Detection and prediction of mean and extreme European summer temperatures with a multimodel ensemble

Citation for published version:
Hanlon, HM, Morak, S & Hegerl, GC 2013, 'Detection and prediction of mean and extreme European summer temperatures with a multimodel ensemble', *Journal of Geophysical Research: Atmospheres*, vol. 118, no. 17, pp. 9631-9641. https://doi.org/10.1002/jgrd.50703

Digital Object Identifier (DOI):
10.1002/jgrd.50703

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Journal of Geophysical Research: Atmospheres

Detection and prediction of mean and extreme European summer temperatures with a multimodel ensemble

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

Received 31 July 2012; revised 30 July 2013; accepted 1 August 2013; published 9 September 2013.

[1] We analyze 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 1960 to 2011, taken from E-OBS an observational dataset created for the European Commission funded project (ENSEMBLES), 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 temperatures during summer. The region of interest is Europe; however, a number of subregions are also studied, which include Western Europe, the British Isles, the Mediterranean, and Central Europe. The observed changes in the analyzed indices are well represented by the multimodel mean and are within the range of the multimodel ensemble for most regions, with the exception of 1 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 three of the four 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 initialization in some cases.

Citation: Hanlon, H. M., S. Morak, and G. C. Hegerl (2013), Detection and prediction of mean and extreme European summer temperatures with a multimodel ensemble, *J. Geophys. Res. Atmos.*, *118*, 9631–9641, doi:10.1002/jgrd.50703.

1. Introduction

[2] Recent years have seen two of the most devastating extreme heat wave events in Eurasia, the 2003 European heat wave [*Schär et al.*, 2004; *Fink et al.*, 2004; *Hanlon*, 2010] and 2010 Russian heat wave [*Barriopedro et al.*, 2011; *Dole et al.*, 2011; *Rahmstorf and Coumou*, 2011; *Otto et al.*, 2012]. The European summer heat wave of 2003 exhibited anomalously hot temperatures, with the European continental mean summer average temperature exceeding the long-term mean (1961–1990) 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*, 2010; *Fischer et al.*, 2007a, 2007b].

[3] These extreme heat wave 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 [*Díaz et al.*, 2006; *Fouillet et al.*, 2006; *Grize et al.*, 2005; *Pascal et al.*, 2006].

[4] 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 occurring, due to internal variability alone, is evaluated to be small [*Hegerl et al.*, 2007, 2010], while attribution analyses evaluate several potential

Additional supporting information may be found in the online version of this article.

[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.

Corresponding author: H. M. Hanlon, Met Office Hadley Centre, Fitzroy Road Exeter, EX1 3PB, UK. (helen.hanlon@metoffice.gov.uk)

2169-897X/13/10.1002/jgrd.50703

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, 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].

[5] 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 heat wave 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 heat wave event; however, there are 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 and Coumou* [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 and Coumou* [2011] and *Dole et al.* [2011] conclusions were not mutually exclusive.

[6] 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 (one in 10) daily temperature extremes on decadal timescales using the Met Office Hadley Centre decadal prediction system (DePreSys). These are initialized decadal predictions which attempt to provide improved predictions of natural internal variability [*Smith et al.*, 2010]. *Hanlon et al.* [2013] have shown that 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 initialization of observations.

[7] 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 multimodel 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 heat wave 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 initialization of these models with observed values.

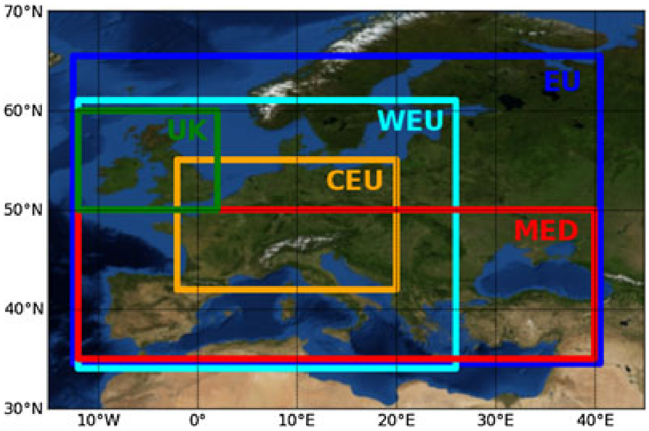

Figure 1. Regions used in this study.

