Can a Decadal Forecasting System Predict Temperature Extreme Indices?*

HELEN M. HANLON, GABRIELE C. HEGERL, AND SIMON F. B. TETT

School of Geosciences, University of Edinburgh, Edinburgh, United Kingdom

DOUG M. SMITH

Met Office Hadley Centre, Exeter, United Kingdom

(Manuscript received 28 July 2012, in final form 20 November 2012)

ABSTRACT

Daily maximum and minimum summer temperatures have increased throughout the majority of Europe over the past few decades, along with the frequency and intensity of heat waves. It is essential to learn whether this rise is expected to continue in the future for adaptation purposes. A study of predictability of European temperature indices with the Met Office Hadley Centre Decadal Prediction System (DePreSys) has revealed significant skill in predictions of 5- and 10-yr average indices of the summer mean and maximum 5-day average temperatures based on daily maximum and minimum temperatures for a large area of Europe, particularly in the Mediterranean. In contrast, the decadal forecasts of winter mean/minimum 5-day average temperature indices show poorer skill than the summer indices. Significant skill is shown for the United Kingdom in some cases but less than for the European/Mediterranean regions.

Comparison of two parallel ensembles, one initialized with observations and one without initialization, has shown that the skill largely originates from external forcing. However, there were a few cases with hints of additional skill in forecasts of decadal mean indices due to the initialization.

Model realizations of extreme indices can have large biases compared to observations that are different from those of the mean climate indices. Several methods were tested for correcting biases, as well as for testing the significance and quantifying uncertainty of the results to rule out cases of spurious skill. Bias correction of each index individually is required as biases vary across different extremes.

1. Introduction

Throughout the majority of Europe there has been an upward trend in daily maximum and daily minimum summer temperatures over the past few decades that has been attributed in part to human influences (Christidis et al. 2012). As a result, the frequency and intensity of heat waves in Europe has also increased. Since climate model projections suggest that this rise will continue, it is important for impacts research to make accurate predictions of changes in heat wave indices. A recent study by Eade et al. (2012) demonstrated skillful predictions of moderate (1 in 10) temperature extremes on decadal time scales. Here we complement that study by assessing less moderate extremes and focusing on the prediction of summer and winter extremes for the United Kingdom and Europe. Precise predictions of future heat waves are not currently possible given the chaotic nature of the climate system, reducing our ability to make effective adaptation decisions. However, if we can effectively quantify uncertainty surrounding these predictions, this would allow an improved understanding of the risks posed by climate change and more effective planning for the future. In this study the level of precision associated with climate predictions is assessed on a decadal time scale using the Met Office Decadal Prediction System (DePreSys).

Following the devastating effect of the 2003 European heat wave (Fink et al. 2004) and 2010 Russian heat wave (Barriopedro et al. 2011), there have been many studies of the seasonal predictability of summer temperature

* Supplemental information related to this paper is available at the Journals Online website: http://dx.doi.org/10.1175/JCLI-D-12-00512.s1.

Corresponding author address: Dr. H. M. Hanlon, School of Geosciences, University of Edinburgh, Grant Institute, The Kings' Buildings, West Mains Road, Edinburgh, EH93JW, United Kingdom. E-mail: h.hanlon@ed.ac.uk

DOI: 10.1175/JCLI-D-12-00512.1

extremes. Seasonal predictability of summer 2-m temperatures in Europe has been found to be less skillful than elsewhere in the world because of the difficulty of model simulation of blocking systems, which is a common cause of heat wave events in this region (Palmer et al. 2008). Following model improvements, Weisheimer et al. (2011) found it possible to make skillful predictions of seasonal mean 2-m temperature, particularly at the upper tail of the distribution, suggesting that some models can perform well for hot extremes. The reason for this is that the warming trend in southern European land temperatures is captured well by seasonal forecasts.

Another possible source of predictability arises from the link between dry springtime soil moisture levels, leading to higher summer temperatures and the occurrence of heat waves. In a study of the effect of soil moisture initial conditions on the summer climate predictability, Conil et al. (2009) found that realistic soil moisture boundary conditions are necessary to predict summer climate anomalies across Europe. This is thought to be due to feedback mechanisms involving land surface–atmosphere interactions whereby the coupled temperature and drought conditions are amplified (Seneviratne et al. 2010). For example, low soil moisture levels in spring lead to reduced evaporation, preventing cloud formation, which allows higher levels of insolation to further warm and dry out the land surface (Fischer et al. 2007; Seneviratne et al. 2006). Vautard et al. (2007) found a link between European summer heat waves (based on daily maximum temperatures) and wintertime precipitation deficit in southern Europe, which may be a useful source of seasonal predictive skill in that area. Similarly, Quesada et al. (2012) determined that the number of hot days (number of days with daily average surface temperature above the 90th percentile) was related to low winter–spring rainfall frequency in southern Europe; however, farther north this relationship was not so robust. This difference between northern and southern Europe is explained by Teuling et al. (2009), who analyzed the main external drivers of evapotranspiration. In northern–central Europe the evapotranspiration shows a greater correlation with radiation than more southerly European regions, which are correlated more with precipitation and therefore soil moisture availability. Hence the seasonal predictive skill obtained from observing a dry winter–spring is useful in southern Europe, but farther north, where the mechanisms driving evapotranspiration are different, this will not be so useful. So it is important to consider the predictive skill assessed over regions within Europe where the driving mechanisms are similar. A possible exception could be the response to different emissions of anthropogenic aerosols, but this is not studied here.

When considering the impact an extreme event, it is often daily maximum (Tmax) or daily minimum temperatures (Tmin) that have a large impact on society than the daily average. Impacts of these extreme temperatures can vary greatly depending on location and the vulnerability of a particular subject. Many studies concerned with the effect of extreme temperatures on human health have focused on the daily extreme temperatures (Tmax and Tmin) rather than daily means (D'Ippoliti et al. 2010; Fouillet et al. 2006; Pascal et al. 2006; Díaz et al. 2006). High daytime temperatures can cause hyperthermia and heat stroke and the body can also be put under additional heat stress by high nighttime temperatures, preventing those suffering heat-related illnesses from recovering during the night. Hence, high nighttime temperatures (High Tmin) can be just as important when considering health effects as high daytime temperatures (Tmax) [as shown by Grize et al. (2005) for heat wave events in Switzerland]. Hamilton et al. (2012) found skill in predicting the number of daily extreme temperatures (Tmin and Tmax outside the 10%–90% range of the distribution in a season), particularly in Northern Hemisphere summer, although this was lower than the skill in predicting the seasonal mean especially in the extratropics.

Considering predictions further into the future—for example, the near-term future (10–20 years ahead)—Meehl et al. (2009) show that pattern and magnitude of surface temperature change is similar for different emissions scenarios. In general, it is not until the latter half of the twenty-first century that different emission pathways cause significantly different temperature responses (Solomon et al. 2007; Hawkins and Sutton 2009). Hence near-term climate prediction is not expected to be too affected by scenario uncertainty.

Evidence of decadal prediction skill resulting from anthropogenic external forcing is shown in Lee et al. (2006). Initialization of models with observations may improve skill both by predicting natural internal variability and by correcting the model's response to previous external forcing factors (e.g., Smith et al. 2010). Idealized studies (e.g., Branstator et al. 2012; Branstator and Teng 2012) suggest that internal variability may be predictable up to a decade ahead in the North Atlantic. In reality there are many technical problems to overcome, especially the lack of subsurface ocean observations for initialization and how to deal with model biases. Nevertheless, several studies have demonstrated improved skill through initialization for SST in the North Atlantic (Keenlyside et al. 2008; Pohlmann et al. 2009; Smith et al. 2010; Chikamoto et al. 2013; van Oldenborgh et al. 2012). However, the impact of initialization on predictions of atmospheric conditions over land is less convincing.

In this paper, a methodology is outlined by which we compute decadal predictions of extreme temperature indices from hindcasts performed with the Met Office Decadal Prediction System and assess the accuracy of these predictions. This assessment is based on two ensembles of hindcasts, one initialized with observed values and one without an assimilation of observations. The ensembles and observed data are discussed in section 2. Section 3 outlines methods and uncertainty estimates. The two predictions are then compared in section 4 to determine whether initializing the model with observations improves the accuracy of the prediction.

2. Data

a. Observations

The observations used to compare the model against are taken from the ENSEMBLES project (http://ensembles-eu.metoffice.com/index.html) observational database (Eobs), which is a high-resolution (0.5° latitude × 0.5° longitude grid) gridded dataset of observations (Haylock et al. 2008). It should be noted that Eobs is a gridded dataset of observations where the observations from individual stations have been interpolated onto a regular grid. As a consequence, there will still be some uncertainty in these observations due to measurement errors, variations in the density of stations, and interpolation methods. However, as there is a large number of stations across the area this dataset covers and we have further regridded this data to the lower resolution of the DePreSys model, these errors are assumed to be small. To make these data directly comparable with that of the model, the Eobs data were first regridded using area-averaging interpolation to the horizontal resolution of the model grid (3.75° longitude × 2.5° latitude), and then the grid points defined as sea in the DePreSys land mask were set as missing values and the value for 29 February in leap years was removed. Finally, as the DePreSys model calendar is 360 days and we require winter [December–February (DJF)] and summer [June–August (JJA)] daily data, the calendar was converted from 365 to 360 days by removing the 31st day of any month with 31 days (January, July, and August) and inserting 1 and 2 March as the 29th and 30th days of February.

b. Model

The Met Office built a decadal prediction system (DePreSys) (Smith et al. 2007, 2010) utilizing the third climate configuration of the Met Office Unified Model (HadCM3) (Pope et al. 2000; Gordon et al. 2000). The atmospheric component of this model has a horizontal resolution of 3.75° longitude by 2.5° latitude and 19 levels in the vertical up to a height of 40 km. This is coupled to an ocean component that has a horizontal resolution of 1.25° longitude by 1.25° latitude and 20 levels in the vertical.

