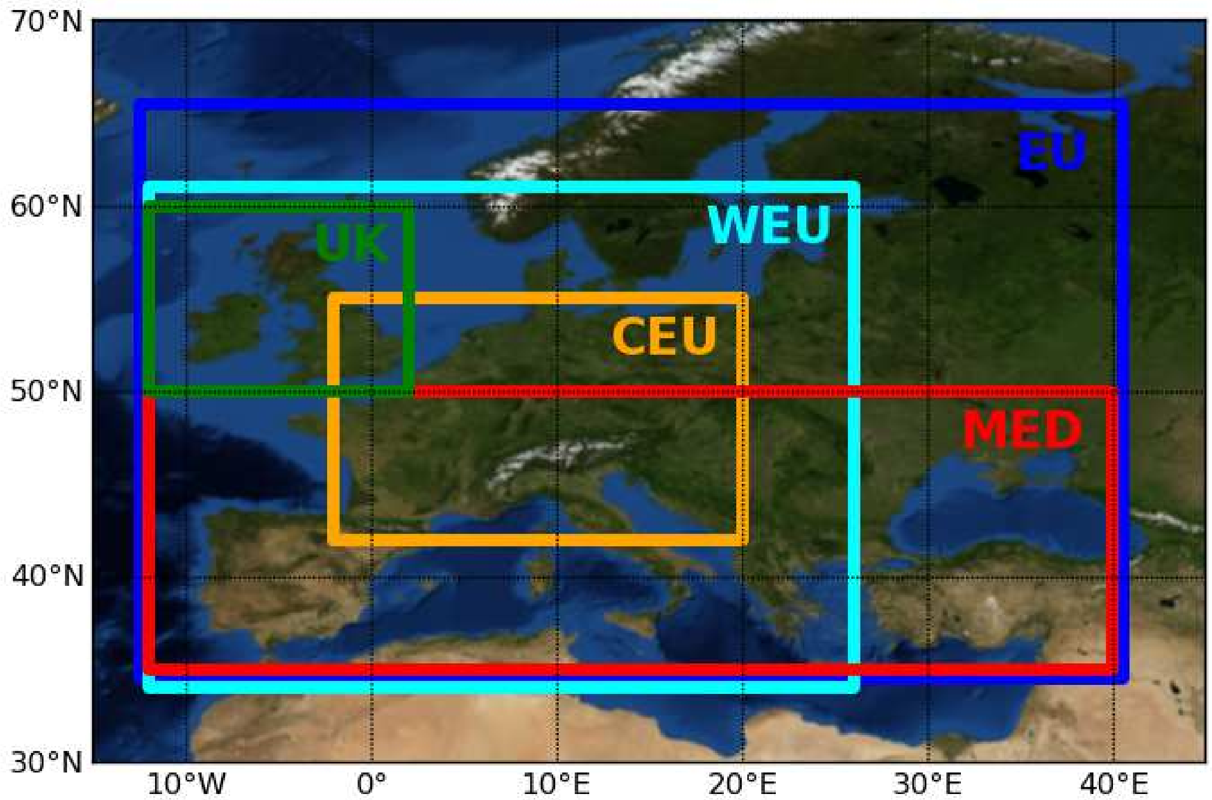

Figure 1. Regions used in this study

Table 1. Model Description

Model	Horizontal resolution	No. of vertical levels	Ocean Coupling	Reference
CanCM4	2.8125° lon x 2.7906° lat	35	"CanOM4" 40 vertical level (*Merryfield et al.* [2013])	(*von Salzen et al.* [2013])
HadCM3	3.75° lon x 2.5° lat	19	"HadOM" 1.25°x1.25° 20 vertical levels	*Collins et al.* [2001], *Smith et al.* [2007], *Smith et al.* [2010])
MIROC5	1.406° lon x 1.4° lat	40	"COCO4.5" 1.4°lat x 0.5-1.4°lon, 50 vertical levels (*Hasumi* [2004])	*Watanabe et al.* [2010]
MPI-ESM-LR	1.875° lon x 1.865° lat	47	"MPIOM", 1.5°lat/lon, 40 vertical levels (*Jungclaus et al.* [2012])	*Raddatz et al.* [2007], *Marsland et al.* [2003]