[8] 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

[9] This study uses gridded observed and model-simulated data sets of 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°N–65°N latitude,12°W–40°E longitude), along with subregions Western Europe (WEU, 34°N–61°N latitude, 12°W–26°E longitude), UK and Ireland (UK, 50°N–60°N latitude, 12°W–2°E longitude), Mediterranean (MED, 35°N–50°N latitude, 12°W–40°E longitude), and Central Europe (CEU, 42°N–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

[10] The observed data originate from the ENSEMBLES project observational database (E-OBS), which is a high-resolution (0.5° latitude by 0.5° longitude grid) gridded data set 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

[11] 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 multimodel information that is much more robust than the use of

Table 1. Model Description

Model	Horizontal Resolution	No. of Vertical Levels	Ocean Coupling	Reference
CanCM4	2.8125° longitude × 2.7906° latitude	35	"CanOM4" 40 vertical level [*Merryfield et al.*, 2013]	*Von Salzen et al.* [2013]
HadCM3	3.75° longitude × 2.5° latitude	19	"HadOM" 1.25°×1.25° 20 vertical levels	*Collins et al.* [2001], *Smith et al.* [2007], *Smith et al.* [2010]
MIROC5	1.406° longitude × 1.4° latitude	40	"COCO4.5" 1.4°latitude × 0.5–1.4°longitude, 50 vertical levels [*Hasumi and Emori*, 2004]	*Watanabe et al.* [2010]
MPI-ESM-LR	1.875° longitude × 1.865° latitude	47	"MPIOM," 1.5°latitude/longitude, 40 vertical levels [*Jungclaus et al.*, 2012]	*Raddatz et al.* [2007], *Marsland et al.* [2003]

single models which is often applied in detection studies for extremes [*Morak et al.*, 2013; *Christidis et al.*, 2005]. For model description, see Table 1.

[12] The forcing of the historical runs includes anthropogenic forcing, such as the observed concentrations of greenhouse gases and aerosols, generally direct as well as indirect forcing, and natural forcing such as the recorded changes in volcanic aerosol or changes in solar activity for the twentieth 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 (four ensemble members), and MPI-ESM-LR (three ensemble members).

[13] 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 initialized 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 (six ensemble members), and MPI-ESM-LR (10 ensemble members).

3. Methodology

3.1. Indices Computation and Processing

[14] The following six indices have been computed and analyzed throughout this study:

[15] 1. Summer average minimum temperature: the mean average daily minimum temperature computed over the summer season June–August.

[16] 2. Summer average maximum temperature: the mean average daily maximum temperature computed over the summer season June–August.

[17] 3. Max 1day Tmin: the highest daily minimum temperature that occurred between 1 April and 30 September.

[18] 4. Max 1day Tmax: the highest daily maximum temperature that occurred between 1 April and 30 September.

[19] 5. Max 5day Tmin: the highest 5 day rolling mean average daily minimum temperature that occurred between 1 April and 30 September.

[20] 6. Max 5day Tmax: The highest 5 day rolling mean average daily maximum temperature that occurred between 1 April and 30 September.

[21] The indices were computed for the observations and the model runs (both historical and decadal simulations) on their respective grids, which were then regridded using nearest neighbor interpolation to the grid of HadCM3, which is the coarsest grid of all data sets (3.75° longitude × 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 are 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 interannual variability. The multimodel mean time series was computed by averaging over all multimodel ensemble members.

3.2. Detection Analysis

[22] 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 follows:

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

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 for the multimodel mean time series for the corresponding index, ν_1 for a realization of the model internal variability, and ν_0 for a realization of the observed variability.

[23] Using this method, we obtain scaling factors α, which are the factors by which the fingerprints (here we have used "non-optimized" 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 optimized 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 modeled 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].

[24] The samples of internal variability (noise) are obtained from the model-simulated variability of each individual model run after subtracting the multimodel 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 5th and 95th percentiles of the scaling factors.

[25] 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 eightieth 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 multimodel mean is statistically consistent with the observations.

3.3. Prediction Analysis

[26] The indices detailed in section 3.1 are also computed for the decadal hindcasts, which are model simulations initialized 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 modeled 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

$$x_{i,t,m=m*} = x_{i,t,m=m*} - \frac{1}{10}\sum_{i=0}^{9}\frac{\sum\limits_{m\notin m*} x_{i,t=0,m}}{n-1} + \frac{1}{40}\sum_{\text{year}=1961}^{2000} y_{\text{year}} \qquad (2)$$

where m is a set of all ensemble members, $m*$ is each individual ensemble member, ($m* = 1$ to n) and n is the number of members. i corresponds to each of the 10 year runs started every 5 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, and then $i = 1$, $t = 0$ is summer 1966, the first summer in the run started in 1965 and so on. We correct the index by removing the mean modelled index. This is calculated by taking the index computed at leadtime 0 (t=0, the index value for the first summer) for all ensemble members but the member being corrected (this is sometimes described as "leave one out"), for each of these the average over each run i is taken, then finally the index is averaged across these selected members. In order to avoid overcorrecting, the member being corrected is left out of this average when the ensemble average of the mean modeled index is calculated. Hence, the correction applied across each 10 year run remains constant across different lead times within that run and as such does not account for drift in the model at later lead times. To perform the correction, the mean modeled index is subtracted from the modeled index for each member individually, and then mean observed index (averaged over all years between 1961 and 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.

