

Near-term prediction of impact-relevant extreme temperature indices

H. M. Hanlon · G. C. Hegerl · S. F. B. Tett · D. M. Smith

Received: 16 January 2013 / Accepted: 16 June 2014 / Published online: 12 August 2014
© Springer Science+Business Media Dordrecht 2014

Abstract A previous study of predictability of European temperature indices revealed significant skill in predictions of 5/10-year average indices of 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. Here, this work is extended to study indices relevant to high heat-related impacts on energy use, human health and maize yields in Europe. The skill of predictions of these indices is assessed using decadal predictions of the number of days above critical thresholds of daily maximum, mean and minimum Summer temperatures. Following comparison of these predictions with observed conditions, there is skill found in parts of Europe where the decadal predictions exceed that of using observed climatology and persisting present conditions. Areas in the Mediterranean show the most skill in near-term predictions, while skill is small in Northern/Central Europe. There is even some evidence of skill on small scales. This system is determined to be not appropriate for predicting indices in the UK as the model significantly overestimates the trend in these indices. A further test studies the effect of initialising the decadal forecasts with observations. Simulations that include external forcing, such as greenhouse gas increases, show better skill in predicting changes in the frequency of hot events than those that do not, and the initialisation of forecasts with the model used here does not improve this skill.

This article is part of a Special Issue on “Managing Uncertainty in Predictions of Climate and Its Impacts” edited by Andrew Challinor and Chris Ferro.

Electronic supplementary material The online version of this article (doi:10.1007/s10584-014-1191-3) contains supplementary material, which is available to authorized users.

H. M. Hanlon (✉) · D.M. Smith
Met Office Hadley Centre, Fitzroy Road, Exeter, Devon EX1 3PB, UK
e-mail: helen.hanlon@metoffice.gov.uk

G. C. Hegerl · S.F.B. Tett
School of Geosciences, University of Edinburgh, West Mains Road, Edinburgh, EH9 3JW, UK

Keywords Extreme temperatures · Adaptation · Decadal prediction · Extreme events · Climate · Impacts

1 Introduction

Extreme temperature events, such as the 2003 European Heatwave (Schär et al. 2004; Fink et al. 2004) and 2010 Russian heatwave (Barriopedro et al. 2011), have had a severe impact on society and nature, in particular the impact on human health was profound. In terms of impacts, it is not extreme seasonal temperatures, but increased daily extreme temperatures which are most damaging (Díaz et al. 2006; Fouillet et al. 2006; Grize et al. 2005; Pascal et al. 2006). Impacts of daily extreme temperatures are now even more important to study. It has been shown that daily extreme temperatures show an upward trend in mean summer daily maximum (Tmax) and daily minimum temperatures (Tmin) in Europe over the past few decades, that has been attributed in part to human influences (Christidis et al. 2012); and similarly, that the frequency (Morak et al. 2013) and intensity (Zwiers et al. 2011) of extreme temperatures has increased.

In light of how severe the impact of high temperatures can be, and because these events may become more frequent in the future, it has become even more important to determine how well we can predict the changing likelihood of such events to enable improved capability for adaptation and planning for the future. Hamilton et al. (2012) found seasonal forecasts of the number of daily extreme temperatures (outside the 10–90 % range) had significantly better skill than persistence, though, lower than the skill in predicting the seasonal mean especially in the extratropics. The summer season was the most skillful in the northern hemisphere. A recent study by Eade et al. (2012) demonstrated significantly skillful predictions of moderate (1 in 10) temperature extremes on decadal timescales, especially for multi-year periods. These assessments of skill were performed using the Spearman rank correlation coefficient and standardised root mean square error.

In this study we build upon the work of Hanlon et al. (2013a) who found significant skill, beyond observed climatology using the mean square skill score (Murphy 1988), in predicting the summer average and hottest 5-day average daily maximum (Tmax) and daily minimum (Tmin) temperatures in Europe with the Met Office Hadley Centre decadal prediction system (DePreSys). Hanlon et al. (2013a) determined that this skill is due almost entirely to the forecast recreating the climate change signal rather than from its initialisation. Subsequently, work shown in Hanlon et al. (2013b) found similar results using four CMIP5 models (CanCM4, HadCM3, MIROC5 and MPI-ESM-LR). However, for some models there was evidence of improved skill when initialising with observations. In particular, the Max Planck Institute Earth System Model (MPI-ESM-LR, Raddatz et al. (2007), Marsland et al. (2003)), shows skill originating both from the external forcing due to climate change and from initialisation with observations.

Following these previous studies, we apply the same methodology as Hanlon et al. (2013a) comparing decadal hindcasts to observations and assessing the accuracy of them compared to observed climatology and persistence. However, here we use indices which are based on exceedance of temperature thresholds which are more directly relevant to impacts and this paper investigates the usefulness of such decadal predictions for informing adaptation decisions.

2 Methodology for predicting extreme temperature indices

The procedure applied follows the methodology described in Hanlon et al. (2013a); now termed H13. That paper presented a method to calculate decadal predictions for climate indices from climate model simulations, and evaluates them based on a comparison to alternative prediction methods. Here, this methodology is adapted to assess the predictability of more impact-relevant indices which count the number of days above a critical threshold.

2.1 Data

Observations The observed data originate from the Ensembles project observational database (Eobs) (see Haylock et al. (2008) for more details). The data set is based on individual station data which have been interpolated to a high resolution (0.5° latitude by 0.5° longitude grid) regular grid. As a consequence, there will still be some uncertainty in these observations due to variations in the density of stations and interpolation methods, along with any measurement errors. We have regridded these observations using area-averaging, to the horizontal resolution of the model (3.75° longitude by 2.5° latitude) prior to computing the indices, to allow a more direct comparison.

Models As in H13, we use the UK Met Office decadal prediction system (DePreSys) (Smith et al. 2007, 2010) which employs the Hadley Centre coupled global climate model version 3 (HadCM3) (Pope et al. 2000; Gordon et al. 2000) at a horizontal resolution of the atmospheric component of 3.75° longitude by 2.5° latitude. Further details of the model are given in H13. We use a set of retrospective forecast experiments (hindcasts) comprising a 9 member perturbed physics ensemble (hereafter referred to as PPE or the initialized ensemble). For each ensemble member a different variant of HadCM3 is used which span a range of different climate sensitivities from 2.6 – 7.1°C and El Niño variabilities, in order to sample model parameter uncertainty. This is achieved by applying different combinations of perturbations to 29 parameters that control sub-grid scale atmospheric and surface processes (Murphy et al. 2004).

