JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 117, D03114, doi:10.1029/2011JD016541, 2012

Forecasting the number of extreme daily events on seasonal timescales

Emily Hamilton,[1] Rosie Eade,[1] Richard J. Graham,[1] Adam A. Scaife,[1] Doug M. Smith,[1] Anna Maidens,[1] and Craig MacLachlan[1]

Received 12 July 2011; revised 2 November 2011; accepted 1 December 2011; published 8 February 2012.

[1] We investigate the potential for skillfully predicting the number of daily temperature extremes over 3 month (seasonal) periods. We use retrospective forecasts from the Met Office seasonal forecasting system, GloSea4, nominally initialized 1 month ahead of the target season. Initially, we define daily extremes to be events outside either the upper or lower deciles of the daily temperature distribution from the relevant season. This definition provides a threshold that is sufficiently "extreme" to be of interest to many users but moderate enough to allow a sufficient sample for verification and to be of regular use to users. We show that skill reduces slightly at more extreme thresholds. Correlations of predicted and observed numbers of upper or lower decile extreme days over a season are significantly greater than zero over much of the globe and, in general, are better than a persistence forecast. Forecast skill for seasonal mean temperature is similar to, but generally greater than, the skill of predictions of the number of extreme days. Observations have a strong relationship between the seasonal mean and the number of extreme days. We show that the skill in predicting the number of extreme days is largely a consequence of this relationship and occurs primarily through a shift in the distribution of the daily data rather than a change of its shape. The ability to predict the El Niño–Southern Oscillation and climate change are both significant contributors to the skill in predicting temperature extremes. In summer, significant skill also comes from initializing soil moisture.

Citation: 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.

1. Introduction

[2] We investigate the skill of the current Met Office seasonal prediction system, GloSea4 [*Arribas et al.*, 2011], in forecasting the number of extreme daily temperature events within a target 3 month season predicted at a nominal 1 month lead. Initially, extreme warm days are defined as all days exceeding the 90th percentile of the climatological daily temperature distribution for the season and extreme cold days are similarly defined as not exceeding the 10th percentile. An investigation using similar methodology (R. Eade et al., manuscript in preparation, 2012) has been carried out for both temperature and precipitation extremes over multiple seasons using the Met Office decadal prediction system (DePreSys) [*Smith et al.*, 2007].

[3] To date, there have been few attempts to predict the number of extreme daily events on a seasonal time scale. Indeed, there are no studies, as far as we are aware, on prediction of daily temperature extremes on seasonal timescales, although *Gershunov and Barnett* [1998] looked at how the number of daily extreme temperatures over the United States in ECHAM3 varied with prescribed El Niño–Southern Oscillation (ENSO) forcings. There have been several attempts to assess the skill in forecasting precipitation extremes at seasonal lead times, for example, studies by *Robertson et al.* [2009] and *Zeng et al.* [2010].

[4] The focus of this article is on moderate extremes, allowing robust validation of results. To give an idea of their magnitude, in northern Europe (land points in 15°W–30°E and 45°N–65°N), these moderate extremes relate to daytime temperatures of approximately 26°C in summer and −9.8°C in winter. These have been calculated as the 90th and 10th percentiles of Tmax and Tmin, respectively, from the observational data set used in this study (HadGHCND from 1989 to 2009) [*Caesar et al.*, 2006]. The thresholds are of great interest for numerous practical applications where threshold exceedance is relevant. In many applications, the threshold of interest will be constant throughout the season, such as in the rail industry where each piece of equipment is designed to withstand particular extreme temperatures [*British Standards Institute*, 2001]. Often susceptibility depends on locally relevant thresholds in a particular region, which may be irrelevant elsewhere in the world. In human health, for example, acclimatization means that people living in warm climates have a higher heat-related mortality threshold than those living in cool climates [*Gosling et al.*,

[1] Met Office Hadley Centre, Exeter, UK.