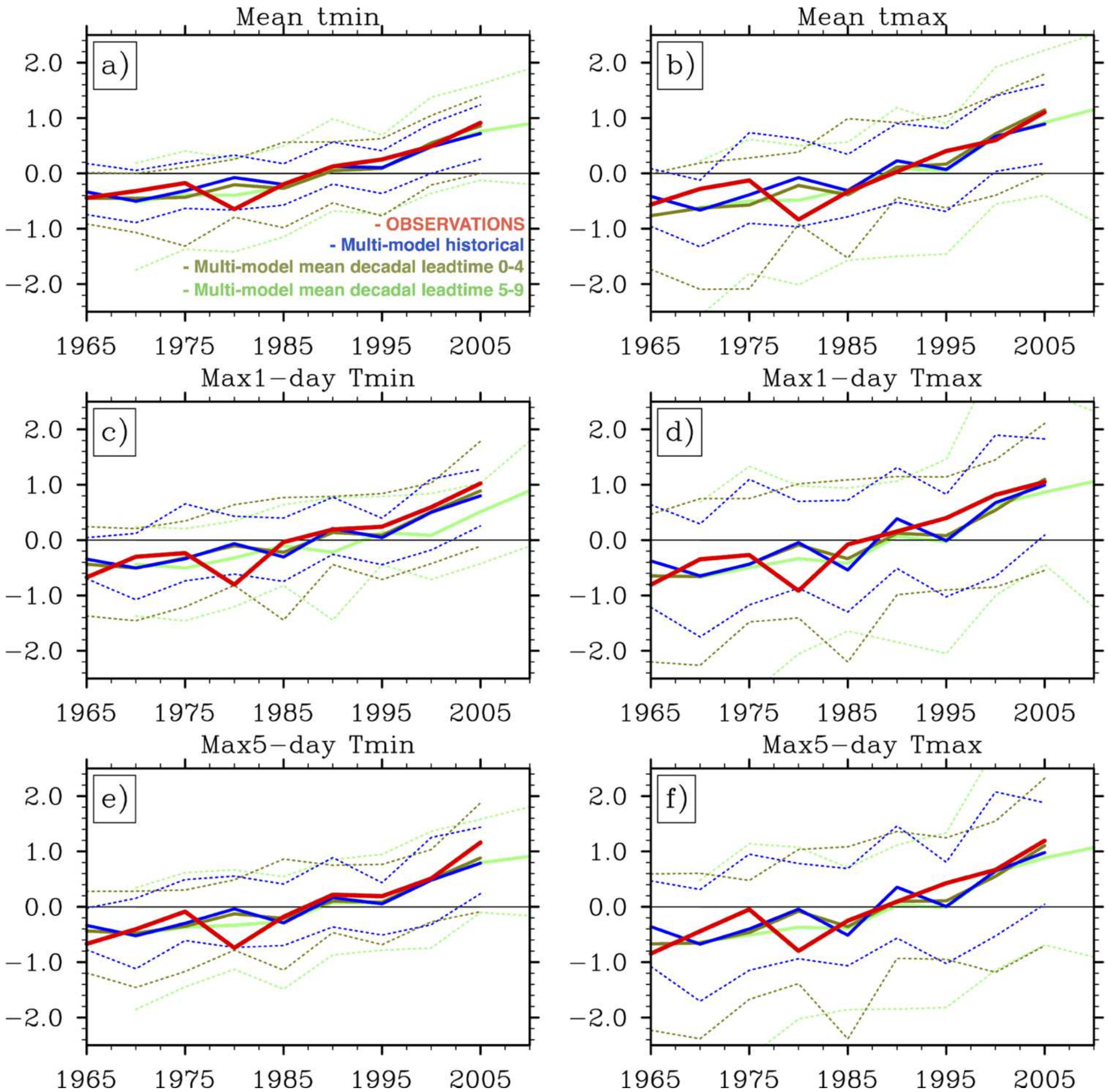

Figure 2. 5-year average time series of the magnitude of anomalies relative to the reference period 1961-2005 in a) mean summer minimum temperature, b) mean summer maximum temperature, c) Max1-day Tmin, d) Max1-day Tmax, e) Max5-day Tmin and f) Max5-day Tmax across Europe. Observations are shown in red. The multi-model mean of the historical runs is shown by the blue lines. A time series consisting of the multi-model mean of the average of the first (last) five years from each decadal run is shown in dark green (light green) . The ensemble spread is shown for each time series by the dashed lines.