Both the atmosphere and ocean components of the system were initialised every November from 1960 to 2005 with anomalies from the observed climatology added onto the model climatology, and then run for 10 years (see Smith et al. (2007, 2010); also H13). The model system has also been run without initialising with observed conditions (non-initialised ensemble). Each individual member of the non-initialised ensemble is performed with the same model variants as the corresponding member of the initialised ensemble, but without assimilation of the observed state of the atmosphere or ocean. This ensemble is used to diagnose whether the initialisation has improved the forecasts, or whether any forecast skill we might diagnose originates from inclusion of external influences on climate, such as increasing greenhouse gases.

2.2 Indices

We quantify the skill in the following impact-relevant indices:

- TX29: Number of days where T_{\max} exceeds 29°C between 1st of April and 30th of September.
- CDD: Cooling Degree Days, the cumulative sum of the number of degrees over which the daily mean temperature exceeds 18°C each day between 1st of April and 30th of September.

- TX25: Number of days where T_{\max} exceeds $25\text{ }^{\circ}\text{C}$ between 1st of April and 30th of September.
- TN18: Number of days where T_{\min} exceeds $18\text{ }^{\circ}\text{C}$ between 1st of April and 30th of September.
- CHDWN: Combined hot days and warm nights, the number of days where both T_{\max} and T_{\min} exceed a set of thresholds. For this index two sets of thresholds were considered, $T_{\max} > 30\text{ }^{\circ}\text{C}$ with $T_{\min} > 15\text{ }^{\circ}\text{C}$ (low threshold) and $T_{\max} > 35\text{ }^{\circ}\text{C}$ with $T_{\min} > 20\text{ }^{\circ}\text{C}$ (high threshold). Please note, a day is only counted in this index if both the T_{\max} and T_{\min} values for that day exceed the corresponding thresholds.

T_{\max} above $29\text{ }^{\circ}\text{C}$ has a detrimental impact on crop yields such as corn (Schlenker and Roberts 2009) as the plants start to die as temperatures exceed this threshold.

The cooling degree day index (CDD) is a measure of power consumption due to use of cooling systems to regulate building temperature. It is an index that is averaged over a large area to give an estimate of energy demand over that area during periods of high temperatures. It does not attempt to predict the power required for small areas or individual buildings in which thermostats may be set to a wide variety of temperatures for operational, comfort or cost-saving reasons. There are less guidelines published for CDD in Europe than for winter heating degree days (HDD) but as summer temperatures display a strong upward trend this index may become more important in the future. Failure to meet increasing demand of power for cooling systems during high-heat events could exacerbate the impacts of those events.

CDD is calculated by taking difference between the threshold value and daily average temperature for all days in April to October and finding the cumulative sum, giving the total number of degrees above the baseline threshold for that season. Here we have used $18\text{ }^{\circ}\text{C}$ ($65\text{ }^{\circ}\text{F}$) as the threshold, which is the baseline regularly used across the US (National Climatic Data Center 2012) based on guidelines for building design to ensure human comfort. It is a rough measure of how much power will be required to cool buildings across large areas (e.g. continents). Essentially, the threshold should be the temperature at which the majority of thermostats are set to turn air conditioning on. This threshold maybe a little low for some parts of Europe, for example, UKCP09 suggest $22\text{ }^{\circ}\text{C}$ as a threshold for the UK (Jenkins et al. 2008). Jenkins et al. (2008) found an increase in average number of CDD defined in all regions of the UK from 1961 to 2006.

TX25 and TN18 are examples of more moderate thresholds for extremes. Despite not being directly linked to dramatic impacts, they may still be useful. $25\text{ }^{\circ}\text{C}$ is the temperature threshold above which increasing cases of health impacts occur in the UK, concurrent with rising hospital admissions UK Department of Health/ NHS (2012). Results for these additional indices are shown in the supplement.

We also looked at combined extremes of T_{\max} and T_{\min} with the CHDWN index, to account for health effects that occur with combined high daytime and night time temperatures. Two sets of thresholds were considered as the low threshold that would be appropriate for the UK region (UK Department of Health/ NHS 2012) is not necessarily appropriate for more southerly regions of Europe that are already adapted to high temperatures. For this second case higher thresholds were chosen following recommendation in Fischer and Schär (2010).

All indices described above are computed for all ten years of the decadal hindcasts with start dates between 1979 and 2005 (inclusive). This time period is chosen to allow us to correct the forecast with out-of-sample data for 30 years previous, and is constrained by

reliable observations which are available from 1950. The skill is assessed based only on the indices computed for runs with start dates between 1979 and 2000, as we require observations for ten years following the start date. We also calculate skill averaged across leadtimes in order to assess if the forecasting system can skilfully predict changes over a longer time: we calculate a pentadal average over the first five years of the forecast; and a decadal average over all ten years of the forecast. This average over leadtime is performed after the index has been computed and bias corrected for each individual leadtime. Furthermore, we assess the prediction of regional averages over land points only. The regions used in this study are: Europe (35–65 °N latitude, 10 °W–40 °E longitude), the British Isles (50–60 °N latitude, 10 °W–2 °E longitude, hereafter termed ‘UK’ but this does still include Ireland), the Mediterranean (35–50 °N latitude, 10 °W–40 °E longitude) and Central Europe (42–55 °N latitude, 2 °W–20 °E longitude).

3 Processing and evaluating predictions

3.1 Bias correction

A bias exists between the modelled and observed extremes which is influenced by small scale parametrised processes and local feedbacks, which are not always well captured by the model and are different for extreme temperatures than for mean temperatures (Hanlon et al. 2013a).

Following the index calculation, this mean bias of the observed index is removed, avoiding the mean difference between the observed and modelled index, leading to much more useful predictions (e.g. Hawkins et al. (2013)). For the regionally averaged indices the bias correction is performed after calculating the regional average. Our bias correction method follows the guidelines set by the World Climate Research Program (WCRP 2011): the bias correction to an index is made by removing the mean index calculated over the 30 years prior to the initialisation from a transient run with the same model variant and adding on the mean index over the same time period calculated with observations. As shown by Eq. 1:

$$\text{model_bc}_{y,l,m} = \text{model_bc}_{y,l,m} - \text{mean}(\text{transient}_{y-30:y-1,0,m}) + \text{mean}(\text{eobs}_{y-30:y-1}) \quad (1)$$

where, y is the year, l is the leadtime from the start of the run and m refers to each ensemble member. For example, the index calculated for the run starting in November 1979 is corrected with the mean over 1950–1979. The reason for correcting with prior data is to allow the same method of bias correction to be used to correct a future forecast. Preventing our skill assessment from being preconditioned with observations that occurred during the in-sample time period, which would not have been available at the time.

We do not apply a time-dependent bias correction both in order to correct as little as possible, and because the model we use is anomaly initialised, leading to forecasts that are less likely to drift than those with full field initialisation (Smith et al. 2013).