2007; *Curriero et al.*, 2002]. Adaptation in agriculture also leads to variations in the relevant absolute temperature threshold of interest; for example, citrus fruits are grown in warm climate and are susceptible to frosts [*Rogers and Rohli*, 1991], whereas cattle can be reared in a climate with extremely harsh winters and can survive temperatures down to around −35°C [*Young*, 1981]. To generalize our investigation to a global assessment, we use a percentile approach, whereby the extreme temperature is relevant to the climatology of the region. A discussion on how skill varies with the chosen threshold is also included.

[5] This article is structured as follows: We begin by describing a simple method to predict extreme temperature days using the daily raw model output from GloSea4. Our methods have the advantage that it is not necessary to apply a bias correction to the data. We then compare the skill of GloSea4 in predicting the proportion of extreme temperature days in a season with the skill of forecasting the seasonal mean temperature. By introducing a second method for predicting the number of extreme temperature days that relies only on the prediction of the seasonal mean temperature, we show that predictability of the extremes is almost entirely a result of predictability of the seasonal mean temperature. We show that it is a change in the mean of the distribution rather than in the shape that produces skill. We also show how this predictability varies with the extremity of the threshold considered.

[6] The ENSO phenomenon is well documented as being associated with extreme climate events [*Philander*, 1990]. Therefore, we assess how much of the skill of the system in predicting the frequency of extreme days is because of the correct prediction of ENSO. We also assess the role played by representation of the temperature response to climate change and by the initialization of summer soil moisture.

[7] Finally, we consider an alternative definition of extreme temperature where the threshold changes daily, rather than being static throughout the season. Prediction of this moving-threshold extreme temperature would be useful for users whose threshold does not fit into a standard season or whose susceptibility or coping capacity changes as their exposure increases (as defined by *Taubenböck et al.* [2008]). An example of this is in human health applications because of acclimatization [*Taylor*, 2006]. The skill in predicting the moving-threshold extreme is compared to that of predicting the static-threshold extreme described previously.

2. Data

2.1. GloSea4

[8] Hindcasts of daily minimum and maximum near-surface temperature (Tmin and Tmax) initialized each week over the period 1989–2009 from the GloSea4 system [*Arribas et al.*, 2011] were used. For each hindcast year and initialization week, there are three ensemble members. The initial conditions of these three members are identical; they are perturbed using stochastic physics [*Bowler et al.*, 2008] during the integration. To take some account of initial condition uncertainty, a time-lagged approach was used: Data from three hindcast start dates (nine members in total) were combined to make predictions for each of four seasons (December–February (DJF), March–May (MAM), June–August (JJA), September–November (SON)). The three start dates were consecutive weeks centered on a 1 month lead (and so are hereafter referred to as 1 month lead forecasts).

[9] The atmosphere model used here has a horizontal resolution of 1.25° latitude × 1.875° longitude. The initialization of the hindcasts is as follows: ERA-Interim [*Dee et al.*, 2009] is used to initialize the atmosphere and land surface. The ocean is initialized using a version of the Met Office optimal interpolation scheme used for short-range ocean forecasting [*Martin et al.*, 2007]. For GloSea4 hindcasts initialized up to and including the year 2005, levels of climate forcings (aerosols, methane, CO_2, etc.) are set equal to observed values. After this, they follow the Intergovernmental Panel on Climate Change A1B scenario [*Intergovernmental Panel on Climate Change*, 2001]. Ozone is fixed to observed climatological values and includes a seasonal cycle. We use the latest upgrade to GloSea4, using model version HadGEM3-AO GA 2.0 (A. Arribas et al., manuscript in preparation, 2012). The main improvements are explicitly initialized sea ice and improved vertical resolution in the atmosphere to represent the stratosphere and in the ocean to better represent the mixed layer and diurnal cycle.

2.2. GloSea3

[10] The predecessor to GloSea4, GloSea3, has a larger number of ensemble members available in its hindcasts (15 per start date compared to a total of 9). For this reason, GloSea3 hindcasts were used here to assess the robustness of the results to the number of ensemble members. GloSea3 hindcasts of daily Tmin and Tmax are available initialized on the first day of each month over the period 1987–2007.