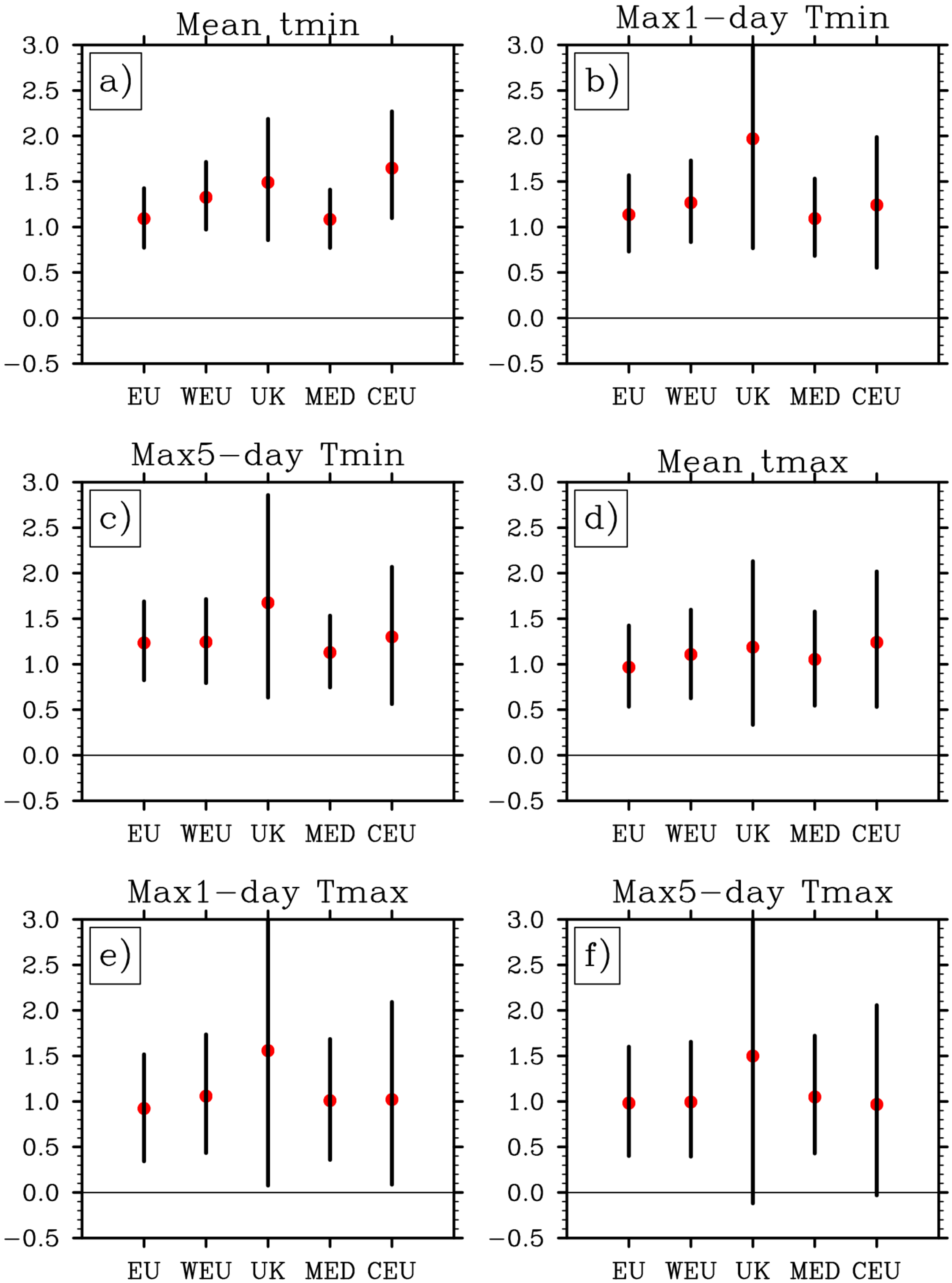

Figure 3. Scaling factors (red dots) plus 5-95% uncertainty range (vertical bars) of changes in the magnitude of a) mean summer minimum temperature, b) mean summer maximum temperature, c) Max1-day Tmin, d) Max1-day Tmin, e) Max5-day Tmin and f) Max5-day Tmax across Europe and sub-regions, WEU, UK, MED and CEU.