[27] 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 multimodel 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 multimodel mean.

[28] 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 modeled forecast is closer to the observation than the alternative method of prediction (e.g., 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.

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

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

where MSE denotes the mean square error calculated across all 10 year runs at individual lead times ($\mathbf{x}_t$) compared to corresponding observations ($\mathbf{y}_t$), $\mathbf{x}_{i,t} = \sum\limits_{m=0}^{n} x_{i,t,m}$ is the ith ensemble mean decadal forecast at lead time t, y_i is the ith observed value corresponding to the same year as lead time t, and 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.

[29] 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 and 2000 (as outlined in section 3.1).

[30] 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 initialized 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 modeled 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 which assimilates observations has, compared to the unassimilated ensemble, that has no initial knowledge of the observed state of the climate.

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

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

[31] The MSSSs (equations (4) and (5)) are calculated for 5 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.

[32] The MSSS is computed from the ensemble average of the regionally averaged index at each lead time 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 realization, 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 realizations of the MSSS, and the 10–90% range from these provide the uncertainty on the MSSS. If the score is significantly above 0, then the forecast has more skill in predicting the index than the reference forecast, for example, the in-sample observed climatology or the uninitialized historical simulations.

[33] An additional method of estimating uncertainty compares 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 normalize the random forecast. One thousand random forecast realizations are generated, and a distribution of MSSSs is computed from these. The 90th percentile of this distribution is taken as a cutoff point, below which the MSSSs for the decadal hindcasts are considered not significantly better than random noise.

4. Results

[34] 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 and 5 day extremes. The observed change is well represented by the multimodel 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 subregions. The Western Europe region (see supporting information Figure S1) and the Mediterranean region (see supporting information Figure S2) show very similar changes to those seen for the European region. Even the Central European (see supporting information Figure S3) and UK (see supporting information Figure S4) 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 Max 1day Tmax and the Max 5day Tmax across the UK region (supporting information Figure S4), 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 Max 1day Tmax and the Max 5day Tmax index.

[35] The results of the detection analysis of all indices show that, with the exception of the changes in Max 5day 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 Figure 2). 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 analyzed, but with large