We correct the index rather than the daily data, as extremes show a different bias from average days (for more discussion, see H13). For clarity, where the index is calculated for each grid point individually the bias correction is also performed at each grid point individually and where the index is a regional average the bias correction is done for the regional average. Also, as each member of the perturbed physics forecasts are created using a different variant of the model, we consider each member as a separate model and

therefore bias correct each individually. The correction applied remains constant across different leadtimes.

4 Evaluating forecast skill

As in H13, the average skill of the initialised decadal forecasts, compared to forecasts made using an alternative method of prediction, the ‘reference forecast’. This is done by computing the Mean Square Skill Score (MSSS) (Murphy 1988), which compares the mean square errors between each forecast and the observations by:

$$MSSS(f, y, x) = \left(1 - \frac{MSE(f, x)}{MSE(y, x)} \right) \times 100 \quad (2)$$

Where MSE, the mean square error is calculated as:

$$MSE(f, x) = \frac{1}{n} \sum (f_i - x_i)^2, \quad (3)$$

f_i is the i th value from forecast f , x_i is the i th observed value, y is the reference forecast and n is the number of forecasts (here the forecasts are the 10-year runs started every year).

In this study we use MSSS to estimate how accurately the ensemble mean of the initialised perturbed physics ensemble (PPE) DePreSys hindcasts recreate the corresponding observed values, compared to alternative forecasts based on observational climatology, the persistence of the previous year/decade value of the index from observations, and equivalent climate model based hindcasts performed with no initialisation of observed values. This follows the best practice guidance for use of the mean square skill score to assess decadal predictions proposed by Goddard et al. (2013).

A forecast is deemed to be skillful if it is closer to the observed value than a reference forecast. The first reference forecast we use is the average of the previous 30-years before the start of the run. Prior climatology is a reasonable benchmark in this case as extreme temperature indices fluctuate year to year as they are affected by weather variability. Also, this method has the benefit of relying only on data available at the start of the run, allowing a fair test of the system forecasting future years, for which no observations are available. The calculations are also repeated, persisting observed conditions as the reference forecast. Persistence refers to the value of the index from the year (or average of years) immediately prior to the start of the run. In both cases the skill score assesses if the model-based forecasting system provides a better forecast than a forecast based on long-term average or preceding observations only.

To determine whether initialising the forecast with observed conditions improves its skill, the skill analysis is also performed using the non-initialised ensemble forecast as the reference forecast. This shows whether the ensemble which assimilates observations is more skillful than the runs which did not use initial conditions based on the observed state of the climate.

It should be noted that where the number of days observed and forecast by the benchmark is equal to zero this score results in zero error for the reference forecast leading to degenerate skill scores (Eq. 2). Thus the MSSS is not useful for assessing very extreme cases of indices based on count data, with no occurrences during the time period yielding the skill score misleading. This is not surprising, as assessing very rare events is difficult with most methods.

The MSSS is calculated for the indices described in Section 2.2 for each leadtime, where the leadtime refers to the time since the initialisation of the prediction. This MSSS is also found for an average over leadtimes 0–4 and 0–9 years, in order to determine if a decadal or semi-decadal average has more skill than that of a prediction for an individual year.

4.1 Uncertainty assessment

Uncertainties arise from the limited ensemble size of nine members and limited number of start dates, which we address in the same way as in H13, estimating the sampling uncertainty by bootstrapping-with-replacement (Efron and Tibshirani 1993, Chapter 6) across these nine members. The 10–90 % range from a thousand samples provides the error on MSSS. If the score and its error are above zero then the forecast has more skill in predicting the index than the reference forecast.

An additional method of estimating uncertainty on the MSSS is also performed, which compares the MSSS computed for the forecast to the MSSS computed with a random forecast, which should have no significant skill apart from coincidence (for detail, see H13). The 90th percentile of a distribution of the random forecasts (normalized to have the same mean and standard deviation as the real forecasting system) is taken as a cut off point. Below which the MSSS is considered to not be significantly better than 90 % of the random forecasts and hence not appropriate for use as prediction.

The forecast skill is limited by inherent predictability, which varies with the index and region studied. Predictability is further decreased by model uncertainty arising from errors in the model inputs such as the initial conditions, boundary conditions, physical constraints and the driving variables, the model structure, stochastic variability and underestimation of the observed variability due to small sample size. Some of this is removed by bias correction, although the forecast will still be affected by errors in the climate model.

Uncertainty in the values of climate model parameters is accounted for in our approach by applying different perturbations to physical parameters in the initialised ensemble. Although the small number of perturbed simulations (nine) may not capture the full uncertainty. Structural uncertainty, arising from inaccuracies in the numerical solution of the equations controlling the model, is not accounted for in this study as there is only one model used. Previous papers studying skill of extreme summer temperatures in Europe have shown that the findings are broadly similar across models, with some exceptions (Hanlon et al. 2013b). In summary, our study accounts for important aspects of model uncertainty, but cannot span the full possible range of climate model forecasting system behaviours.

As the forecast skill is based on a comparison to observations there is also the possibility of added uncertainty due to error in the observations. As discussed above, this may have arisen by measurement error or during the regridding and area averaging process. However, as the station density in Europe is high, we believe observation errors to be small so have neglected it.

This study does not rely on expert judgement but verifies the accuracy and validation of predictions using quantitative methods based on mean square errors between models and observations. Any judgments made have related to the suitability of the methods for this purpose. Where there has been more than one choice of method we have chosen the most conservative, for example the bootstrapping procedure for sampling the model spread.

5 Results

5.1 $T_{\max} > 29^{\circ}\text{C}$

The regional decadal average forecasts show a steady increase in the number of days exceeding the 29°C threshold for the Europe and Mediterranean regions, consistent with observations (Fig. 1, lower). The difference between the observed timeseries and the model ensemble mean shows that the TX29 index is biased slightly low, even after bias correction. Despite this small bias the model captures most of the observed variability, with the exception of a late 1990s decadal averages, increasing confidence in the prediction.