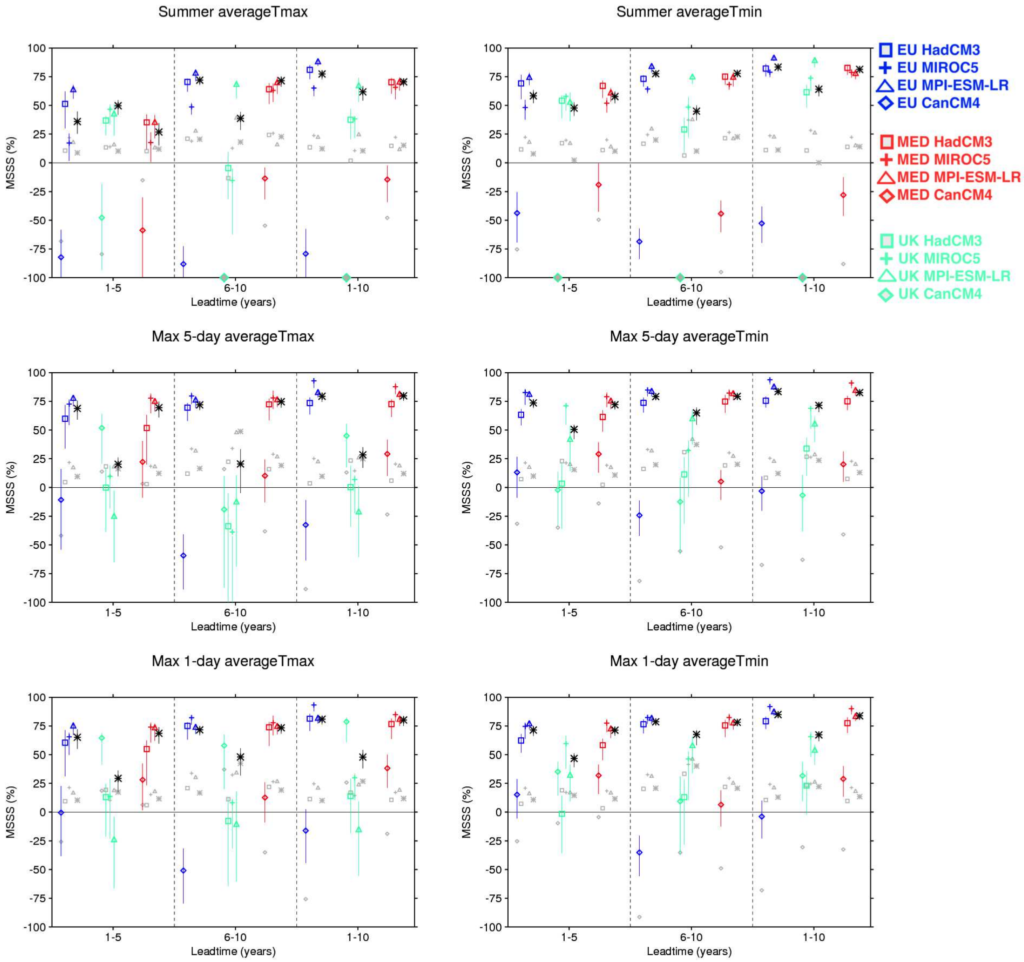

Figure 4. Mean Square Skill Score (MSSS) of the summer average (top), Max5-day average (middle) and Max1-day average (bottom) Tmax(left) and Tmin (right) averaged over 5/10 years for each model (CanCM4(diamond), HadCM3(square), MIROC5(cross) and MPI-ESM-LR(triangle) and the multi-model average (black star)) compared to E-OBS observed climatology (1961-2000). These scores are computed with regionally averaged indices for EU(Blue), UK (green) and MED(red). WEU and CEU were found to be very similar to EU and MED respectively so are omitted from this figure. To be skillful, the MSSS and its associated 10-90% error bar (calculated using bootstrapping with replacement) must be above zero and to be significantly different to noise, the model MSSS must be greater than MSSS obtainable with 90th percentile of realisations of random noise (shown by a smaller grey symbol), see 3.3. Where the MSSS is below -100 the forecast is particularly unskillful compared to climatology, an enlarged symbol filled with grey shading is placed at the bottom of the plot to highlight these cases.