Retrospective forecast experiments (known as hindcasts) were performed by the UK Met Office Hadley Centre as a contribution to the EU ENSEMBLES project (van der Linden and Mitchell 2009) and consist of an ensemble of nine members. Each member uses a different variant of HadCM3 obtained by perturbing poorly constrained parameters in the model physics schemes. The parameter perturbations were selected from a set of 128 model variants created by applying different combinations of perturbations to 29 parameters that control subgrid-scale atmospheric and surface processes. The perturbations are described in more detail in Murphy et al. (2004) along with an analysis of which parameters are most related to the uncertainty in global climate sensitivity. Of these 128 HadCM3 model variants, eight of them, along with the unperturbed model, are selected for use in the DePreSys hindcasts. These variants span a wide range of climate sensitivity, from 2.6° to 7.1°C, in order to sample model uncertainty.

The DePreSys hindcasts were initialized every November from 1960 to 2005 and run for 10 years. Both the atmosphere and ocean components of the system were initialized with values calculated as anomalies from the observed climatology added onto the model climatology, in order to reduce model drift after assimilation (Smith et al. 2007). The climatological period used to compute the anomalies is 1958–2001 for the atmosphere and 1951–2006 for the ocean. The atmospheric anomalies were taken from the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40; Uppala et al. 2005) and ECMWF operational analysis, while ocean anomalies were taken from a Met Office Hadley Centre ocean analysis (Smith and Murphy 2007) but updated to produce a better fit to observations. These anomalies were then assimilated into an integration of each of the perturbed model variants, run from December 1958 to November 2007, producing initial conditions for each start dates. The initialized hindcasts are referred to as the perturbed physics ensemble (PPE) forthwith.

A parallel set of uninitialized hindcasts, referred to as NoAssim, was performed alongside the PPE hindcasts to allow a diagnosis of whether the initialization has improved the forecasts. Each individual member of the NoAssim ensemble is performed with the same model variants as the corresponding member of the PPE ensemble but without assimilation of the observed state of the atmosphere or ocean from the analyses as performed for the PPE ensemble.

3. Methodology

Several indices are computed to characterize the average and extreme temperatures in Europe, which measure large-scale heat wave and extreme cold events. The indices are calculated for each grid point over land for each model run and the regridded observations.

To define a heat wave or extreme cold event it is often necessary to look beyond the seasonal mean of these daily extremes. We define an index that can provide information on a shorter time scale than the seasonal mean along with a measure of the intensity of extreme temperatures. It is the maximum (minimum) 5-day average and is calculated by taking a rolling 5-day average throughout the year and finding the maximum (minimum) value. It has the benefit of being less noisy than the annual maximum (minimum) while still allowing the extremity of the hottest temperatures of a given year to be seen, and is almost as detectable as seasonal mean changes (Hegerl et al. 2004).

a. Indices

The following indices have been computed and analyzed throughout this study:

- winter average minimum temperature (WTmin): the mean average daily minimum temperature computed over the winter season January–March,
- winter average maximum temperature (WTmax): the mean average daily maximum temperature computed over the winter season January–March,
- summer average minimum temperature (STmin): the mean average daily minimum temperature computed over the summer season June–August,
- summer average maximum temperature (STmax): the mean average daily maximum temperature computed over the summer season June–August,
- minimum 5-day average Tmin (Min5day-Tmin): the lowest 5-day mean average daily minimum temperature that occurred between 1 January and 30 December (using a 360-day calendar),
- minimum 5-day average Tmax (Min5day-Tmax): the lowest 5-day mean average daily maximum temperature that occurred between 1 January and 30 December (using a 360-day calendar),
- maximum 5-day Tmin (Max5day-Tmin): the highest 5-day mean average daily minimum temperature that occurred between 1 January and 30 December (as on a 360-day calendar), and
- maximum 5-day Tmax (Max5day-Tmax): the highest 5-day mean average daily maximum temperature that occurred between 1 January and 30 December (as on a 360-day calendar).

The seasonal mean indices are found by taking the mean average over all daily values in the summer–winter months. The maximum and minimum 5-day averages are calculated by taking each day of the year and finding the mean of a 5-day period surrounding that day (2 days on either side) and calculating the maximum and minimum values of that 5-day average for each year. It is expected that the maximum value will occur during a summer month and the minimum during the winter but that is not necessarily the case, so we do not restrict this calculation to particular seasons.

These extreme temperature indices are computed for both the PPE and NoAssim decadal runs and are bias corrected compared to the 30-yr antecedent observational climatology as described below. Then a mean square skill score as described in Murphy (1988) and discussed in Goddard et al. (2013) is used to assess the skill at individual lead times and at 5- and 10-yr lead time averages. In the case of individual lead times the indices computed for summers at different positions throughout the run are assessed separately to show how skill varies with lead time during a 10-yr run. Then 5- and 10-yr averages of each index are computed from each run as a prediction of the average for the next 5–10 years and compared to the corresponding 5- and 10-yr averages from the observations, to assess how skillful these semidecadal–decadal predictions are. Because of the period of the observations and the model runs, the skill is assessed for the runs starting from 1980 to 2000 inclusive, for which we have both a 30-yr prior observational climatology and concurrent observations with which to compare all lead times.

b. Regions

After the index is calculated a regional average over land grid points is performed, as we are concerned with the predictability of large-scale heat wave events, similar in magnitude to the 2003 European event. The area of interest for this study is Europe (35°–65°N, 10°E–40°W). Along with this region we also compute averages for two subregions, the British Isles (50°–60°N, 10°E–2°W, hereafter termed ''UK'' but this does still include Ireland) and the Mediterranean (35°–50°N, 10°E–40°W). We also performed this analysis for a central Europe (42°–55°N, 2°E–20°W) region; however, the results for this region were very similar to that of the Europe region so they are not shown here. In addition to these regions, the northern Europe region was also considered (50°–65°N, 10°E–40°W), and the results for this region are shown in the supplementary material (available at the Journals Online website: http://dx.doi.org/10.1175/JCLI-D-12-00512.s1).

c. Correlation of DePreSys with observations

In general, heat wave indices predicted by the model will contain biases relative to the observations, and these must be corrected to obtain a forecast. Since bias correction could potentially affect the forecast skill, we first assess the Pearson correlation coefficient for the summer average values of Tmax and Tmin (STmin, STmax) before any bias correction. For this we compare the DePreSys PPE and NoAssim ensembles to Eobs observations over the years 1960–2000. The significance of these correlation coefficients was determined by calculating a lower threshold for the correlation coefficient that captures the 95th percentile of the null hypothesis of no linear relationship. This threshold was calculated to be 0.312 using the two-tailed Student's t test where the number of degrees of freedom is assumed to be equal to the number of years minus 2 and is displayed on the plots by a gray shaded region. This neglects autocorrelation due to internal variability, which is small over land regions. The uncertainty within the ensemble was estimated by bootstrapping with replacement to obtain a number of realizations of the ensemble mean (Efron and Tibshirani 1993, chapter 6). Each of these realizations was used to calculate the correlation coefficient of that realization with observed values. This produced a range of possible values of this correlation coefficient. The 10%–90% interval of this range was taken as the uncertainty on the correlation due to ensemble spread and is shown as vertical bars through each point.

d. Bias correction

The bias correction method applied is in line with the guidelines set by the World Climate Research Program (WCRP) for anomaly initialized model results [for details, see WCRP (2011)]. Since the model was initialized with an anomaly compared to climatology, as opposed to being initialized with observed values, the forecasts are less likely to drift. As such, the correction applied is only required to adjust the forecasts back to observed climatology, without accounting for drift, and is calculated as in Eq. (1):

$$\text{model} = \text{model} - \text{model climatology} + \text{observed climatology}. \quad (1)$$

The bias in temperature extremes is not the same as the bias in the seasonal mean as a result of processes influencing the frequency of extremes that are different to those affecting the mean. The reason for this is that in models, these processes (usually small-scale parameterized processes or local feedbacks that are influential on extremes) are not always as well captured as those that govern more large-scale processes that determine mean climate. Bias correcting daily data and then averaging to find a seasonal average leads to a good correction for the seasonal mean data. For an extreme index (e.g., Max5day-Tmax and Max5day-Tmin) this is not appropriate as the indices we compute are measures of extremes that can occur on different calendar days. Applying a different bias correction for each calendar day individually does not solve this problem because daily biases cannot be diagnosed accurately enough. It would also inject more variability in the computed index. We find that bias correcting the index, rather than the raw daily data, successfully removes almost all of the mean bias between the observed index and the modeled index and is therefore the appropriate bias correction method for this purpose. This is more important for the extreme indices than seasonal averages. We apply the same method across both indices. The effect of different bias corrections is illustrated in Fig. 1, where the time series of the Max5day-Tmax and Max5day-Tmin are computed with uncorrected data, data with a daily mean bias removed from each daily value before the index is calculated, and, finally, the data, corrected with a 30-yr prior climatology after the index, is computed.

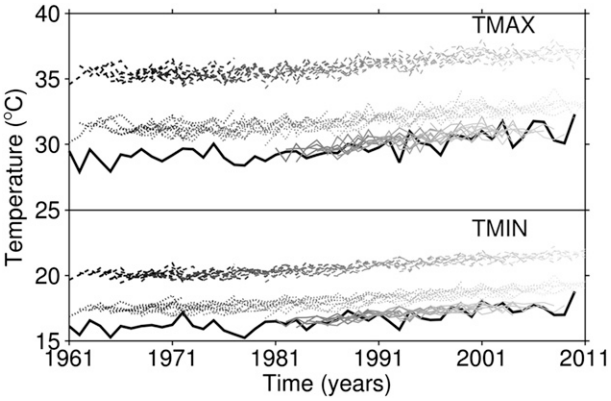

FIG. 1. Time series of (top) Max5day-Tmax and (bottom) Max5day-Tmin averaged over Europe from the DePreSys (PPE) ensemble compared to Eobs observations (thick solid black lines). Each line shows the annual index calculated for each year of the 10-yr runs, started every year. As such the time series overlap, so the lines are shaded progressively lighter for runs with later start dates. The three time series shown are computed with the uncorrected data (dashed lines), with the bias correction performed after the index calculation (solid lines), and with bias correction applied to the daily Tmax and Tmin data (dotted lines).