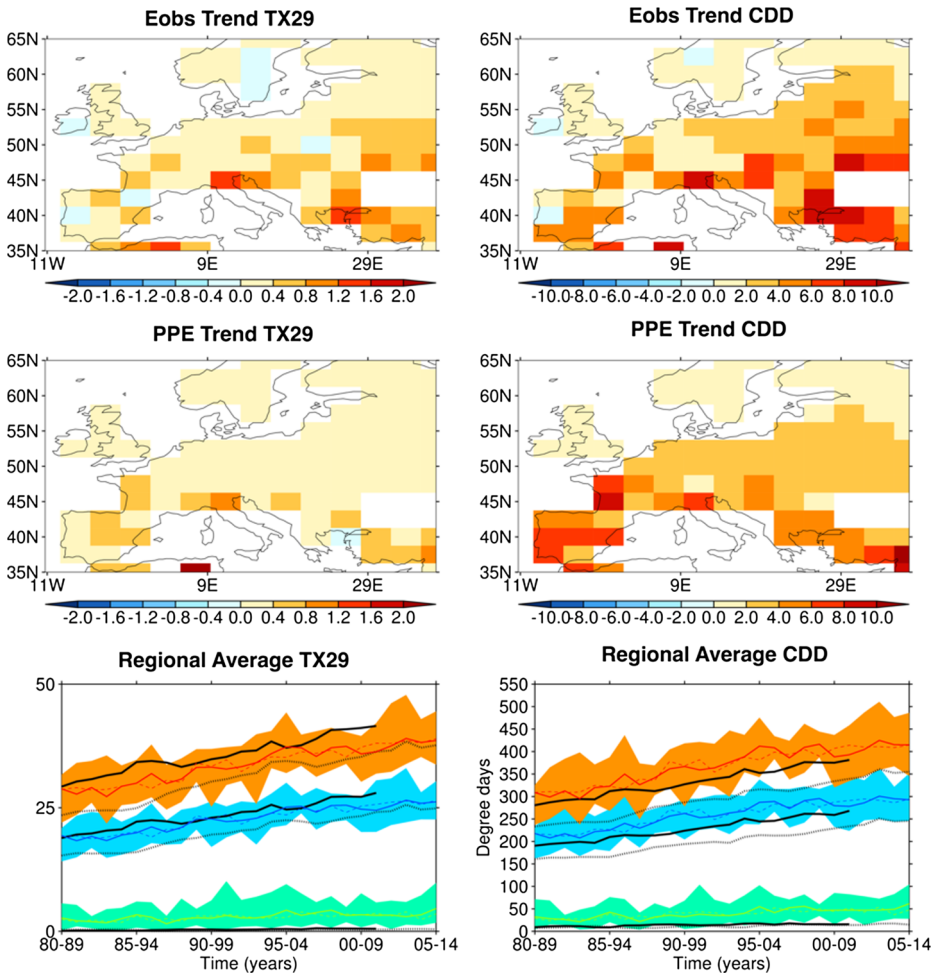


Fig. 1 Trend in decadal averaged TX29 ($^{\circ}\text{C}/\text{year}$, left) and CDD (days/year, right) of 1981–2010 computed with observations (*upper*), initialised ensemble (PPE) mean leadtime 0–9 average (*middle*). Timeseries of each index over 1981–2010 (lower) computed with observations (*black lines*), persistence (*dashed black lines*) and initialised predictions (averaged over lead-times 0–9 years) are shown by the coloured lines and non-initialised predictions by the dashed coloured lines (*red* is Mediterranean, *Blue* is Europe and *Green* is UK). The shading represents the uncertainty from bootstrapped estimates of initialised ensemble spread

An exception is the regional decadal average timeseries for the UK, which overestimates the number of exceedances for these thresholds. The observations show very few occurrences where this threshold is exceeded and zero trend in this index (Fig. 1, lower left). Hence this system is not reliably predicting changes in these indices over the UK.

Throughout most of Europe there is an increasing trend in the observed decadal average of TX29 (Fig. 1, upper left). For grid points that show a decrease in the number of days exceeding the 29 °C threshold, the trend is small compared to the increases seen elsewhere. The spatial pattern of these changes is not recreated perfectly by the model, although it does have similar large-scale features, for example, increasing trend in TX29 throughout Europe with larger changes in Southern Europe than further North (Fig. 1, middle left).

It is unlikely that a global model would be skillful at a spatial scale equivalent to an individual grid point due to the low signal to noise ratio of these indices at small spatial scales. However, there are some grid points which do show the decadal average of the model is more skillful than observed climatology (Fig. 2, upper right) and observed persistence (Fig. 2, lower right), especially in Southern Europe. Then the skill scores obtained with runs initialised with and without observations were compared, to determine if there is a benefit to initialisation. For TX29 there is no evidence to suggest the skill here is due to initialisation of the model (Fig. 2, middle right).

As the skill scores are not statistically significant across all grid points it is likely we are asking too much of our Global model in this case. Hence, this analysis is likely to be more robust when performed at a lower spatial resolution. So we take the regional average of the index over 4 spatial domains of varying size, Europe, Central Europe, Mediterranean and UK and repeat the skill analysis with these new averages (note that skill scores displayed in the left-hand panels of Fig. 2 are calculated for the regional average of the index rather than taking a regional average of the skill scores, as shown in the right-hand panels). There should be more skill for the large area average index than that calculated for each grid point, due to chaotic behaviour of the system and the model's limited ability to resolve and recreate small scale features.

Initialised predictions of TX29 for the Europe and Mediterranean regional averages are significantly more skillful than observed climatology and random noise at all leadtimes and for the 5-year and 10-year averages (Fig. 2, upper left). The 5 and 10-year averages of TX29 for the Central Europe region are more skilful than observed climatology but this is not consistently true for all the individual years (Fig. 2, upper left). The UK regional average does not show any skill for TX29, as the model overestimates the trend in these indices when averaged over the UK. This region is less prone to exceeding the 29 °C threshold than more southerly regions, so predicting these very rare events is a challenge.

Forecasts based on initialised predictions of TX29 for Europe and Mediterranean are significantly more skilfull than those based on persisting preceding 5-year and 10-year averages (just 10-year average for Central Europe region) but this is not consistent across the individual leadtimes (Fig. 2, lower left). As there is no significant improvement in the skill of these initialised predictions compared to predictions that are not initialised, the skill is due to the forcing of in the model, not the initialisation of the model with observations (Fig. 2, middle left).

5.2 TX25, TN18 and CHDWN

We also studied other daily temperature based indices, including TX25 where the daily maximum temperature exceeds a more moderate threshold of 25 °C and TN18 where the daily minimum temperature exceeds 18 °C. Similar figures, as displayed above for TX29,