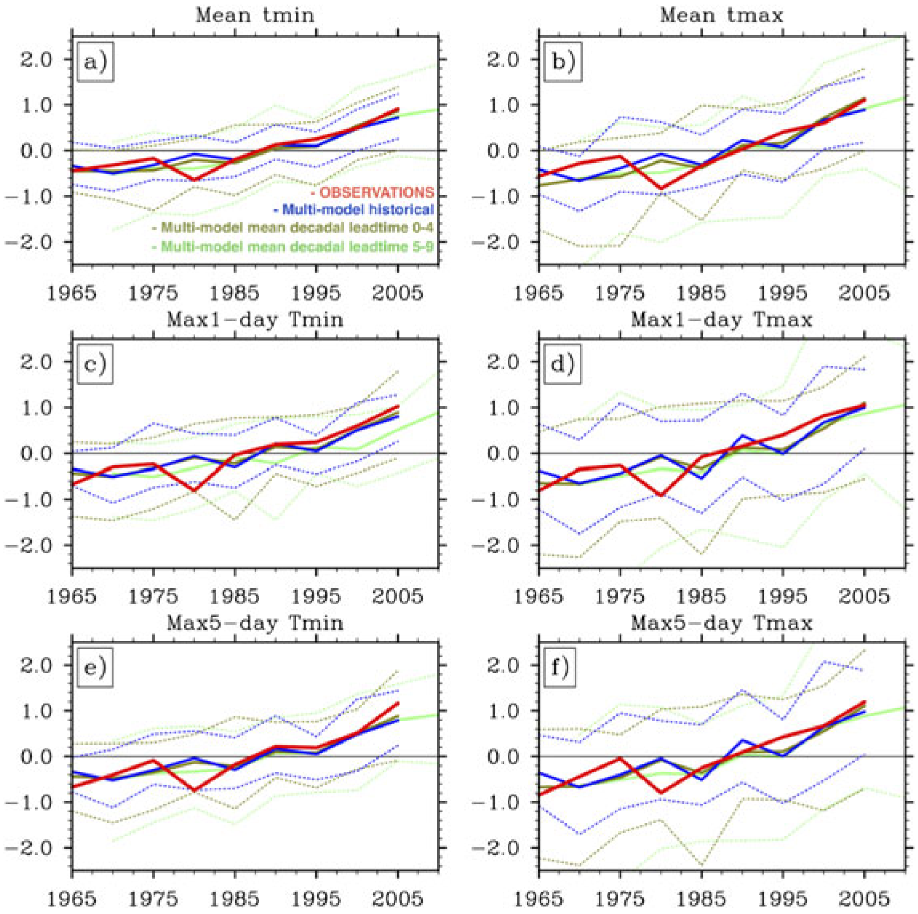

Figure 2. Five-year average time series of the magnitude of anomalies relative to the reference period 1961–2005 in (a) mean summer minimum temperature, (b) mean summer maximum temperature, (c) Max 1day Tmin, (d) Max 1day Tmax, (e) Max 5day Tmin, and (f) Max 5day Tmax across Europe. Observations are shown in red. The multimodel mean of the historical runs is shown by the blue lines. A time series consisting of the multimodel mean of the average of the first (last) 5 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.

uncertainty ranges. For all regions and all indices considered, the multimodel 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 analyzing 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.

[36] We have repeated the detection analysis with annual data (not shown) which shows very similar results to those obtained by analyzing the indices smoothed by 5 years. The only exceptions were that in contrast to the analysis based on 5 year 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 multimodel data as done here makes this result more robust to model uncertainty.

[37] 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 modeled 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