The model climatology of the index for bias correction is taken from transient runs of the same model and the observed climatology is computed from the Eobs dataset. The climatology here is defined as the 30-yr mean of the index prior to the start of the run (i.e., the 1980 run is

corrected with climatologies averaged over 1950–79). The benefits of using a prior climatology are that the data are not preconditioned on observations that occurred during the in-sample time period, allowing this correction to be applied systematically to runs that extend into the future.

The DePreSys decadal forecasts are created using a perturbed physics ensemble, so each member of this ensemble must be considered as a different model. To account for this, the mean bias between the modeled index and the observed index is removed for each member of the ensemble separately. The correction applied remains constant across different lead times.

e. Calculation of MSSS

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. To compare the accuracy of predictions made with two different methods we use the mean square skill score (MSSS) (see Murphy 1988). It compares the mean square errors between each forecast with the observations. This skill score was used to estimate how accurately the DePreSys hindcasts (PPE) recreate the corresponding observed values of the regional average indices, compared to Eobs observational climatology. It is also used to test if the uninitialized runs (NoAssim) are any more skillful than the initialized PPE runs. The method of using the mean square skill score to assess decadal predictions has been evaluated by Goddard et al. (2013), who has also laid out best practice guidance for use with this method:

$$\mathrm{MSE}(f,x) = \frac{1}{n}\sum(f_i - x_i)^2 \quad \text{and} \tag{2}$$

$$\mathrm{MSSS}(f,y,x) = 1 - \frac{\mathrm{MSE}(f,x)}{\mathrm{MSE}(y,x)} = 1 - \left[\frac{\frac{1}{n}\sum(f_i - x_i)^2}{\frac{1}{n}\sum(y_i - x_i)^2}\right], \tag{3}$$

where MSE is the mean square error, f_i is the ith forecast, x_i is the ith observed value, y is the reference forecast, and n is the number of forecasts (here the forecasts are the 10-yr runs started each year). The forecasts here refer to the regional average of the indices described above. We perform the regional average before the skill is assessed because the skill in these indices (especially the more extreme indices) can vary greatly over the larger regions and as a consequence the average of a noisy spatial pattern of skill is less meaningful than the skill of the regional average.

A skillful prediction is considered to be a forecast that is closer to the observed value than our observed climatology (here the average of the previous 30 years before the start of the run). A prior climatology is deemed to be a reasonable benchmark in this case as extreme temperature indices can randomly fluctuate year to year as they are affected by weather variability. This benchmark also has the advantage that it could be used as an alternative method of prediction as it does not require any knowledge of the future. Comparing to an in-sample climatology does not have this advantage because one needs to know the value of the index for the entire period tested, so it could not be used as a method of prediction. Hence the 30-yr prior climatology is used as the reference forecast y.

This is repeated using the NoAssim forecast as the reference forecast [as in Eq. (4)] and the results used to determine whether PPE is more skillful than NoAssim (termed noa in the equations). The reason for computing the difference in skill with this method, as opposed to repeating the computation in Eq. (A1) for NoAssim, is to remove 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. The difference in skill between the two ensembles shows how much more skill the ensemble that assimilates observations has over the unassimilated runs that had no initial knowledge of the observed state of the climate:

$$\text{ppe/noa Skill Difference} = 1 - \frac{\mathrm{MSE}(\mathrm{ppe},x)}{\mathrm{MSE}(\mathrm{noa},x)} = 1 - \left[\frac{\frac{1}{n}\sum(\mathrm{ppe}_i - x_i)^2}{\frac{1}{n}\sum(\mathrm{noa}_i - x_i)^2}\right]. \tag{4}$$

The MSSSs [Eqs. (3) and (4)] are also calculated for 5- and 10-yr averages of the annual indices. The purpose of this was to determine whether long-term averages show more skill than predictions of individual years.

f. Estimation of error on the MSSS

The MSSS is computed from the ensemble average of the regional average of a given index at each lead time for a particular year. Uncertainties arise from the limited ensemble size of nine members. An estimate of this sampling uncertainty is made using bootstrapping with replacement (Efron and Tibshirani 1993, chapter 6). For each realization, nine members of the ensemble are drawn with replacement, from the entire nine-member

ensemble. Then the same computations are done on the sample as performed for the ensemble average. A thousand samples are generated and the 10%–90% range from these provides the error on MSSS. Although this error estimate is biased low, being cautious we use it. If the score and its error are above zero then the forecast (whether the forecast from the PPE ensemble or the NoAssim ensemble) has more skill in predicting the index than the 30-yr prior observed climatology.

An additional method of estimating uncertainty is to compare a random forecast, which should have no significant skill apart from coincidence, with climatology. A random forecast is generated assuming a different normal distribution for each period (seasonal, 5-yr average, and 10-yr average), PPE member, and index. The mean and standard deviation for the normal distribution is estimated from each member of the perturbed-physics and NoAssim ensembles separately and used to normalize the random forecast. One thousand 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 MSSS is considered not significantly better than random noise; this is shown on the figures as a gray point.

4. Results

It can be seen that the summer average indices (STmax and STmin) in the modeled regions, Europe, and the Mediterranean (MED) are significantly correlated with observations at almost all lead times (Fig. 2). This shows that the changes in the modeled temperatures are recreating the observed changes reasonably well in those regions. The UK does not show significant correlations for STmax, suggesting that the changes in the maximum daily temperatures are not captured well by the model over the UK, whereas the STmin does show significant correlations at some lead times, albeit not consistently. This suggests that UK summer daily maximum temperatures are not as predictable as those in other European regions and also not as predictable as UK summer daily minimum temperatures.

The correlation coefficient does not seem to increase or decrease as lead time increases, suggesting that model drift is not having much of an effect on the relationship between modeled and observed temperatures on this 10-yr time scale. However, the correlation coefficients do not decrease with lead time, which suggests that the initialization of the model is not having a noticeable influence on the correlation, since otherwise the strength of the correlation would reduce as the model evolves away from the initial state.

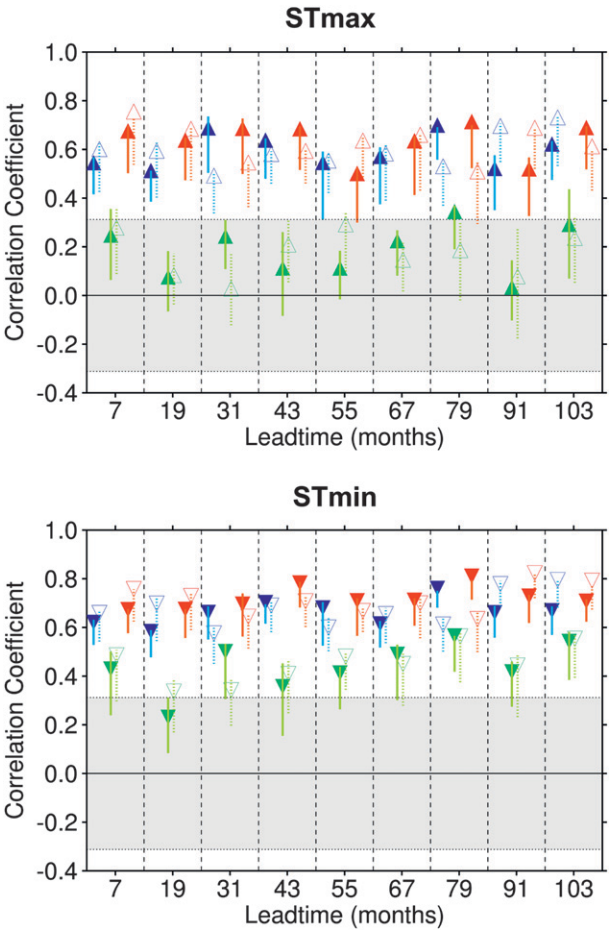

FIG. 2. Correlation of the summer average indices (top) STmax and (bottom) STmin from PPE (solid) and NoAssim (dotted) compared to observations. These are regionally averaged indices for Europe (blue), the Mediterranean (red), and the UK (green). The gray shaded region shows the values of correlation coefficient that do not show a statistically significant correlation (at the 95% level). These indices are computed with the uncorrected model data from runs starting in 1960–99 at each lead time and the correlation with the same indices computed for the corresponding years from the observational record (Eobs).

Are these correlations, found for the summer average indices (STmax and STmin), due to similar trends in the data? The decadal average STmax and STmin over 1961–2000, as calculated from Eobs observations, have increased over the 40-yr period in almost all areas (Fig. 3, top panels), with larger warming trends in the more southerly regions. This widespread increase is also seen in the 10-yr averages from the modeled results for both ensembles (Fig. 3, middle and bottom panels). There is also a similar north–south contrast in the modeled trends, as seen in the observed trend, although the magnitude of the trends in the south is not so large.