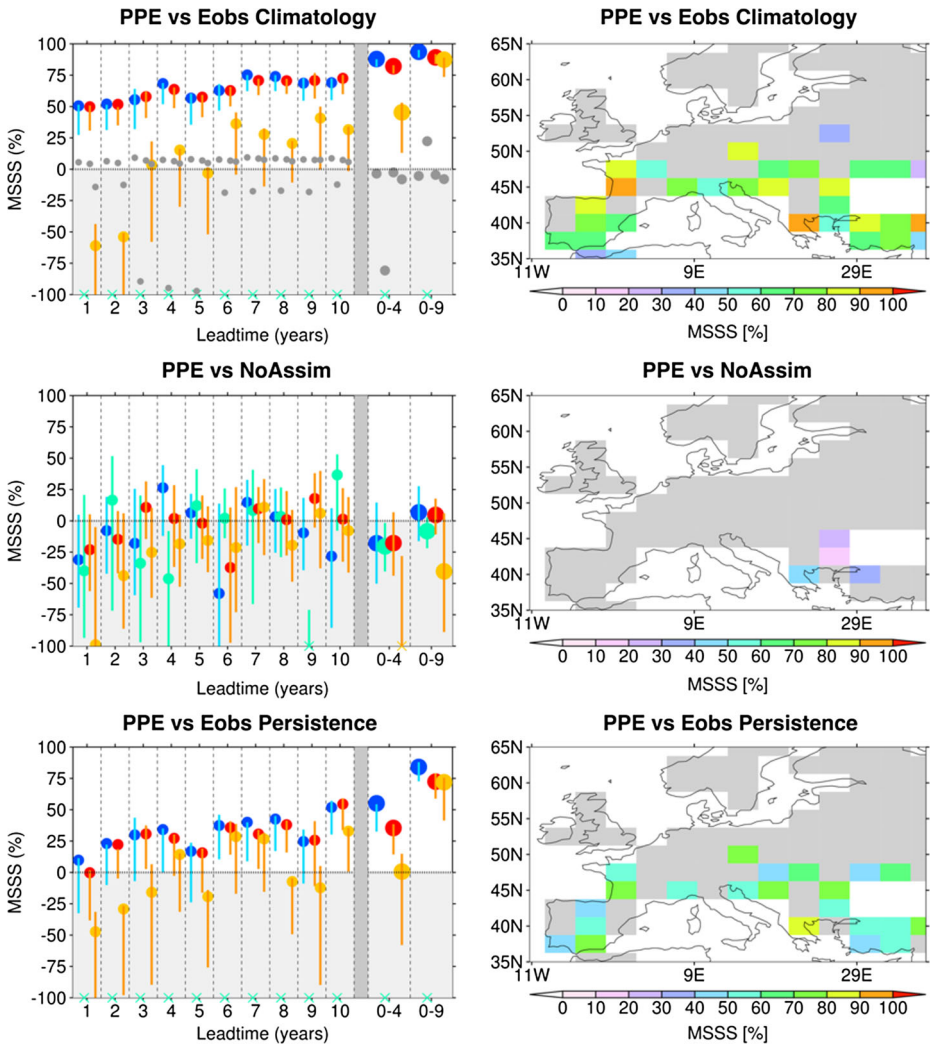


Fig. 2 Mean Square Skill Score of the TX29 computed with the initialised ensemble (PPE) mean compared to observed climatology (*upper*), non-initialised ensemble (*middle*) and persistence (*lower*), for all leadtimes and 0–4, 0–9 yr average leadtimes. Each symbol represents the individual score for each regional average, which have associated error bars (Europe is *blue*, Mediterranean is *red*, Central Europe is *orange* and UK is *green*). The *smaller grey symbols* in the left hand panels show the MSSS obtained by a random forecast. The *right hand panels* show a map of the scores calculated for each grid point for the 0–9 leadtime average, any grid points for which the score is not statistically significant are *shaded grey*

for these further indices are shown in the supplementary material as they are not directly relevant to a specific impact but do highlight some interesting features.

The more moderate Tmax index TX25 shows more grid points (mostly in central/southern Europe) with significantly skilful scores (see supplement). The regional average of TX25 shows more skill for Central Europe than TX29 but the other regions similar skill scores. The likely reason for these differences is that the trend in TX25 is larger than TX29 giving a stronger signal.

TN18 has a weaker trend, especially in the UK, and the model predictions are biased higher than observations rather than lower as seen for TX25 and TX29. This has resulted in lower skill scores for TN18 across the regions with only skill displayed for Europe and Mediterranean regions, none for the smaller regions Central Europe and UK (see supplement).

The variation in indices based on T_{min} (eg. TN18) rather than those based on T_{max} (eg. TX29 and TX25) is an interesting feature and is likely due to different physical processes governing daytime/night-time temperatures. T_{max} is usually recorded during day when temperatures are driven by the incoming short-wave solar radiation. Where as, T_{min} is usually recorded during the night and is governed by the outgoing long-range radiation which allows the surface to cool. The amount of incoming solar radiation which reaches the earth is dependent on the level of cloud cover and the radiative cooling during the night. Both are affected by the amount of moisture in the atmosphere, so these processes can be further affected by land surface-atmosphere interactions involving water exchange. The difference in skill for these two variables could be a result of the model reproducing one of these processes better than the other, hence it is important to consider these variables separately and to not assume skill in one index implies skill in the other.

Analysis of the combined hot days and warm nights index has shown very few occurrences of the relevant thresholds being exceeded across all the regions. The main issue with indices that have very few occurrences arises when assessing the skill of predicting an event that does not occur often, in either the model or observations. It is possible to obtain a very high skill but all that explains is that the index is mostly zero in both datasets. It yields no evidence as to whether the model could skillfully predict a future change.

Due to there also being a lack of spatial consistency across the region and large variation in the vulnerability at smaller scales, the CHDWN index is not considered further in this study. The decadal prediction system is not designed to look at such small scales, so trying to predict this index with the model would not be appropriate. To study the predictability of such an index relevant to human health impacts, it would be beneficial to use a prediction system with considerably higher spatial resolution or an additional downscaling method. For more detailed analysis and figures showing the climatology and trend in this index please refer to the supplement.

5.3 Cooling degree days (CDD)

From investigation of the TX29 and CHDWN indices, it is clear that the predictions made with a global model of this resolution should consider indices which are relevant when the average over a large spatial scale is taken. A good example of such an index is heating degree days, as used to estimate energy consumption in Winter. It is calculated based on the number of degrees the daily mean temperature is below an absolute temperature threshold and this is averaged over a large area for the season. However, as this index is already regularly used and understood for Europe, and because we are concerned with hot temperature extremes in this study, we investigate Cooling Degree Days (CDD). CDD is an index widely used to estimate power consumption in the US during warm seasons, calculated as the sum of degrees where the daily mean temperature exceeds an absolute temperature threshold (of 18 °C (65 °F)).

The trend of the index calculated with observed temperatures is slightly negative in the 1950–1980 period (Fig. 3 middle left) to largely positive in the 1980–2012 period (Fig. 3, middle right). This change in trend can also be clearly in the time series for the regional average CDD (Fig. 3, lower left) for the Mediterranean, Central Europe and Europe regions.