[11] In GloSea3, initial condition uncertainty is represented using wind stress and sea surface temperature perturbations designed to represent observed uncertainties in these parameters. There is no further representation of model uncertainty. The atmospheric component of the GloSea3 system [*Pope et al.*, 2000] has a horizontal resolution of 2.5° latitude × 3.75° longitude. The system is based on the HadCM3 model [*Gordon et al.*, 2000]. Unlike GloSea4, climate forcings are constant for all hindcast years in GloSea3.

2.3. Observations

[12] Observational data used were the daily observed Tmin and Tmax fields from the HadGHCND data set [*Caesar et al.*, 2006]. This covers the period 1989–2010 on a 2.5° latitude × 3.75° longitude grid. The coverage is incomplete both spatially and temporally, so skill assessment has not been carried out at any grid point with more than 10% of data missing over the entire period. Moreover, at each grid point, any season in any year with more than 10% of the data missing is excluded from the analysis. We call the remaining days “nonmissing.” No data is available over the oceans, and these restrictions have resulted in no analysis being carried out for the Southern Hemisphere.

3. Methodology

3.1. Static Threshold or Moving Threshold

[13] Two definitions of an extreme daily temperature are used in this article. In the first definition, thresholds are constant throughout a season, resulting in the probability of

exceedance varying throughout the season (e.g., 90th percentile exceedances are less likely at the beginning of spring than at the end). However, in total, 10% of the days in the season exceed (do not exceed) the 90th (10th) percentile thresholds. Days that meet this criterion are called "static-threshold exceedances." The threshold is calculated for observations and hindcasts separately using all of the daily data available over the hindcast period, 1989–2009, for the relevant season. Separate calculation of the thresholds in the observations and the hindcasts means that no further action need be taken to remove model biases.

[14] We also define "moving-threshold exceedances." Here, each day in a season is defined to be extreme if it exceeds the relevant percentile of its own daily temperature distribution. This means that a priori each day during a season has an equal chance of being extreme. For the same percentile, the expected number of exceedances in a complete season remains the same as for static-threshold exceedances.

[15] To create these moving thresholds, the temperatures relating to the relevant percentile are found from the daily data for each calendar month for all years covered by the hindcast period, separately for observations and hindcasts. These monthly thresholds are then smoothed into daily thresholds using a fast Fourier transform with a half-power of 8 days (Eade et al., manuscript in preparation, 2012). This results in daily thresholds where the daily probability of exceedance is approximately constant throughout the year, and the thresholds are continuous at the annual and seasonal boundaries.

3.2. Calculating the Number of Extremes in Observations and Hindcasts

[16] The numbers of extreme days observed during each season of the hindcast period were counted by first calculating the relevant threshold from the observations at each grid point. For example, the static thresholds for the SON season were obtained at each grid point by calculating the 90th percentile of the observed distribution of daily Tmax from all days in September, October, and November for all years of the hindcast period. Then, for each year we calculated the proportion, P, of the observed days in the season that exceeded this threshold. At each grid point

$$P = \frac{\text{Number of days exceeding threshold}}{\text{Number of “nonmissing” days in season}}.$$

For static-threshold extremes, the predicted proportion of the season that exceeds the same percentile was calculated using one of the two methods (described in the following). Only the first of these methods is used for moving-threshold extremes.

[17] In Method 1, the value corresponding to the observed percentile threshold was calculated in the hindcasts using all daily data from the relevant season in the hindcast period. Then for each year in the hindcast set, the proportion of the days in the season exceeding this threshold was calculated.

[18] In method 2, the predicted proportion of the season exceeding the hindcast threshold was inferred using the forecasted seasonal mean temperature anomaly and the historic observed relationship between the observed seasonal mean temperature anomaly and the number of threshold exceedances. The derivation of this relationship is described later.