It is not reasonable to assess the skill of predictions of individual grid points because of a low signal-to-noise

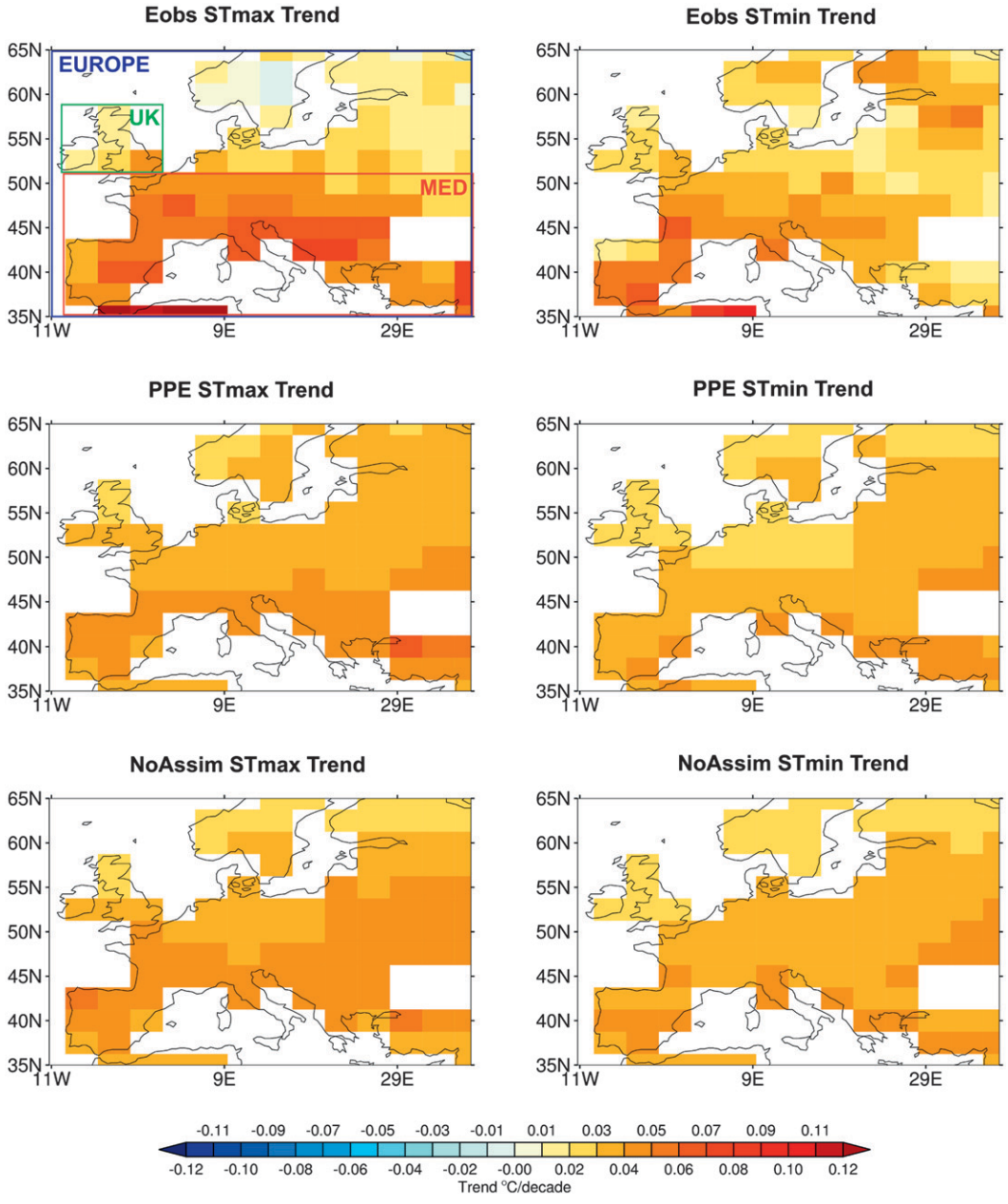

FIG. 3. Trend in decadal average (average over lead times 1–10 yr) of summer average indices (left) STmax and (right) STmin from (top) Eobs observations over Europe (land only) over 1961–2000 compared to the same period in the (middle) DePreSys and (bottom) NoAssim. The temporal average is performed over the values at different lead times that correspond to the same period (1961–2000). To calculate the trend in the indices with time, an ordinary least squares linear regression of the temperature index against time was performed for each grid point. The colored boxes in the top left panel define the regions used to calculate the regional averages of the indices. To view modeled trends as anomalies from the observed trend, see supplementary Fig. 1.

ratio of these indices at small spatial scales, and because the spatial pattern in the observed trend is not recreated well by the model across all indices, for example the Max5day-Tmax and Max5day-Tmin (Fig. 4). Instead, a regional average of these indices is made to capture the trends in indices on a larger spatial scale. Nevertheless, similarities between the modeled and the observed spatial patterns of these trends show there is greater warming in the more southerly regions (i.e., the Mediterranean) than in the north, in both observations and the modeled ensembles. Also, the decadal average trends of Max5day-Tmax and Max5day-Tmin (Fig. 4, middle

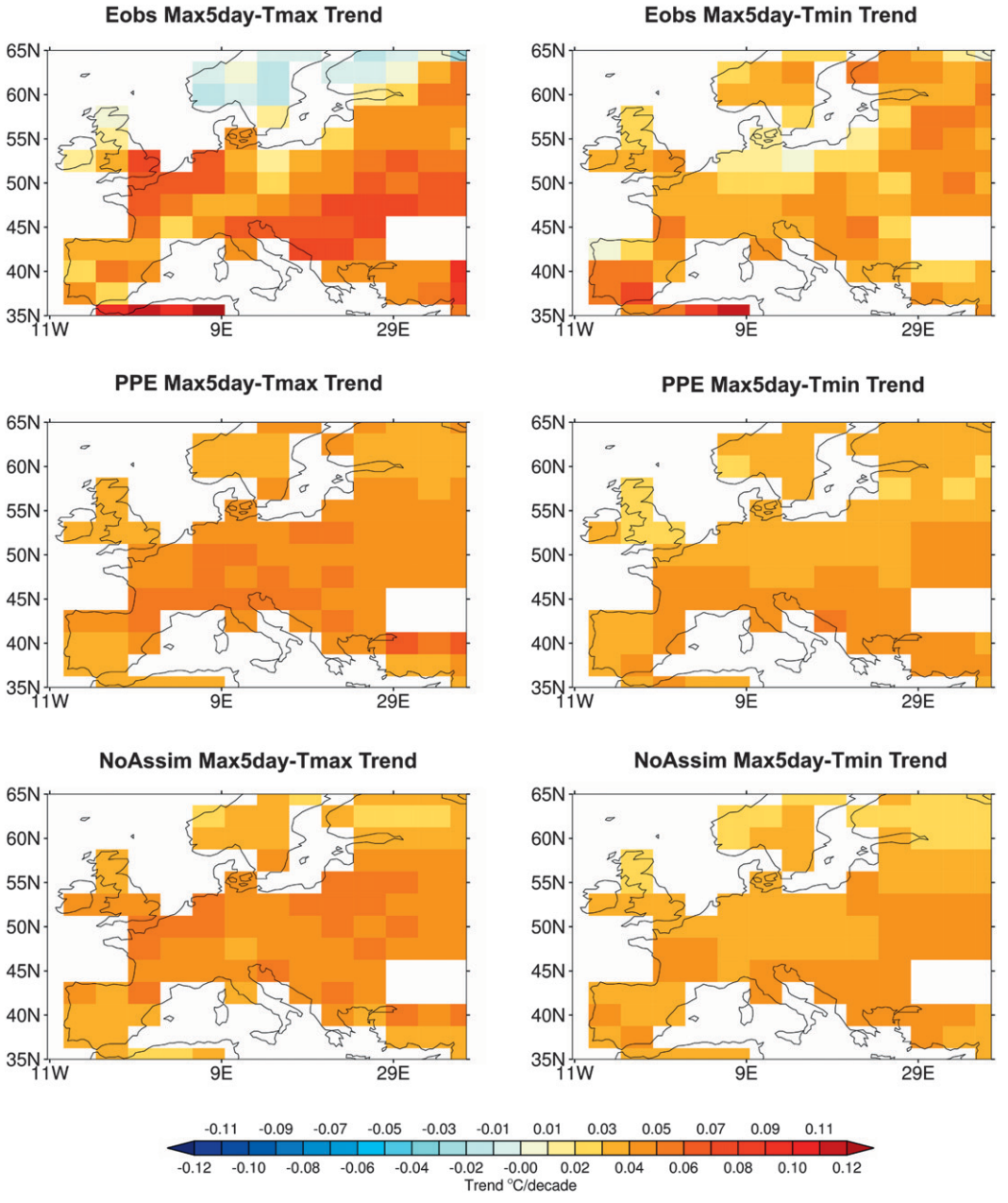

FIG. 4. As in Fig. 3, but for the maximum 5-day average index Max5day-Tmax and Max5day-Tmin. To view modeled trends as anomalies from the observed trend, see supplementary Fig. 2.

and bottom panels) show similar features to that of the decadal average STmax/STmin trends; they are positive throughout most of Europe but are slightly greater in magnitude than the summer average indices. There is also more variation between PPE and NoAssim for Max5day-Tmax.

A comparison of the decadal trends in the initialized (DePreSys) and the uninitialized (NoAssim) runs show similar spatial patterns of decadal average trends in summer average temperatures (see middle and bottom panels of Figs. 3 and 4).

Figures 5 and 6 show that the trend in the model seems to follow the observations well and the spread encompasses most of the year-to-year variability. The UK indices are the exception, as the trend is somewhat overestimated, especially STmax. The time series shown in Figs. 5 and 6 also include the decadal average predictions for subsequent decades, the latest being the 2006–15 decadal average.