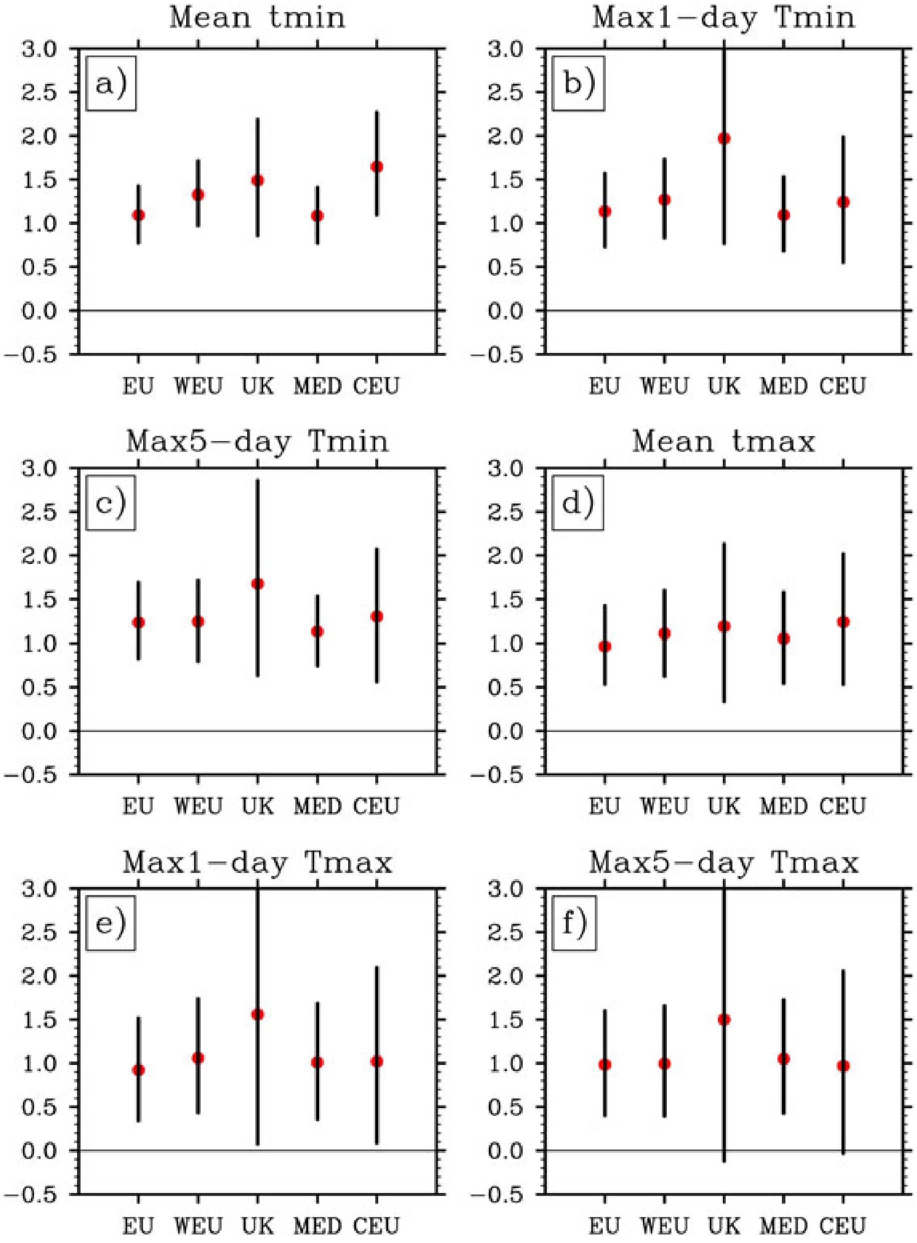

Figure 3. Scaling factors (red dots) plus 5–95% uncertainty range (vertical bars) of changes in the magnitude of (a) mean summer minimum temperature, (b) Max 1day Tmin, (c) Max 5day Tmin, (d) mean summer maximum temperature, (e) Max 1day Tmax, and (f) Max 5day Tmax across Europe (EU) and subregions, WEU, UK, MED, and CEU.

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 exceeding not only that of using climatology but also persistence, we do not further investigate persistence here.

[38] 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 lead time 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 multimodel average also varies. Further investigation could inquire as to whether excluding models with lower skill would allow for more skillful multimodel predictions than that obtained when all are included. EU is predicted skillfully at all lead times. Over the UK the predictions are only skillful for the average of the first 5 years, and MED is skillful for the last 5 years (6–9 years) of the