[19] For comparison, persistence predictions were made with an equivalent lead time; that is, by persisting the 3 month period that ends before the beginning of the forecast start date. For example, the persistence forecast for the proportion of hot days in JJA was taken as equal to the proportion of hot days in February, March, and April, the last complete 3 month period before the forecast issue in May. A persistence prediction where the proportion of days exceeding the threshold was taken from the same season in the previous year was also tested. The skill of this latter persistence approach was much lower (not shown).

3.3. Skill Calculation

[20] In this article, we assess the "meteorological" skill in forecasting the number of extreme days, rather than trying to address the value of the forecast to users. A simple skill measure, the Spearman's rank correlation coefficient (hereafter, correlation) was therefore chosen. This correlation coefficient was chosen, rather than the more well-known Pearson's correlation coefficient, as it is more appropriate for count data.

[21] At each grid-square, the correlation was calculated from the prediction fields as follows. The proportion of threshold exceedances (or the seasonal mean temperature anomaly) was calculated for each ensemble member for every year in the hindcast period, and an ensemble mean was calculated for each hindcast year. The corresponding proportions of threshold exceedances (seasonal mean temperature anomaly) were then calculated in the observations. These fields of predictions and observations were then smoothed in space (see section 3.4). The correlation between these two smoothed fields was then calculated at grid point level (where each point on the 2.5° × 3.75° grid lattice represents a region of 17.5° latitude × 18.75° longitude on the smoothed grid). This resulted in a global field of correlation for each forecast system and each type of exceedance event (later referred to as "combination"). Unless otherwise stated, we only consider 10th and 90th percentile extremes, so there are 16 possible extreme combinations (10th or 90th percentile with Tmin or Tmax for each of four seasons) and 8 possible combinations for seasonal mean temperature (pairs of Tmin and Tmax with each of four seasons). To summarize the information from these fields, the skill of each method was represented as a single field of mean correlation (where the mean was calculated over all combinations at grid point level; later called "grid point skill") and the global area–weighted averages of mean correlation (global average skill). Here, "global" refers only to grid points with nonmissing data, thus, corresponding to the Northern Hemisphere land. The differences between methods have been similarly represented as the differences of the grid point skill and the differences in global average skill. A bootstrapping technique [*Wilks*, 1995] has been used to assess whether skill and differences in skill between the methods discussed differ significantly from zero. This has been carried out at both the grid point and global average levels. If the correlation is greater than zero and the 90% confidence interval of the bootstrapped correlation does not include zero, then the correlation is judged significantly greater than zero at the 5% level (one-tailed test). Similarly,

if the 95% confidence interval of the differences in correlation does not include zero then the absolute differences in correlation are said to be significantly different from zero at the 5% level (two-tailed test). The p values for one-tailed tests have been calculated as $\alpha/2$ where $100(1 - \alpha)$ is the smallest confidence interval that contains zero; similarly, for two-tailed tests, the p value is α. The bootstrapping technique applied takes account of spatial correlations and correlations between types of extremes in calculating the significance of the global average skill. The effect on significance levels of temporal correlation was found to be negligible.

3.4. Smoothing

[22] To reduce noise and obtain robust results, smoothing of the observed and hindcast fields has been carried out. The procedure is as follows: First, if necessary, the data are regridded to the 2.5° × 3.75° grid using bilinear interpolation and masked so that missing data areas are identical in observations and hindcasts and only include land points. They are then smoothed using a 7 × 5 box (of size 17.5° × 18.75°); we apply the average of the 7 × 5 box to the grid point in the center point of the box. Grid points are set to "missing" if more than half of the data in the 7 × 5 box is missing.

4. Results

4.1. Does the Skill of Predicting the Number of Extreme Days Exceed That of Predicting the Seasonal Mean?

[23] First, we compare the skill of predicting the proportion of extreme days in the season (Figure 1a) with skill of predicting the seasonal mean temperature (Figure 1b).