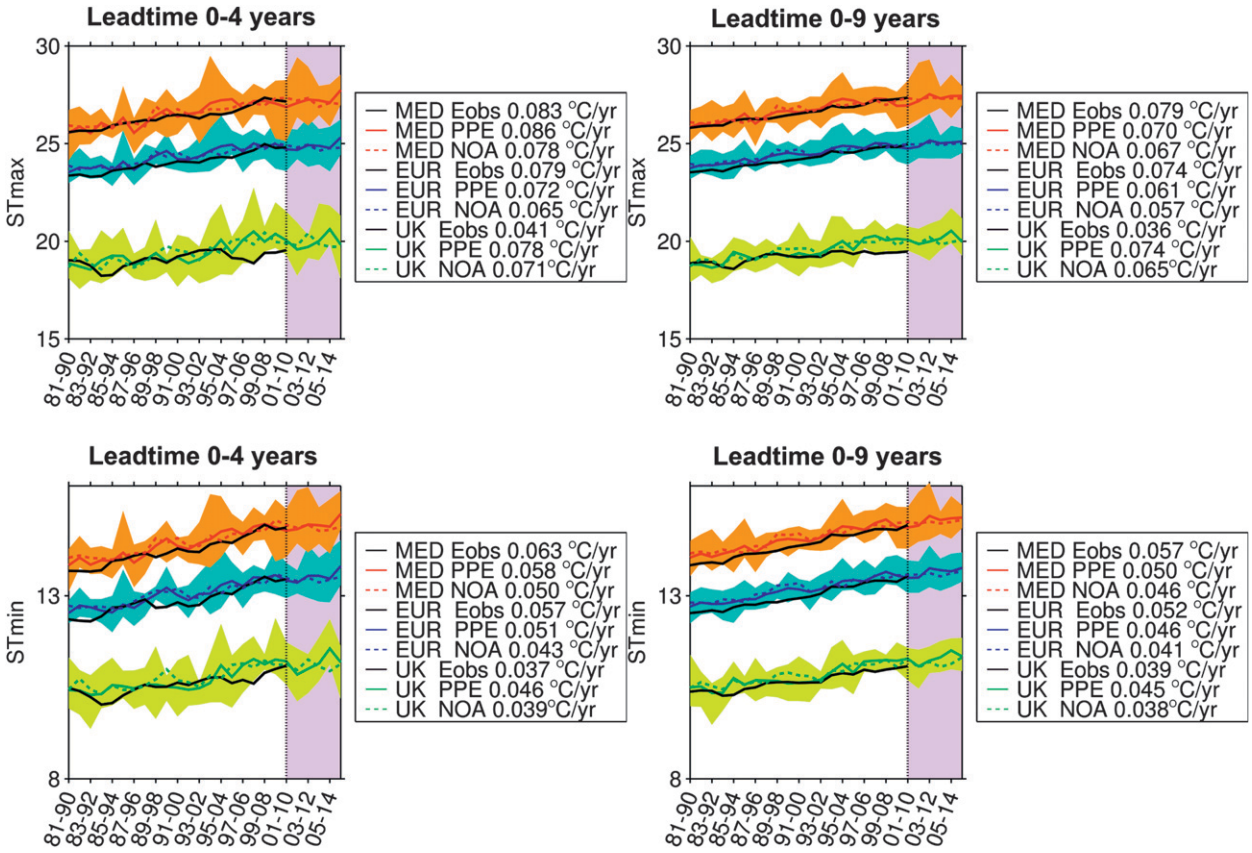

FIG. 5. Time series showing the mean and spread of summer average indices (top) STmax and (bottom) STmin from runs starting in 1981 to 2000, from the PPE (solid colored lines) and NoAssim (dotted colored lines) ensembles after bias correction for (left) 5- and (right) 10-yr lead time averages and compared to Eobs observations (black solid line). The lead times are averaged over 5- and 10-yr periods. The time series shown are the average over the first five summers of each run in the left panels and the decadal summer average of the entire 10-yr run in the right panels. To ensure that the observed values correspond to the same average of years as shown for the modeled results, a 5-yr smoothing was applied to the observations for the time series in the 0–4-yr lead time plot and 10-yr smoothing applied to the observed time series in 0–9-yr lead time plot. The colored shaded region represents 10%–90% range of the ensemble spread. This is shown for each region: Europe (blue), the Mediterranean (red), and the UK (green). The lilac region shows further decadal forecasts not included in the skill score analysis.

a. Is there skill beyond climatology?

Where the MSSS is above zero this means there is more skill in the model than in the observed prior climatology; in other words, the modeled forecasts are closer to the actual observations than the average of the 30 years prior to the start of the runs. Alternatively, if the MSSS is negative this means that Eobs 30-yr prior climatology is closer to the observed value than the forecast and therefore there is negative skill in the model. It should be remembered that this measure of skill is purely the ability of one method beyond another to predict the index. Even if skill is found in predictions it can still be a poor prediction, so the skill scores should be considered alongside analysis of the time series, trends, and spread to ensure the predictions are appropriate for use (as shown in Figs. 5 and 6).

To test the statistical significance of the skill scores, the error bars are calculated using the 10%–90% range of a bootstrapped estimate of ensemble variability and are shown by vertical lines through each MSSS value (for details of the bootstrapping method, see section 3f). If the error bar crosses zero the result is deemed to not be significant (90% level), since existence of skill beyond the benchmark is uncertain for a different combination of ensemble members (see section 3f for more details). The MSSS is also tested for significance beyond that of a random forecast. The MSSS must be above the shaded region that represents the 90% range of the MSSS obtained from 1000 realizations of random noise with equivalent variability to that of the model. So, the MSSS value and the associated error must be above zero and above the shaded region (determined by the random forecasts) for the model to be deemed skillful compared

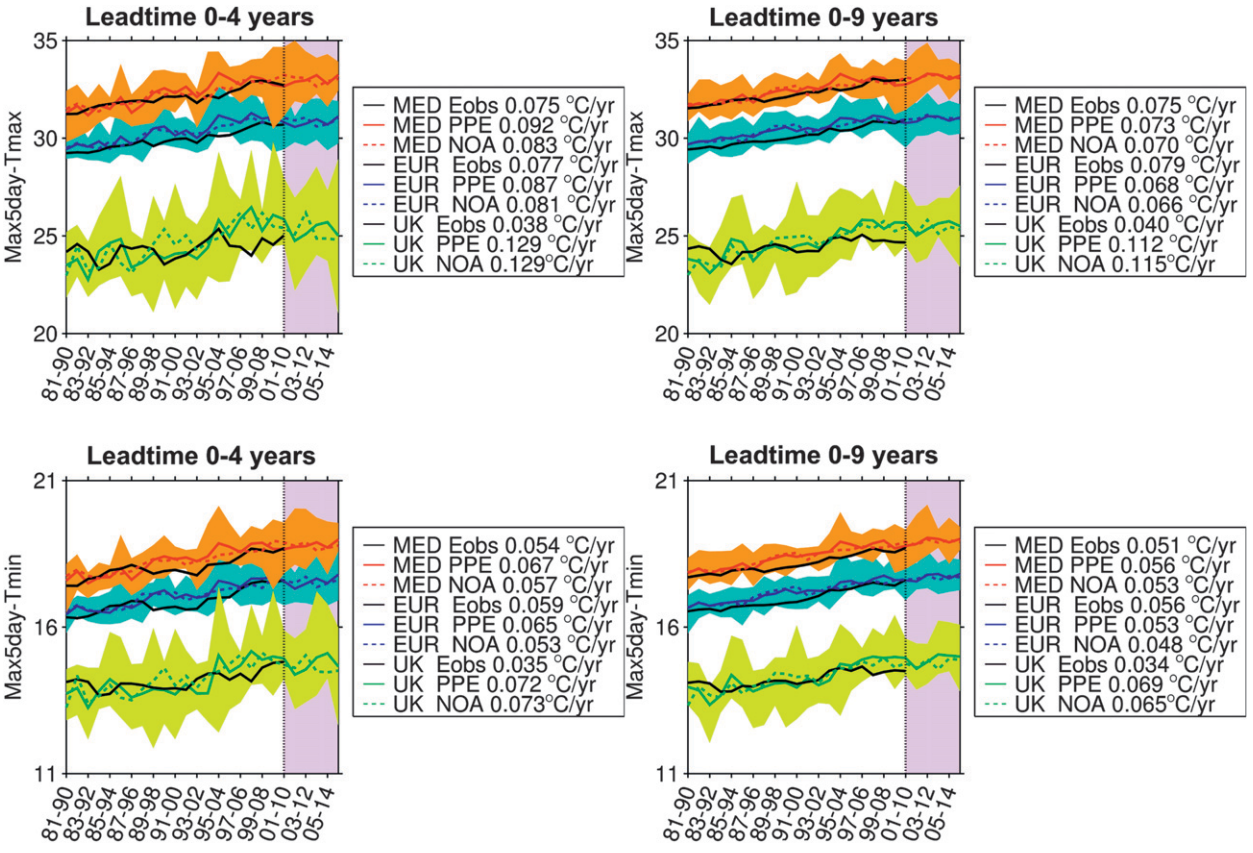

FIG. 6. As in Fig. 5, but for maximum 5-day average indices Max5day-Tmax and Max5day-Tmin.

to observed climatology and significantly beyond random noise, although small overlaps may still indicate significant results given the variances are approximately additive.

First, the MSSSs are computed for STmax and STmin at individual lead times from the PPE ensemble compared to the 30-yr prior Eobs climatology. Forecasts of STmax and STmin do show some skill beyond the 30-yr prior climatology, especially for the Europe and Mediterranean regions. However, this is not significantly greater from the skill obtainable from a random forecast (gray points) for any regions, at any lead times (Fig. 7). This is because the model cannot adequately recreate the year-to-year variability of these indices. That is not to say the model predictions are worthless, though. It has already been shown that the 40-yr trend in these indices is well correlated with observations so the model predictions are providing some useful information on longer time scales. However, an average over lead times may be more skillful while still providing useful information on how indices could change in the next few decades.

Another feature seen in Fig. 7 is the MSSS increasing with lead time. This feature arises because the length of the period analyzed for each lead time stays the same, so later lead times include analysis over later years, a few of which were much warmer than average. The impact of this is an increased trend at later times and hence the signal-to-noise ratio also increases, providing slightly more skill over these later periods. It also highlights again the lack of skill coming from initialization as there is no reduction in skill with lead time but rather an increase.