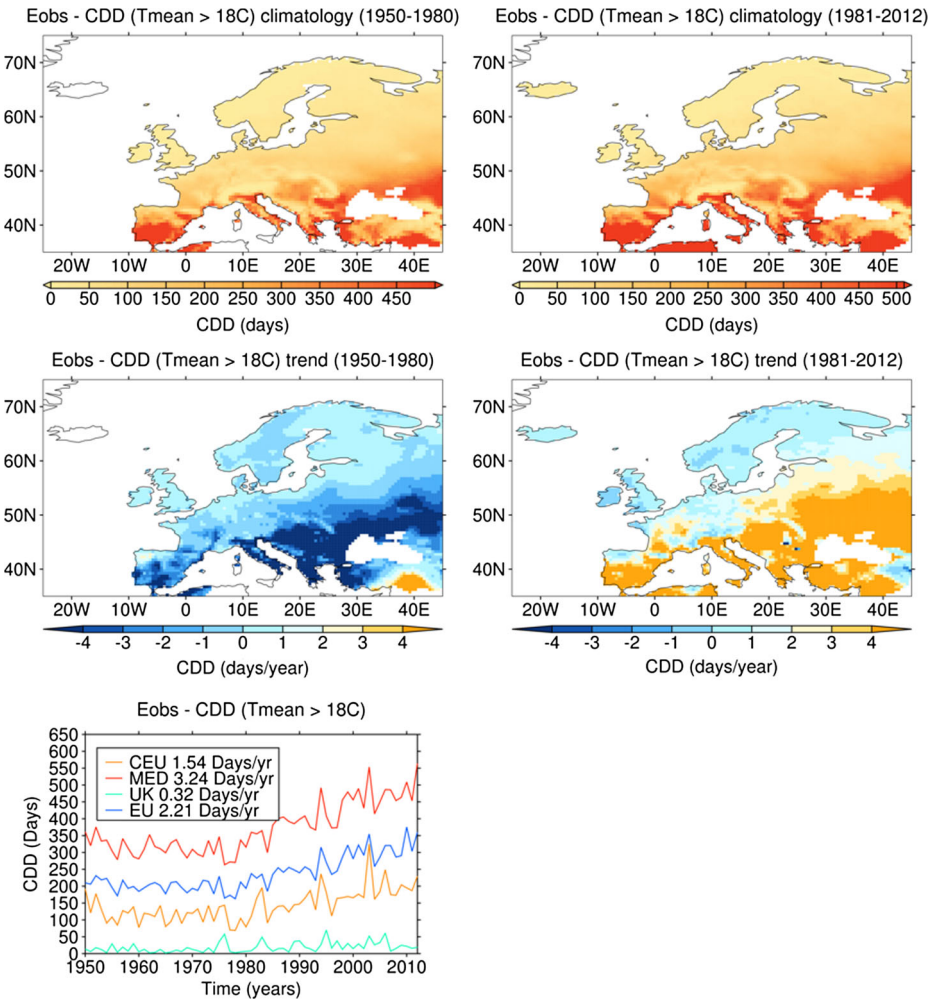


Fig. 3 Cooling Degree Days (CDD) computed with observations. *Upper plots* show maps of climatology of CDD index (1950-1980 (left) 1981-2012(right)). Cooling degree days is the cumulative sum of degrees above a daily threshold (Tmean greater than 18 °C) (Upper plot). *Middle plots* show maps of the trend in CDD over the two periods (1950-1980 (left) 1981-2012(right)). *Lower plot* shows a timeseries of CDD over 1950-2012 computed with observations. *Red* is the regional average over the Mediterranean region, *Blue* is Europe, *Orange* is Central Europe and *Green* is the UK

The climatology of the two periods maintain the same spatial pattern, the later period (1981-2012) showing a slight increase in average CDD (compare Fig. 3 upper left and upper right). The index remains small in magnitude for UK across both periods compared to more southerly regions. The initialised predictions show similar trends (see Fig. 1, right panels).

Similar to the TX29 index there are some grid points that show significant skill when compared to observed climatology (Fig. 4, upper right) and persistence (Fig. 4, lower right), these are mostly in Southern Europe. As there is no statistically significant skill consistently across grid points, relying on information from a single grid point would not be advised