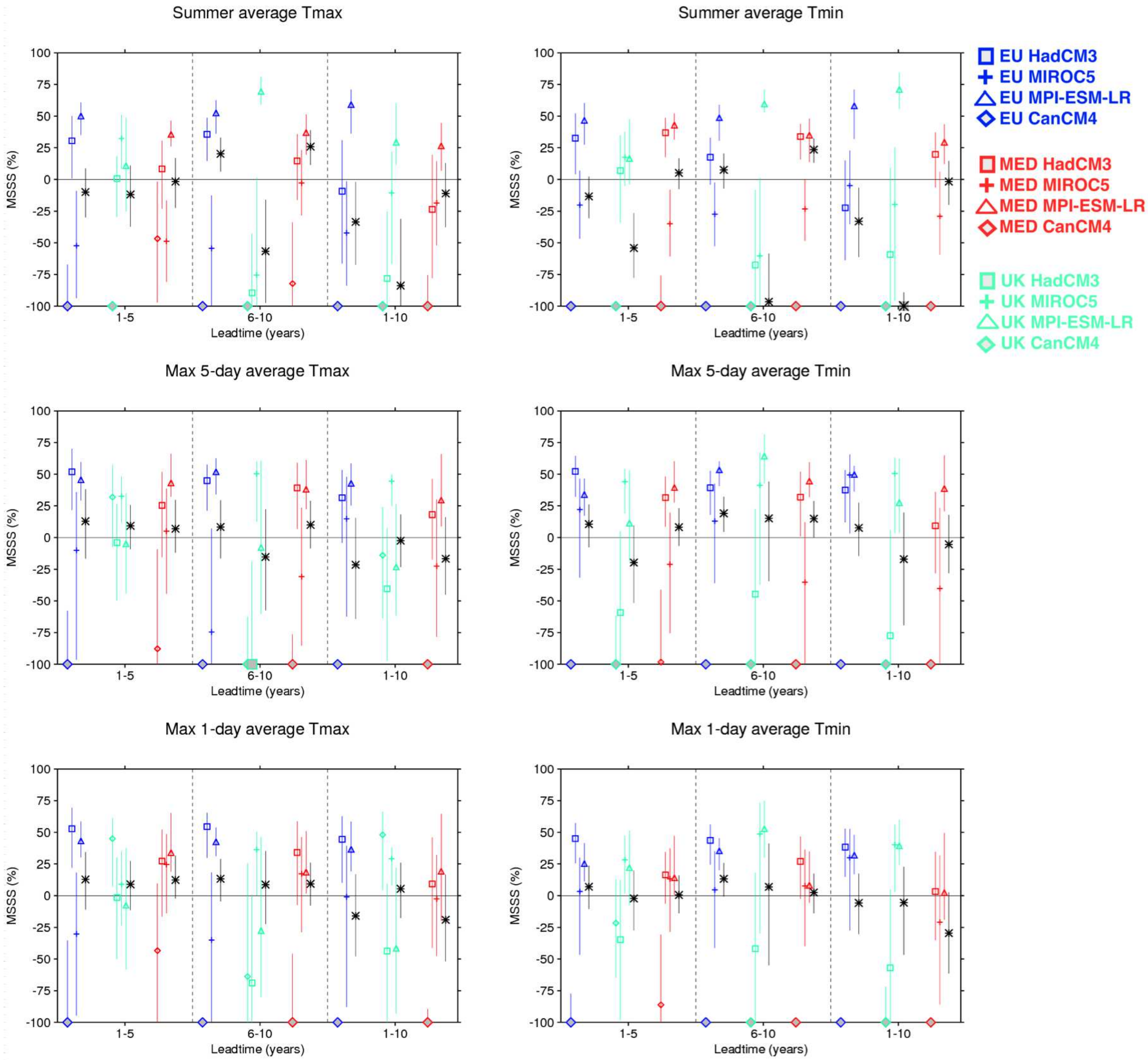

Figure 5. As in Figure 4 but the MSSS for the indices computed with decadal simulations is compared to the equivalent indices computed with the historical simulations instead of observed climatology. Positive significant skill indicates the decadal forecasting system has higher skill than the historical uninitialised runs.