To determine if longer-term averages are more greatly influenced by larger-scale processes (including ocean dynamics) and forced trends than the annual indices, we look at skill in the 5- and 10-yr averages of these indices as predicted from each model run. The MSSS, as applied in the top left panel of Fig. 8, assesses the amount of skill a model has in predicting the 5-yr and decadal averages of STmax/STmin beyond the observed 30-yr prior climatology. There is significant skill in 5-yr and 10-yr average STmax and STmin for the Europe, Mediterranean, and UK regions (Fig. 8, top-left panel) but not for the UK 1–5-yr average STmax, which does not show significant skill beyond the climatological average.

When repeating the analysis applied to the seasonal mean to the maximum 5-day average indices we see significant skill for the 5-yr and decadal average Max5day-Tmax

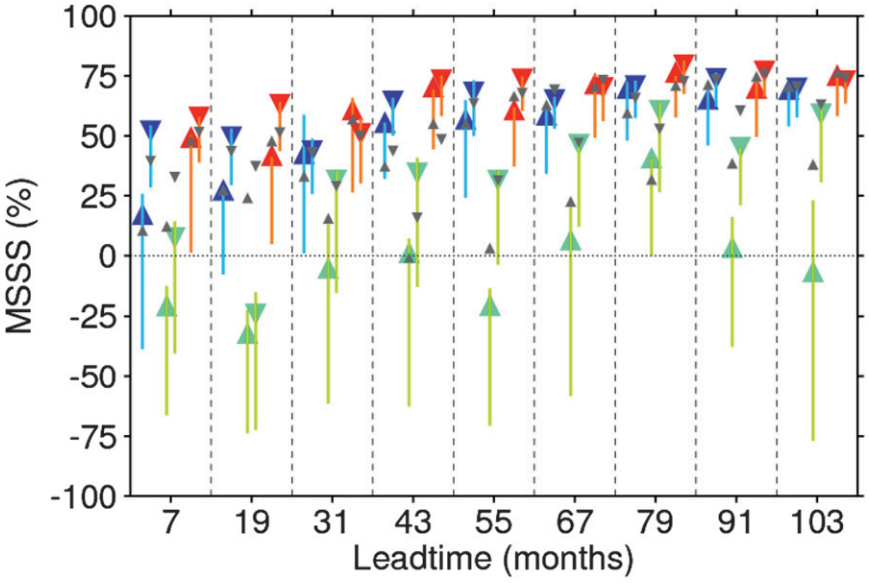

FIG. 7. Mean square skill score of summer average indices STmax (up arrow) and STmin (down arrow) for individual lead times from DePreSys PPE ensemble averaged over several regions including, Europe (blue), the UK (green), and the Mediterranean (red). This was calculated with indices that had the bias of the Eobs climatology of that index removed before calculation. Error bars are calculated using bootstrapping with replacement and the gray point represents the 90th percentile of the range of MSSS obtainable with realizations of random noise.

and Max5day-Tmin (Fig. 8, top-right panel) in Europe and the Mediterranean but not consistently in the UK. This is promising as this index requires the model to capture changes in daily Tmax and Tmin extremes rather than just the seasonal mean warming.

Similarly for the winter indices, the winter average WTmax (Fig. 8, bottom-left panel) shows skill for all regions at all lead times, while winter average WTmin only shows skill in Europe and the UK, not the Mediterranean. Also, the 5-day minimum indices (Fig. 8, bottom-right panel) are not skillful for Europe and Mediterranean, with the predictions for the Mediterranean becoming particularly unskillful at later lead time averages. The only skill in this index is the UK decadal average and 6–10-yr average Min5day-Tmin and Min5day-Tmax.

b. Is there skill beyond persistence?

There are methods of prediction other than the modeled prediction or climatology that could be used as benchmarks, such as, predictions with other climate models, statistical models or by simply persisting the previous years' or decades' value. The alternative option for a baseline tested here is the observation from the previous 5-yr average or decade. This removes the need to recreate a trend in the data due to external forcings that are unlikely to affect the climate on such as short time scales. Although in this case, there is no measure of variability of the index from past experience. In this case, we may find this prediction is closer to the observation if the following year is similar to the previous, which could be purely due to chance or because the index follows a variability cycle or mode that is autocorrelated. Therefore, it is important to test different methods of prediction to ensure the best prediction method for the specific index in question is being employed, as it will not always be the same across indices.

Figure 9 shows the MSSS for the modeled results compared to the previous 5-yr/decadal average. This is done using the same method as employed when the PPE hindcasts are compared to NoAssim, as detailed by Eq. (6). However, here the reference forecast is the average of the 5 years (or 10 years for the decadal average indices) immediately prior to the start of the run. As such, this represents the known value of the index at the start of the simulation, which we persist to obtain the next values of this index, hereafter termed "persistence." Where the value of this MSSS is greater than zero along with its associated error, the model is considered to be a better prediction than persistence. The model predictions are more skillful predictions than persistence for the majority of Europe and Mediterranean

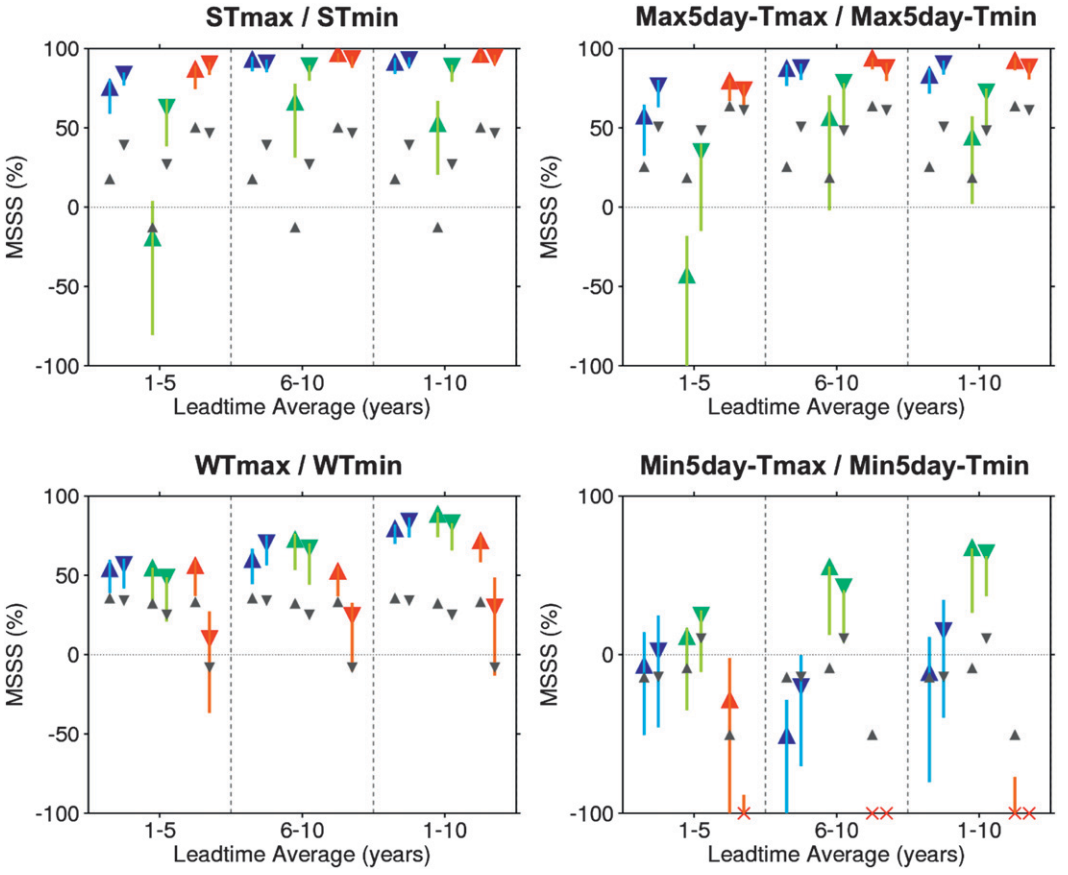

FIG. 8. MSSS of 5-yr–10-yr averaged indices from DePreSys PPE compared to Eobs 30-yr prior climatology, with lead time averaged over several regions in Europe, including Europe (blue), the UK (green), and the Mediterranean (red). The up arrows correspond to the indices computed from daily maximum temperatures (Tmax) and the down arrows are the indices computed with daily minimum temperatures (Tmin). The indices shown are: (top left) STmax and STmin and (top right) Max5day-Tmax and Max5day-Tmin; and (bottom left) WTmax and WTmin, and (bottom right) Min5day-Tmax and Min5day-Tmin. Error bars are calculated using bootstrapping with replacement and the gray point represents the 90th percentile of the range of MSSS obtainable with realizations of random noise. For further explanation of the uncertainty estimation, see section 3f. Where skill score is below −100, showing the forecast is particularly unskilled compared to climatology, a times sign is placed at the bottom of the plot.

indices. The exceptions are European STmax (1–5-yr average only; Fig. 9, top-left panel), European Max5day-Tmax (1–5-yr average only; Fig. 9, top-right panel), Mediterranean WTmin (Fig. 9, bottom-left panel), and Min5day-Tmax and Min5day-Tmin for European–Mediterranean at lead time averages of 6–9 and 0–9 years (Fig. 9, bottom-right panel). However of these, it is only the Mediterranean WTmin, Min5day-Tmax, and Min5day-Tmin indices for which observed persistence (Fig. 9, bottom-left panel) and climatology (Fig. 8, bottom-left panel) are both more skillful than the PPE predictions.