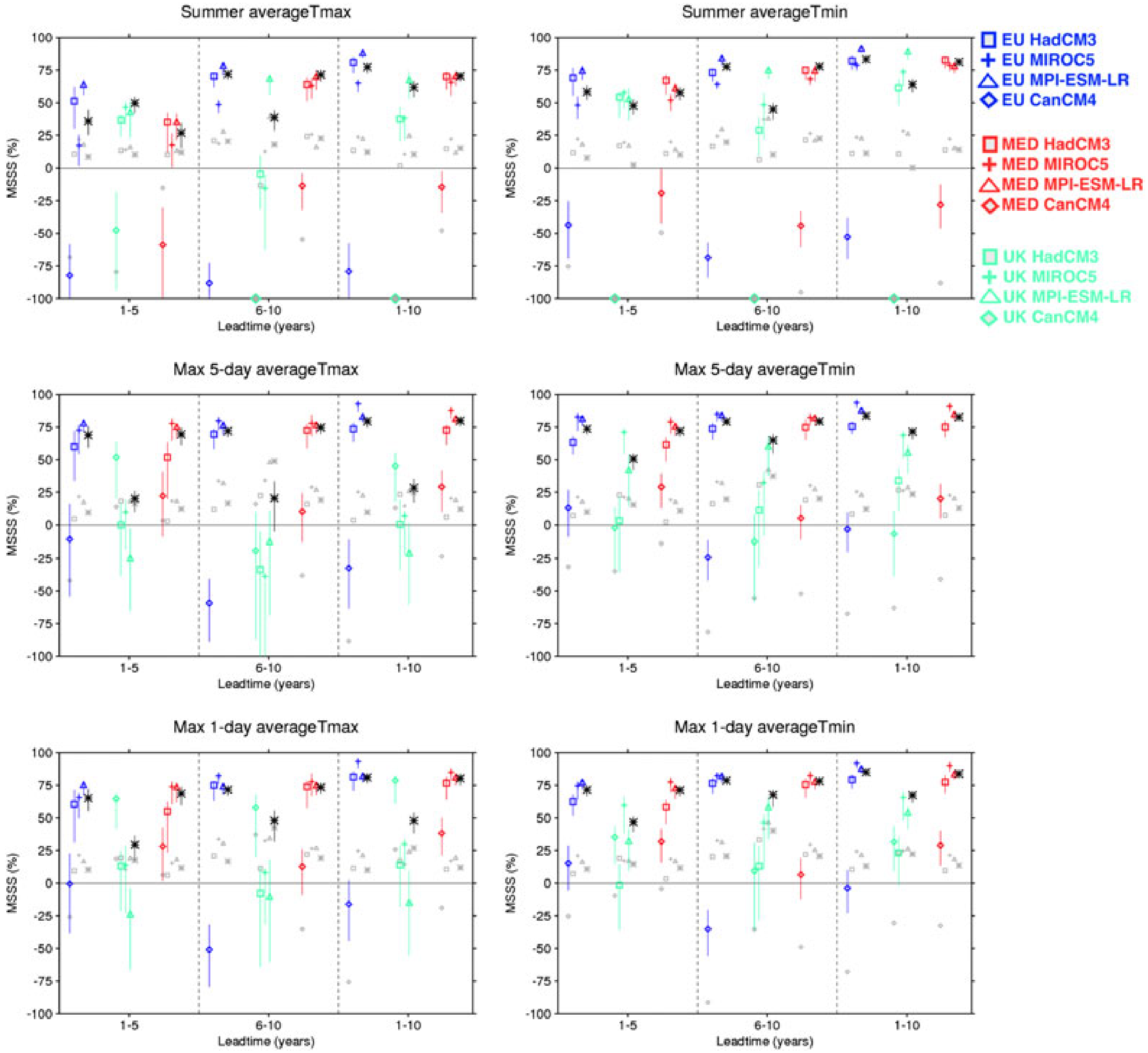

Figure 4. Mean square skill score (MSSS) of the (top) summer average, (middle) Max 5day average, and (bottom) Max 1day average Tmax (left) and Tmin (right) averaged over 5/10 years for each model (CanCM4 (diamond), HadCM3 (square), MIROC5 (cross), and MPI-ESM-LR(triangle) and the multi-model average (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, and 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 0, and to be significantly different to noise, the model MSSS must be greater than MSSS obtainable with 90th percentile of realizations 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.

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 regions. As such, the decadal trend produced is not as close to that observed and affects how skillful the prediction is. This can be seen in the time series for the UK region, shown in supporting information Figure S4, and echoes what was concluded in *Hanlon et al.* [2013] for the HadCM3 (DePreSys) model.

[39] 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 interannual variance than the observed index. As the bias correction applied has only corrected for the bias in the mean index over time, not the interannual variability, some significant bias remains. Since even small remaining biases influence the mean square error highly, this has a