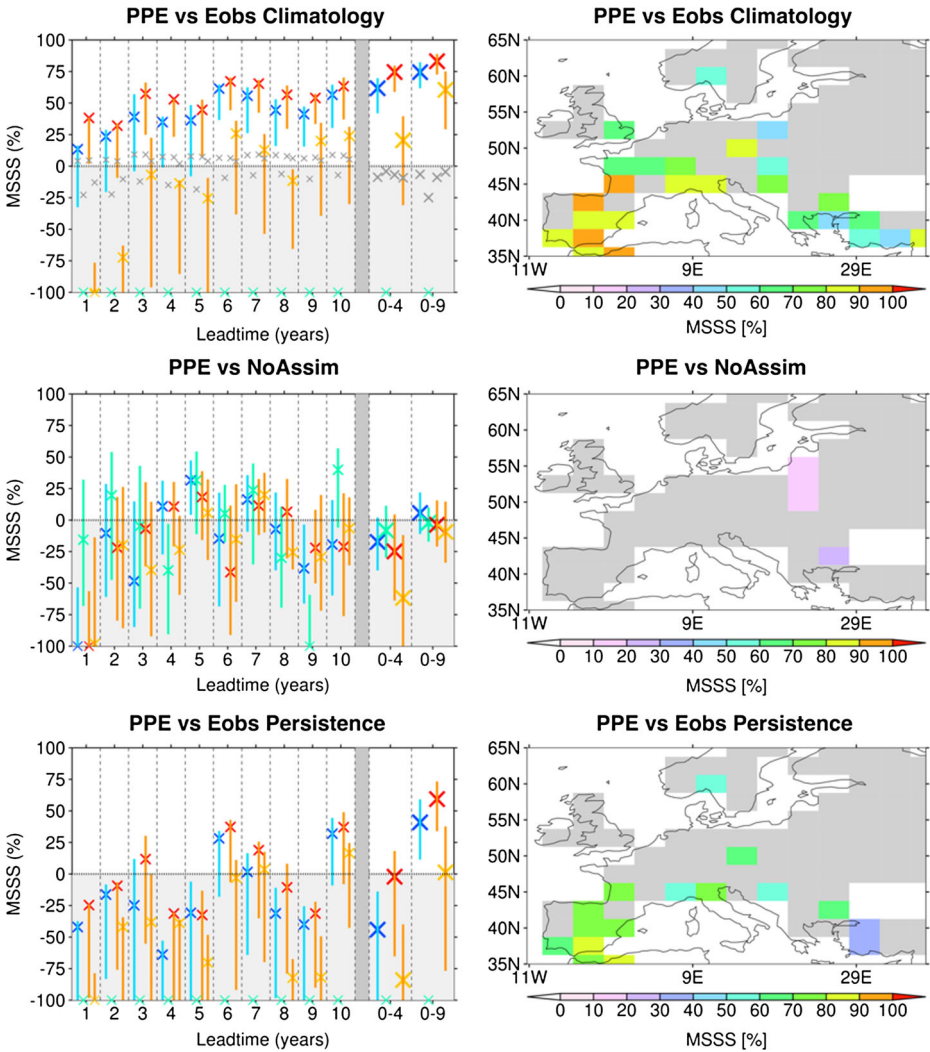


Fig. 4 Mean Square Skill Score of the index cooling degree days (CDD) computed with the initialised ensemble (PPE) mean compared to observed climatology (*upper*), non-initialised ensemble (*middle*) and persistence (*lower*), for all leadtimes and 0–4, 0–9 yr average leadtimes. Each symbol represents the individual score for each regional average, which have associated error bars (Europe is *blue*, Mediterranean is *red*, Central Europe is *orange* and UK is *green*). The *smaller grey symbols* in the left hand panels show the MSSS obtained by a random forecast. The *right hand panels* show a map of the scores calculated for each grid point for the 0–9 leadtime average, any grid points for which the score is not statistically significant are *shaded grey*

here either. Also, comparison of skill with the non-initialised ensemble shows no benefit here coming from initialisation of the model with observations (Fig. 4, middle right).

Regional average CDD is significantly skilful compared to observed climatology for the European and Mediterranean regions at later lead-times and also the 5 and 10 year averages (Fig. 4, upper left). This is most likely due to the steady positive trend in the Southern Europe providing a strong signal, the magnitude of which is recreated well by the model.

In the other regions the magnitude of this index and the trend are smaller and therefore harder to resolve from variability. So the only other area we see with statistically significant skill is the 10-year average for Central Europe. Comparison with the non-initialised ensemble runs shows that this skill is coming purely from the trend rather than the initialisation (Fig. 4, middle left). Alternatively, when the skill compared to the observed persistence of CDD is considered (Fig. 4, lower left), it is found that the only case in which the model is more significantly skilful is for the 10-year averages for Europe and the Mediterranean. Which suggests, in this case, observed persistence (the previous year's value) may be a better predictor for the index than the decadal model prediction.

6 Conclusions

To summarise the results of this study, we find that climate model-based near-term predictions have skill that exceeds that of forecasts based on observed climatology and persistence for indices of hot extremes over parts of Europe, and the European and Mediterranean regional averages. Less skill is found for the Central Europe region with only the decadal average of daily maximum temperatures greater than 25 °C and 29 °C, and CDD, showing significant skill. There is no significant skill found for the UK indices as the observations mostly lie outside the range of the model and the model overestimates the trends in the UK indices. So to conclude, it would be appropriate to use this model to predict the exceedances of these thresholds for the Mediterranean and Central Europe for the next ten years, but not for the UK.

Predictions of rarer events are less skillful than predictions of more moderate events, hence the predictability of the exceedance of absolute thresholds can vary with location depending on how rare an extreme it is in those areas. It is still possible to obtain skillful predictions for more moderate extremes for the UK region, as demonstrated by H13 and Eade et al. (2012).

There is no significant improvement of the skill in the prediction of these indices due to initialisation with this model. However it has been previously shown that initialisation in the MPI model has increased the skill in predictions of European summer temperatures (Matei et al. 2012) and temperature extremes (Hanlon et al. 2013b).

H13 considered several aspects of the methodology for deciding how best to evaluate the decadal predictions of extreme indices. We have found the same issues important in this study also. Firstly, the model prediction needs to be bias corrected for each index as this is more effective than applying the correction to the daily data. Secondly, the skill should be computed for the index in question, rather than relying on skill assessments on mean climate conditions (eg. regional average, seasonal average Tmean). Performing the skill assessment for each case individually, allows for the skill to vary depending on the capabilities of the model which vary with spatial scale, location, variable and how rare the event is. Following this advice also mean that the conclusions made in this paper should not be transferred to more local scales, they are valid for the regional averages over which they have been computed. However, that does not mean knowledge of the performance of the global model for the large scale cannot be helpful for more local scales. This method can help inform which global model may be the most appropriate to drive downscaling/impact models and in turn, the same skill assessment method could also be used to assess those downscaled results.

Following on from this study, to better quantify/constrain the uncertainty on these results there are several improvements that could be made to the methodology. Firstly, study a wider range of models to better quantify the model uncertainty or learn which models are most

skillful for particular cases. It would also be interesting to study how well the large-scale circulation is reproduced as this strongly affects the extremes (Kenyon and Hegerl 2008). As such, the multi-model average does not necessarily give a more skillful result than an individual model (Hanlon et al. 2013b). Larger ensembles would provide a better sampling of the variability and therefore a better quantification of the uncertainty. It would also be advantageous to repeat the study with a higher resolution model, possibly a regional model or employ some downscaling techniques to better recreate the physical processes at smaller spatial scales so more precise predictions can be made.

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 DECC/Defra Met Office Hadley Centre Climate Programme (GA01101). 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. Along with the UK Met Office Hadley Centre for the DePreSys dataset. Also, 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, along with two anonymous reviewers for providing useful advice.