The UK region indices show improved skill in the modeled results compared to persistence for most indices except for the STmax and Max5day-Tmax and Max5day-Tmin (1–5-yr average only; Fig. 9, top panels). Interestingly, the modeled winter and minimum 5-day averages are more skillful than persistence for the UK (Fig. 9, bottom panels). However, this is not so meaningful for the 1–5-yr lead time average Max5day-Tmax, Min5day-Tmax and Min5day-Tmin as these indices are not skillful beyond climatology (Fig. 8, bottom-right panel).

c. *Source of predictive skill*

To assess where the skill in the model is originating, we ask whether it comes from initialization. This is done by comparing the skill obtained for the indices computed with the initialized (PPE) ensemble with the same indices computed with the uninitialized (NoAssim)

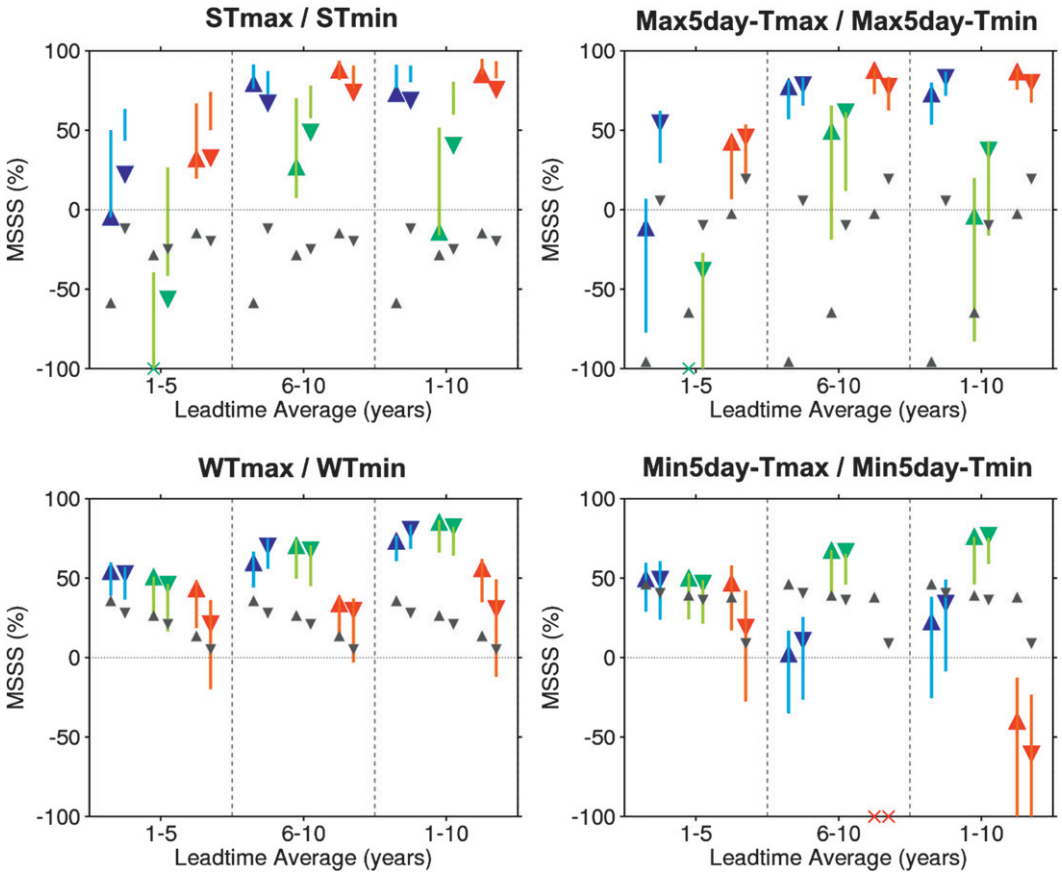

FIG. 9. As in Fig. 8, but with DePreSys PPE compared to persistence (the observed value of the previous 5-yr–decadal average of the index) rather than climatology.

model. Figure 10 shows there is no consistent and significant difference in skill between PPE and NoAssim across indices but there is the odd case. These results suggest there may be some benefit to the skill of the forecasts of the decadal average European STmax/STmin, decadal average Mediterranean STmin, and decadal average Mediterranean WTmin. However, for the maximum and minimum 5-day average indices (Fig. 10, right panels) there is no improvement of skill in forecasting these indices from the initialization. Where it has been determined that there is no skill coming from the initialization, the alternative source of skill is due to the model forcing, which is allowing the model to recreate the observed trend in temperatures over time irrespective of initialization.

5. Conclusions

This study outlines and applies a framework for use when evaluating the skill of predictions of seasonal mean and extreme temperature indices. It has the benefit of being applicable to any index as the computation is performed after the index is calculated. It has been important to bias correct the index before the skill is tested and this correction needs to be applied after the index is calculated, not before, as the bias in an index for extremes can vary substantially from that of the mean. Another benefit of this method is that it can be used to compare the skill in different models.

We compare an observationally initialized perturbed physics GCM to its uninitialized counterpart, observed climatology, and also to a statistical model in the form of persistence. We find significant and robust skill that exceeds persistence and climatology for many of the temperature extremes studied—in particular, the 5-yr/decadal average seasonal means and the 5-yr/decadal average maximum 5-day mean in Europe and Mediterranean.

It would seem that DePreSys performs better than alternative prediction methods for a number of the indices studied. However, as it is not consistently better across all indices and regions, when a model produces a skillful prediction for an index in a given region we cannot assume that it will also be as skillful for

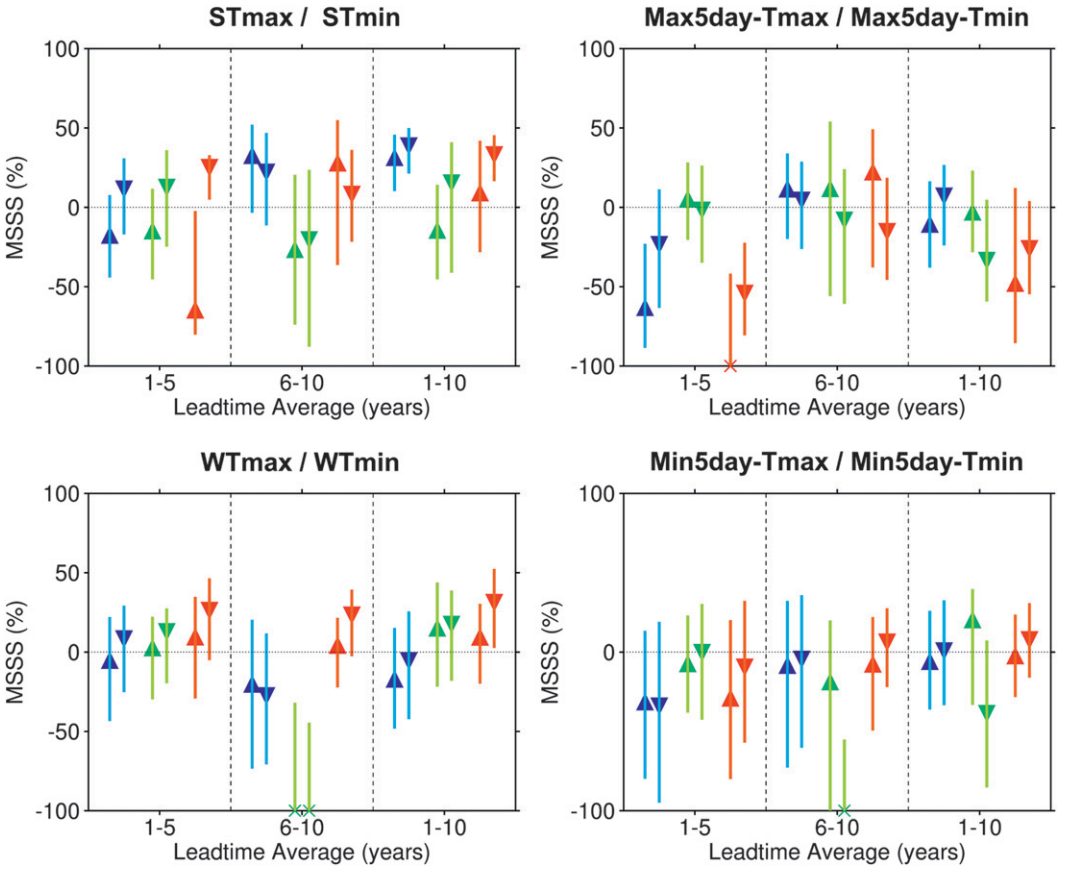

FIG. 10. As in Fig. 8, but with DePreSys PPE compared to NoAssim rather than climatology. Comparison to the skill generated by random noise is excluded from this plot as it is already shown by the gray shading in Fig. 8.

other regions or other indices. Hence it is useful to test the skill for each index and region individually in order to determine the best method of prediction for each case.

The predictions appear most skillful for the seasonal average indices (summer and winter) and the extreme summer indices (Max5day-Tmax and Max5day-Tmin) and are due to external forcing. The extreme winter temperatures (Min5day-Tmin and Min5day-Tmax) show the lowest predictive skill. This is the case for the European and Mediterranean regions, for which the signal-to-noise ratio of the trend compared to variability is stronger than that of the UK, which shows poorer skill (except for some indices based on Tmin). The reason for lower skill in the UK could be due to skill varying with location or a low signal-to-noise ratio as this region has a much smaller spatial area than the other regions. In addition to these regions the northern Europe region was also considered (50°–65°N, 10°E–40°W, which includes the UK). This region displays similar skill to that of the European region (see Figs. 3–5 in the supplement), and hence it is more likely that the lower skill in the UK is due to the variability being greater on this smaller spatial scale.

The skill in the summer average temperatures is due to the model forcing recreating the trend in seasonal averages due to external forcings well. There is poorer but still significant skill seen in the maximum 5-day average temperatures as the processes governing maximum temperatures are harder to model and tend to be nonlinear, so much so that the way extreme temperatures propagate can be heavily dependent on the land and atmospheric conditions present.