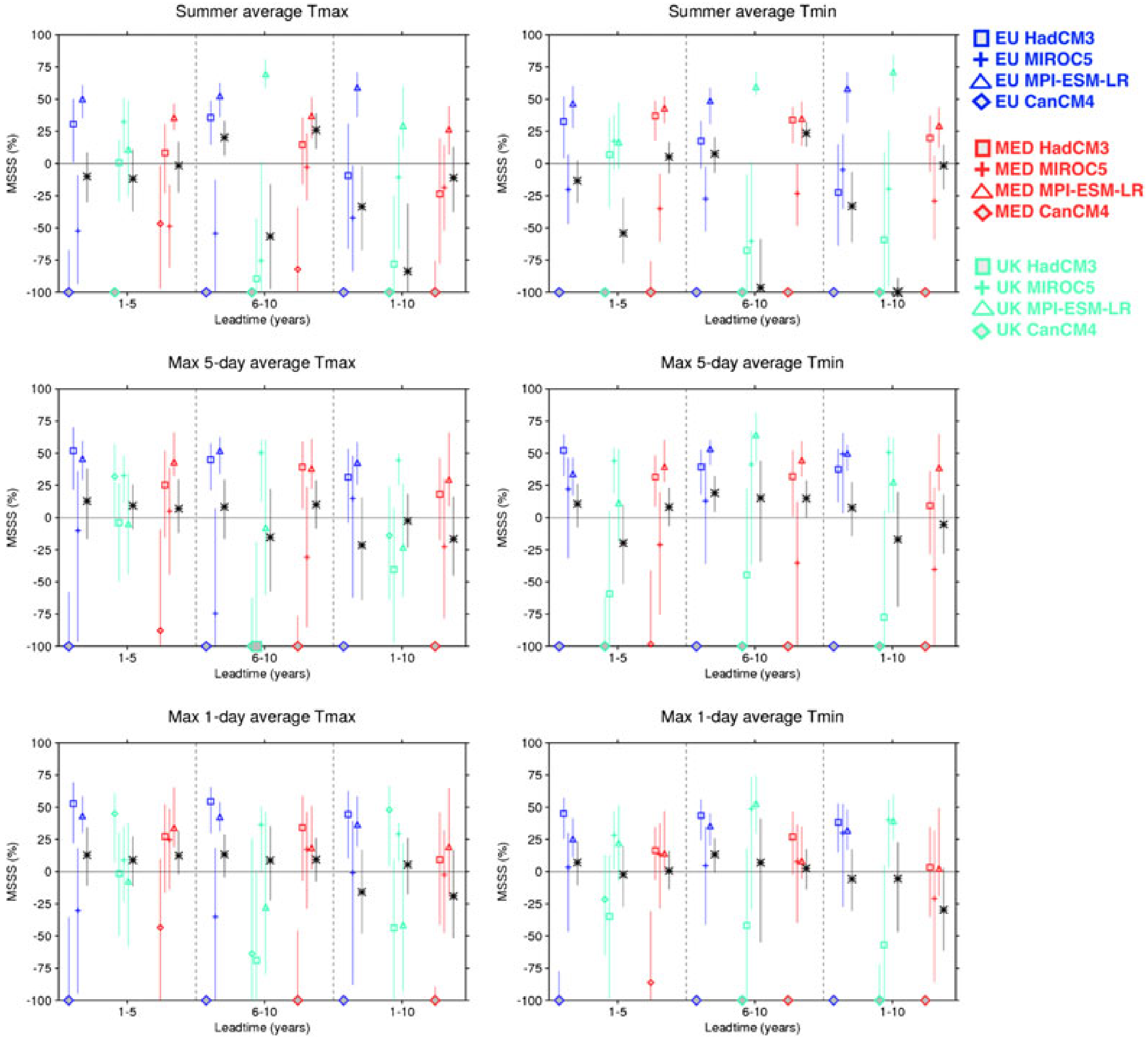

Figure 5. As in Figure 4 but the MSSS for the indices computed with decadal simulations is compared to the equivalent indices computed with the historical simulations instead of observed climatology. Positive significant skill indicates that the decadal forecasting system has higher skill than the historical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

large negative impact on the skill of the CanCM4 model and also on the skill of the multimodel 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.

[40] 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 Max 5day and Max 1day Tmin and Tmax indices (Figure 4, middle and bottom panels, respectively). This is reflected by the multimodel 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 Max 5day/Max 1day Tmin (MIROC5 and MPI-ESM-LR) and the CanCM4 decadal average Max 5day/Max 1day Tmax.

[41] The majority of models and the multimodel average indices do not show any improvement of skill of the initialized decadal runs over the historical runs which do not assimilate observations (Figure 5). There are exceptions 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.*, 2012]. As these runs were also skillful beyond climatology (Figure 4), the initialization is improving the prediction in this case. Other cases which hint at some improvement due to initialization include HadCM3 Europe average extreme indices, HadCM3 Europe 5 year average summer average Tmax, HadCM3 Mediterranean summer average, and Max 5day Tmin, MIROC5 UK Max 5day extremes, and MIROC5 UK decadal average Max 1day extremes. However, since

not all models show this improvement by initialization, the multimodel mean does not either, in general. Where the skill seen in Figure 4 is not added to by the initialization, the alternative source of skill is due to the model forcing, recreating the observed trend in temperatures over time. This could originate either 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

[42] 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 multimodel 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 subregions. 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.

[43] Analysis of the decadal simulations has confirmed this potential for skill: predictions from three out of the four 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 initialized simulations relative to the non-initialized 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 initialization of this model. The other models do not consistently show that the skill of predictions increases due to the initialization 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 multimodel mean prediction.

[44] 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, and 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 modeling 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.

[45] **Acknowledgments.** H.M.H. was supported by the UK Natural Environment Research Council through the EQUIP project (grant NE/H003525/1), and all authors H.M.H., G.C.H. and S.M. were supported by a NERC CMIP5 project. We acknowledge the E-OBS data set 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 modeling 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 (2000), Quantifying the uncertainty in forecasts of anthropogenic climate change, *Nature*, *407*, 617–620.

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

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

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

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

Christidis, N., P. A. Stott, S. Brown, G. C. Hegerl, and J. Caesar (2005), Detection of changes in temperature extremes during the 20th century, *Geophys. Res. Lett.*, *32*, L20716, doi:10.1029/2005GL023885.

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

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

Dole, R., M. Hoerling, J. Perlwitz, J. Eischeid, P. Pegion, T. Zhang, X.-W. Quan, X. Taiyi, and D. Murray (2011), Was there a basis for anticipating the 2010 Russian heat wave?, *Geophys. Res. Lett.*, *38*, L06702, doi:10.1029/2010GL046582.

Eade, R., E. Hamilton, D. M. Smith, R. J. Graham, and A. A. Scaife (2012), Forecasting the number of extreme daily events out to a decade ahead, *J. Geophys. Res.*, *117*, D21110, doi:10.1029/2012JD018015.

Efron, B., and R. J. Tibshirani (1993), *An Introduction to the Bootstrap*, Chapman and Hall.

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

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

Fischer, E. M., S. I. Seneviratne, D. Lüthi, and C. Schär (2007a), Contribution of land-atmosphere coupling to recent European summer heat waves, *Geophys. Res. Lett.*, *34*, L06707, doi:10.1029/2006GL029068.

Fischer, E. M., S. I. Seneviratne, P. L. Vidale, D. Lüthi, and C. Schär (2007b), Soil moisture-atmosphere interactions during the 2003 European summer heat wave, *J. Clim.*, *20*, 5081–5099.

Fouillet, A., G. Rey, F. Laurent, G. Pavillon, S. Bellec, C. Guihenneuc-Jouyaux, J. Clavel, E. Jougla, and D. Hémon (2006), Excess mortality related to the August 2003 heat wave in France, *Int. Arch. Occ. Env. Hea.*, *80*(1), 16–24.

Goddard, L., et al. (2012), A verification framework for interannual-to-decadal prediction experiments, *Clim. Dyn.*, *40*, 245–272.