References

- Barriopedro D, Fischer EM, Luterbacher J, Trigo RM, Garcia-Herrera R (2011) The hot summer of 2010: redrawing the temperature record map of Europe. *Science* 332:220–224
- Christidis N, Stott PA, Jones GS, Shioyama H, Nozawa T, Luterbacher J (2012) Human activity and anomalously warm seasons in Europe. *Int J Climatol* 32(2):225–239
- Díaz J, Linares C, Tobías A (2006) Impact of extreme temperatures on daily mortality in Madrid (Spain) among the 45–64 age-group. *Int J Biometeorol* 50(6):342–348
- UK Department of Health/ NHS (2012) Heatwave Plan for England 2012, <http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance>
- Eade R, Hamilton E, Smith DM, Graham RJ, Scaife AA (2012) Forecasting the number of extreme daily events out to a decade ahead. *J Geophys Res* 117(D21110)
- Efron B, Tibshirani RJ (1993) *An introduction to the bootstrap*. Chapman And Hall
- van der Linden P, Mitchell JFB (2009) ENSEMBLES: climate change and its impacts: summary of research and results from the ENSEMBLES project <http://ensembles-eu.metoffice.com/index.html>. Accessed 12 Apr 2012
- Fink AH, Brücher T, Krüger A, Leckebusch GC, Pinto JG, Ulbrich U (2004) The 2003 European summer heatwaves and drought - synoptic diagnosis and impacts. *Weather* 59(8):209–216
- Fischer EM, Schär C (2010) Consistent geographical patterns of changes in high-impact European heatwaves. *Nat Geosci* 3:398–403
- Fouillet A, Rey G, Laurent F, Pavillon G, Bellec S, Guihenneuc-Jouyaux C, Clavel J, Jouglé E, Hémon D (2006) Excess mortality related to the August 2003 heat wave in France. *Int Arch Occup Env Health* 80(1):16–24
- Grize L, Hussa A, Thommen O, Schindler C, Braun-Fahrlander C (2005) Heat wave 2003 and mortality in Switzerland *Swiss medical weekly*, vol 135
- Goddard L, Kumar A, Solomon A, Smith D, Boer G, Gonzalez P, Deser C, Mason S, Kirtman B, Msadek R, Sutton R, Hawkins E, Fricker T, Kharin S, Merryfield W, Hegerl G, Ferro C, Stephenson D, Meehl GA, Stockdale T, Burgman R, Greene A, Kushnir Y, Newman M, Carton J, Fukumori I, Vimont D, Delworth T (2013) A verification framework for interannual-to-decadal predictions experiments. *Clim Dyn* 40: 245–272
- Gordon C, Cooper C, Senior CA, Banks H, Gregory JM, Johns TC, Mitchell JFB, Wood RA (2000) The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. *Clim Dyn* 16(2):245–272
- Hamilton E, Eade R, Graham RJ, Scaife AA, Smith DM, Maidens A, MacLachlan C (2012) Forecasting the number of extreme daily events on seasonal timescales, vol 117, D03114
- Hanlon H, Hegerl GC, Tett SFB, Smith DM (2013a) Can a decadal forecasting system predict temperature extreme indices? *J Clim* 26(11):3728–3744

- Hanlon H, Morak S, Hegerl GC (2013b) Detection and prediction of mean and extreme European summer temperatures with a CMIP5 multi-model ensemble. *J Geophys Res Atmospheres* 118(17):9631–9641
- Haylock MR, Hofstra N, Klein Tank AMG, Klok EJ, Jones PD, New M (2008) A European daily high-resolution gridded data set of surface temperature and precipitation for 1950. *J Geophys Res* 113:D20:D20119:10
- Hawkins E, Osborne TM, Ho CK, Challinor AJ (2013) Calibration and bias correction of climate projections for crop modelling: an idealised case study over Europe. *Agric For Meteorol* 170:19–31
- Jenkins GJ, Perry MC, Prior MJ (2008) The climate of the United Kingdom and recent trends. UKCP09 scientific reports
- Kenyon J, Hegerl GC (2008) Influence of modes of climate variability on global temperature extremes. *J Clim* 21:3872–3889
- Marsland SJ, Haak H, Jungclaus JH, Latif M, Roske F (2003) The Max-Planck-Institute global ocean/sea ice model with orthogonal curvilinear coordinates. Technical Report 5. *Ocean Model*. 5: 91–127
- Matei D, Pohlmann H, Jungclaus J, Muller W, Haak H, Marotzke J (2012) Two tales of initializing decadal climate prediction experiments with the ECHAM5/MPI-OM model. *J Clim* 25(24):8502–8523
- Morak S, Hegerl GC, Christidis N (2013) Detectable changes in the frequency of temperature extremes. *J Clim* 26:1561–1574
- Murphy AH (1988) Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Weather Rev.* 16:2417–2424
- Murphy JM, Sexton DMH, Barnett DN, Jones GS, Webb MJ, Collins M, Stainforth DA (2004) Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature* 430(7001):768–772
- National Climatic Data Center (2012) HISTORICAL CLIMATOLOGY SERIES 5-2 (January 2010 - December 2011) NCDC, <http://www1.ncdc.noaa.gov/pub/data/hcs/cdd.201001-201112.pdf>. Accessed 26 Feb 2014
- Pascal M, Laaidi K, Ledrans M, Baffert E, Caserio-Schönemann C, Le Tertre A, Manach J, Medina S, Rudant J, Empereur-Bissonnet P (2006) France's heat health watch warning system. *Int J Biometeorol* 50(3):144–153
- Pope VD, Gallani ML, Rowntree PR, Stratton RA (2000) The impact of new physical parametrizations in the Hadley Centre climate model: HadAM3. *Clim Dyn* 16(2):123–146
- Raddatz TJ, Reick CH, Knorr W, Kattge J, Roeckner E, Schnur R, Schnitzler KG, Wetzel P, Jungclaus J. (2007) Will the tropical land biosphere dominate the climate-carbon cycle feedback during the twenty-first century? *Clim Dyn* 29(6):565–574
- Schär C, Vidale PL, Lüthi D, Frei C, Häberli C, Liniger M, Appenzeller C (2004) The role of increasing temperature variability for European summer heat waves. *Nature* 427(6972):332–336
- Schlenker W, Roberts MJ (2009) Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *PNAS* 106(37):15594–15598
- Smith DM, Murphy JM (2007) An objective ocean temperature and salinity analysis using covariances from a global climate model. *J Geophys Res* 112(C2):C02022
- Smith DM, Cusack S, Colman AW, Folland CK, Harris GR, Murphy JM (2007) Improved surface temperature prediction for the coming decade from a global climate model. *Science* 317(5839):796–799
- Smith DM, Eade R, Dunstone NJ, Fereday D, Murphy JM, Pohlmann H., Scaife AA (2010) Skilful multi-year predictions of atlantic hurricane frequency. *Nat. Geosci.* 3:846–849
- Smith DM, Eade R, Pohlmann H (2013) A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction. *Clim Dyn* 41:3325–3338
- Uppala SM, Kallberg PW, Simmons AJ, Andrae U, da Costa Bechtold V, Fiorino M, Gibson JK, Haseler J, Hernandez A, Kelly GA, Li X, Onog K, Saarinen S, Sokka N, Allan RP, Andersson E, Arpe K, Balmaseda MA, Beljaars ACM, Vande Berg L, Bidlot J, Bormann N, Caires S, Chevallier F, Dethof A, Dragosavac M, Fisher M, Fuentes M, Hagemann S, Holm E, Hoskins BJ, Isaken L, Janssen PAEM, Jenne R, McNally AP, Mahfouf JF, Morcrette JJ, Rayner NA, Saunders RW, Simon P, Sterl A, Trenberth KE, Untch A, Vasiljevic D, Viterbond P, Woollen J (2005) The ERA-40 re-analysis. *Q J Roy Meteorol Soc* 131:2961–3012
- WCRP (2011) Data and bias correction for decadal climate predictions International CLIVAR Project Office Publication Series no.150, <http://www.wcrp-climate.org/decadal/publications.shtml>. Accessed 30 Aug 2011
- Zwiers FW, Zhang X, Feng Y (2011) Anthropogenic influence on long return period daily temperature extremes at regional scales. *J Clim* 24(3):881–892