For the indices assessed here we find little impact on skill due to the initialization beyond the first year, consistent with a complimentary study by Eade et al. (2012). However, there is the odd case where there may be some skill in the decadal average prediction coming for the initialization, specifically the European summer average STmax–STmin, decadal average European WTmin, and decadal average summer and winter average Mediterranean STmin and WTmin. These hints of predictability are consistent with other recent studies (Matei et al. 2012; D. Matei 2012, personal communication).

Acknowledgments. HMH, GCH, and SFBT were supported by the UK Natural Environment Research Council through the EQUIP project (Grant NE/H003525/1). DMS was supported by the Joint UK DECC/Defra Met Office Hadley Centre Climate Programme (GA01101) and the EU FP7 COMBINE 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), along with the Met Office Hadley Centre for the DePreSys dataset and Edinburgh Compute and Data Facility (ECDF) for providing computer resources. In addition we would also like to thank fellow EQUIP members Chris Ferro, Tom Fricker, and Emma Suckling for providing useful advice.

APPENDIX

Skill Score Calculation

We rearrange Eqs. (2) and (3) by taking the forecast f = ppe, the value obtained from the ensemble mean of annual index from the perturbed physics ensemble, and the reference forecast $y = \mu_i$ is the 30-yr prior climatology for each run i. By using a dynamic 30-yr prior climatology [$y = \mu_i$, a multivalued external climatology, as described in Murphy (1988)], the rearrangement gives Eq. (A1), where r_{fx} is the correlation coefficient between the forecast (f) and the observations (x), s_f is the standard deviation of the forecast (f), s_x is the standard deviation of the observations (x), $\overline{f}$ is the mean of the forecast, and $\overline{x}$ is the mean of the observations:

$$\mathrm{MSSS}(\mathrm{ppe},\mu,x) = \frac{r^2_{\mathrm{ppe},x} - \left(r_{\mathrm{ppe},x} - \frac{s_{\mathrm{ppe}}}{s_x}\right)^2 - \left(\frac{\overline{\mathrm{ppe}} - \overline{x}}{s_x}\right)^2 - r^2_{\mu\mathrm{x}} + \left[r_{\mu\mathrm{x}} - \left(\frac{s_\mu}{s_x}\right)\right]^2 + \left[\frac{(\overline{\mu} - \overline{x})}{s_x}\right]^2}{1 - r^2_{\mu\mathrm{x}} + \left[r_{\mu\mathrm{x}} - \left(\frac{s_\mu}{s_x}\right)\right]^2 + \left[\frac{(\overline{\mu} - \overline{x})}{s_x}\right]^2}. \quad \text{(A1)}$$

Similarly the equation used to calculate MSSS of PPE compared to the NoAssim forecast, shown in Eq. (4), is rearranged to give Eq. (A2):

$$\mathrm{MSSS}(\mathrm{ppe},\mathrm{noa},x) = \frac{r^2_{\mathrm{ppe},x} - \left(r_{\mathrm{ppe},x} - \frac{s_{\mathrm{ppe}}}{s_x}\right)^2 - \left(\frac{\overline{\mathrm{ppe}} - \overline{x}}{s_x}\right)^2 - r^2_{\mathrm{noa},x} + \left[r_{\mathrm{noa},x} - \left(\frac{s_{\mathrm{noa}}}{s_x}\right)\right]^2 + \left[\frac{(\overline{\mathrm{noa}} - \overline{x})}{s_x}\right]^2}{1 - r^2_{\mathrm{noa},x} + \left[r_{\mathrm{noa},x} - \left(\frac{s_{\mathrm{noa}}}{s_x}\right)\right]^2 + \left[\frac{(\overline{\mathrm{noa}} - \overline{x})}{s_x}\right]^2}. \quad \text{(A2)}$$

REFERENCES

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

Branstator, G., and H. Teng, 2012: Potential impact of initialization on decadal predictions as assessed for CMIP5. *Geophys. Res. Lett.,* **39,** L12703, doi:10.1029/2012GL051974.

——, ——, G. Meehl, M. Kimoto, J. Knight, M. Latif, and A. Rosati, 2012: Systematic estimates of initial-value decadal predictability for six AOGCMs. *J. Climate,* **25,** 1827–1846.

Chikamoto, Y., and Coauthors, 2013: An overview of decadal climate predictability in a multi-model ensemble by climate model MIROC. *Climate Dyn.,* **40,** 1201–1222.

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

Conil, S., H. Douville, and S. Tyteca, 2009: Contribution of realistic soil moisture initial conditions to boreal summer climate predictability. *Climate Dyn.,* **32,** 75–93.

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. Biometeor.,* **50,** 342–348.

D'Ippoliti, D., and Coauthors, 2010: The impact of heat waves on mortality in 9 European cities: Results from the EuroHEAT project. *Environ. Health,* **9** (37), doi:10.1186/1476-069X-9-37.

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, 436 pp.

Fink, A. H., T. Brucher, A. Kruger, G. C. Leckebusch, J. G. Pinto, and U. Ulbrich, 2004: The 2003 European summer heatwaves and drought—Synoptic diagnosis and impacts. *Weather,* **59,** 209–216.

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

Fouillet, A., and Coauthors, 2006: Excess mortality related to the August 2003 heat wave in France. *Int. Arch. Occup. Environ. Health,* **80,** 16–24.

Goddard, L., and Coauthors, 2013: A verification framework for interannual-to-decadal predictions experiments. *Climate Dyn.,* **40,** 245–272.

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

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:10.1029/2011JD016541.

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

Haylock, M. R., N. Hofstra, A. M. G. K. Tank, E. J. Klok, P. D. Jones, and M. New, 2008: A European daily high-resolution gridded data set of surface temperature and precipitation. *J. Geophys. Res.,* **113,** D20119, doi:10.1029/2008JD010201.

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

Keenlyside, N., M. Latif, J. Jungclaus, L. Kornblueh, and E. Roeckner, 2008: Advancing decadal-scale climate prediction in the North Atlantic sector. *Nature,* **453,** 84–88.

Lee, T., F. Zwiers, X. Zhang, and M. Tsao, 2006: Evidence of decadal climate prediction skill resulting from changes in anthropogenic forcing. *J. Climate,* **19,** 5305–5318.

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

Meehl, G., and Coauthors, 2009: Decadal prediction. *Bull. Amer. Meteor. Soc.,* **90,** 1467–1485.

Murphy, A. H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.,* **116,** 2417–2424.

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

Palmer, T. N., F. J. Doblas-Reyes, A. Weisheimer, and M. J. Rodwell, 2008: Toward seamless prediction: Calibration of climate change projections using seasonal forecasts. *Bull. Amer. Meteor. Soc.,* **89,** 459–470.

Pascal, M., and Coauthors, 2006: France's heat health watch warning system. *Int. J. Biometeor.,* **50,** 144–153.

Pohlmann, H., J. Jungclaus, A. Köhl, D. Stammer, and J. Marotzke, 2009: Initializing decadal climate predictions with the GECCO oceanic synthesis: Effects on the North Atlantic. *J. Climate,* **22,** 3926–3938.

Pope, V. D., M. L. Gallani, P. R. Rowntree, and R. A. Stratton, 2000: The impact of new physical parametrizations in the Hadley Centre climate model: HadAM3. *Climate Dyn.,* **16,** 123–146.

Quesada, B., R. Vautard, P. Yiou, M. Hirschi, and S. I. Seneviratne, 2012: Asymmetric European summer heat predictability from wet and dry southern winters and springs. *Nat. Climate Change,* **2,** 736–741.

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

——, 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.

Smith, D., and J. Murphy, 2007: An objective ocean temperature and salinity analysis using covariances from a global climate model. *J. Geophys. Res.,* **112,** C02022, doi:10.1029/2005JC003172.

——, S. Cusack, A. Colman, C. Folland, G. Harris, and J. Murphy, 2007: Improved surface temperature prediction for the coming decade from a global climate model. *Science,* **317,** 796–799.

——, R. Eade, N. J. Dunstone, D. Fereday, J. M. Murphy, H. Pohlmann, and A. Scaife, 2010: Skilful multi-year predictions of Atlantic hurricane frequency. *Nat. Geosci.,* **3,** 846–849.

Solomon, S., and Coauthors, 2007: Technical summary. *Climate Change 2007: The Physical Science Basis,* S. Solomon et al., Eds., Cambridge University Press, 19–91.

Teuling, A. J., and Coauthors, 2009: A regional perspective on trends in continental evaporation. *Geophys. Res. Lett.,* **36,** L02404, doi:10.1029/2008GL036584.

Uppala, S. M., and Coauthors, 2005: The ERA-40 Re-Analysis. *Quart. J. Roy. Meteor. Soc.,* **131,** 2961–3012, doi:10.1256/qj.04.176.

van der Linden, P., and J. F. B. Mitchell, Eds., 2009: ENSEMBLES: Climate change and its impacts: Summary of research and results from the ENSEMBLES project. Met Office Hadley Centre, 160 pp. [Available online at http://ensembles-eu.metoffice.com/docs/Ensembles_final_report_Nov09.pdf.]

van Oldenborgh, G., F. Doblas-Reyes, B. Wouters, and W. Hazeleger, 2012: Decadal prediction skill in a multi-model ensemble. *Climate Dyn.,* **38,** 1263–1280.

Vautard, R., and Coauthors, 2007: Summertime European heat and drought waves induced by wintertime Mediterranean rainfall deficit. *Geophys. Res. Lett.,* **34,** L07711, doi:10.1029/2006GL028001.

WCRP, 2011: Data and bias correction for decadal climate predictions. International CLIVAR Project Office Publication Series No. 150, 3 pp. [Available online at http://www.wcrp-climate.org/decadal/references/DCPP_Bias_Correction.pdf.]

Weisheimer, A., F. Doblas-Reyes, T. Jung, and T. Palmer, 2011: On the predictability of the extreme summer 2003 over Europe. *Geophys. Res. Lett.,* **38,** L05704, doi:10.1029/2010GL046455.