Grize, L., A. Hussa, O. Thommena, C. Schindlera, and C. Braun-Fahrländera (2005), Heat wave 2003 and mortality in Switzerland, *Swiss Med. Wkly.*, *135*, 200–205.

Hamilton, E., R. Eade, R. J. Graham, A. A. Scaife, D. M. Smith, A. Maidens, and C. MacLachlan (2012), Forecasting the number of extreme daily events on seasonal timescales, *J. Geophys. Res.*, *117*, D03114, doi:10.1029/2011JD01654.

Hanlon, H., G. C. Hegerl, S. F. B. Tett, and D. M. Smith (2013), Can a decadal forecasting system predict temperature extreme indices?, *J. Clim.*, *26*, 3728–3744.

Hanlon, H. (2010), An investigation of causes of the 2003 heatwave in Europe using an atmospheric climate model, PhD thesis, University of Oxford.

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

Hasumi, H., and S. Emori (2004), *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 (2008), A European daily high-resolution gridded data set of surface temperature and precipitation for 1950, *J. Geophys. Res.*, *113*, D20119, doi:10.1029/2008JD010201.

Hegerl, G. C., F. Zwiers, S. Kharin, and P. Stott (2004), Detectability of anthropogenic changes in temperature and precipitation extremes, *J. Clim.*, *17*, 3683–3700.

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

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 (2007), 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*, edited by D. J. Karoly, L. Ogallo, and S. Planton, 663–745, Cambridge Univ. Press, Cambridge, United Kingdom and New York.

Jungclaus, J. H., et al. (2012), Characteristics of the ocean simulations in MPIOM, the ocean component of the MPI-Earth System Model, *J. Adv. Model. Earth Syst.*, *5*, 422–446, doi:10.1002/jame.20023/abstract.

Karoly, D. J., and P. A. Stott (2006), Anthropogenic warming of central England temperature, *Atmos. Sci. Lett.*, *7*(4), 81–85.

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

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

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

Meehl, G. A., et al. (2009), Decadal prediction: Can it be skillful?, *Bull. Am. Meteorol. Soc.*, *90*, 1467–1485.

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, and S. Polavarapu (2013), The Canadian Seasonal to Interannual Prediction System. Part I: Models and initialization, *Mon. Weather Rev.* e-View, *141*, 2910–2945, doi:10.1175/MWR-D-12-00216.

Morak, S., G. C. Hegerl, and J. Kenyon (2011), Detectable regional changes in the number of warm nights, *Geophys. Res. Lett.*, *38*, L17703, doi:10.1029/2011GL048531.

Morak, S., G. C. Hegerl, and N. Christidis (2013), Detectable changes in the frequency of temperature extremes, *J. Clim.*, *26*, 1561–1574.

Murphy, A. H. (1988), Skill scores based on the mean square error and their relationships to the correlation coefficient, *Mon. Weather Rev.*, *16*, 2417–2424.

Otto, F. E. L., N. Massey, G. J. van Oldenborgh, R. G. Jones, and M. R. Allen (2012), Reconciling two approaches to attribution of the 2010 Russian heat wave, *Geophys. Res. Lett.*, *39*, L04702, doi:10.1029/2011GL050422.

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

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

Rahmstorf, S., and D. Coumou (2011), Increase of extreme events in a warming world, *P. Natl. A. Sci.*, *108*, 17,905–17,909.

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

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

Seneviratne, S., D. Lüthi, M. Litschi, and C. Schär (2006), Land-atmosphere coupling and climate change in Europe, *Nature*, *443*, 205–209.

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

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

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

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

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

Taylor, K. E., R. J. Stouffer, and G. A. Meehl (2012), An overview of CMIP5 and the experiment design, *Bull. Am. Meteorol. Soc.*, *93*, 485–498.

Vautard, R., P. Yiou, F. D'Andrea, N. de Noblet, N. Viovy, C. Cassou, J. Polcher, P. Ciais, M. Kageyama, and Y. Fan (2007), Summertime European heat and drought waves induced by wintertime Mediterranean rainfall deficit, *Geophys. Res. Lett.*, *34*, L07711, doi:10.1029/2006GL028001.

Von Storch, H., and F. W. Zwiers (2000), *Statistical Analysis in Climate Research*, Cambridge Univ. Press, Cambridge, United Kingdom and New York.

Watanabe, M., et al. (2010), Improved climate simulation by MIROC5: Mean states, variability, and climate sensitivity, *J. Clim.*, *23*(23), 6312–6335.

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

Von Salzen, K., et al. (2013), The Canadian fourth generation atmospheric global climate model(CanAM4). Part 1: Representation of physical processes, *Atmos. Ocean*, *51*(1), 104–125.