[24] While small, the grid point skill in both the extreme and the mean is locally, statistically, and significantly positive throughout the majority of the assessed Northern Hemisphere (Figure 1). When averaged over all extreme definitions, there is also significant improvement over a persistence forecast for both the extreme and means (Table 1, last row). However, when broken down by season, only JJA and MAM exhibit skill significantly greater than a persistence forecast for all types of extremes (Table 1).

[25] Sensitivity of the global skill in predicting the means and in predicting the number of extreme days to the number of ensemble members is displayed in Figure 2. In addition to GloSea4 hindcasts, hindcasts from the GloSea3 prediction system have been used for this purpose, as there are a greater number of ensemble members available from this system. The results for both GloSea3 and GloSea4 show that global skill improves more noticeably when adding an extra ensemble member to a small ensemble. However, the skill appears to converge within the range of ensemble sizes considered. Of particular interest is that for both the mean and the extremes, the correlations are similar for an ensemble size of 9 and an ensemble size of 15. This gives some confidence that the results from the GloSea4 hindcast ensembles (of size 9) will be applicable to operational forecasts of size 42. It is also worth noting that improvements in skill have been made between GloSea3 and GloSea4.

4.2. Does the Skill in the Mean Explain the Skill in Predicting the Number of Extreme Days?

[26] The skill of method 2 shows to what extent the seasonal mean temperature can be used to predict the number of extreme days in a season. Comparison of the skill from method 1 with that from method 2 shows how much skill the daily data from the model is responsible for. Various monotonic increasing functions were considered for mapping the seasonal mean temperature on to the number of exceedances. These included the following: (1) an empirical CDF created by centralizing and combining the empirical CDFs of each season about the all-season climatological mean, (2) a Gaussian CDF whose parameters are estimated from the centralized sample in item 1, (3) the empirical CDF from item 1 is smoothed using a kernel density, and (4) a regression technique. Because of their construction, the correlation-based skill scores considered in this article are robust to the choice of function. The cross-validated trials method (denoted by item 1) had the smallest bias and root mean square error (not shown) and so is used in the remainder of this article.

[27] Methods 1 and 2 exhibit very similar levels of skill in predicting the number of extremes in a season. Indeed, there is no significant difference at the 15% level ($p = 0.16$) between the global skill obtained using methods 1 and 2. By comparing Figure 3a with 3b it can be seen that the spatial variations in skill are similar for methods 1 and 2 (with increased grid point skill around western Canada, southern USA and Southeast Asia). Table 1 shows that the global skill variations for methods 1 and 2 are similar for all combinations: Pearson's correlation coefficient between the skill scores of the two methods is 0.93 across the 16 types of extreme events.

[28] Analysis was also done using additional percentiles (1st, 2.5th, 10th, 25th, 50th, 75th, 90th, 97.5th, and 99th in total). Figure 4a shows the global average skill (averaged over all 16 combinations) of method 1 (red) and method 2 (black) plotted against the threshold percentile. This shows that the similarity in global skill between methods 1 and 2 holds at the full range of percentile thresholds. Note that there is an asymmetry in the skill of both methods, with greater skill demonstrated at the warm extremes than the cold extremes. Asymmetry of this nature has been seen in the simulation of extremes in global climate models, for example, as shown by *Kiktev et al.* [2003, 2007].

[29] The similarity in skill of the two methods suggests that skill in predicting the number of extreme days is very strongly driven by the ability to predict the seasonal mean temperature, even for very extreme thresholds. The daily data (used in method 1) do not improve the ability to rank the seasons successfully (or increase the skill score). This is partly caused by the strong relationship between the number of extremes and the mean in observations (Figure 4a, green). This strong relationship means that a change in the shape of the distribution of daily data over the season does not play a large role in determining the number of extreme days and that mean climate shifts from year to year dominate in determining the likelihood of extreme days.

[30] Using idealized analysis, we can estimate the potential skill benefit achieved by the following: (1) a perfect

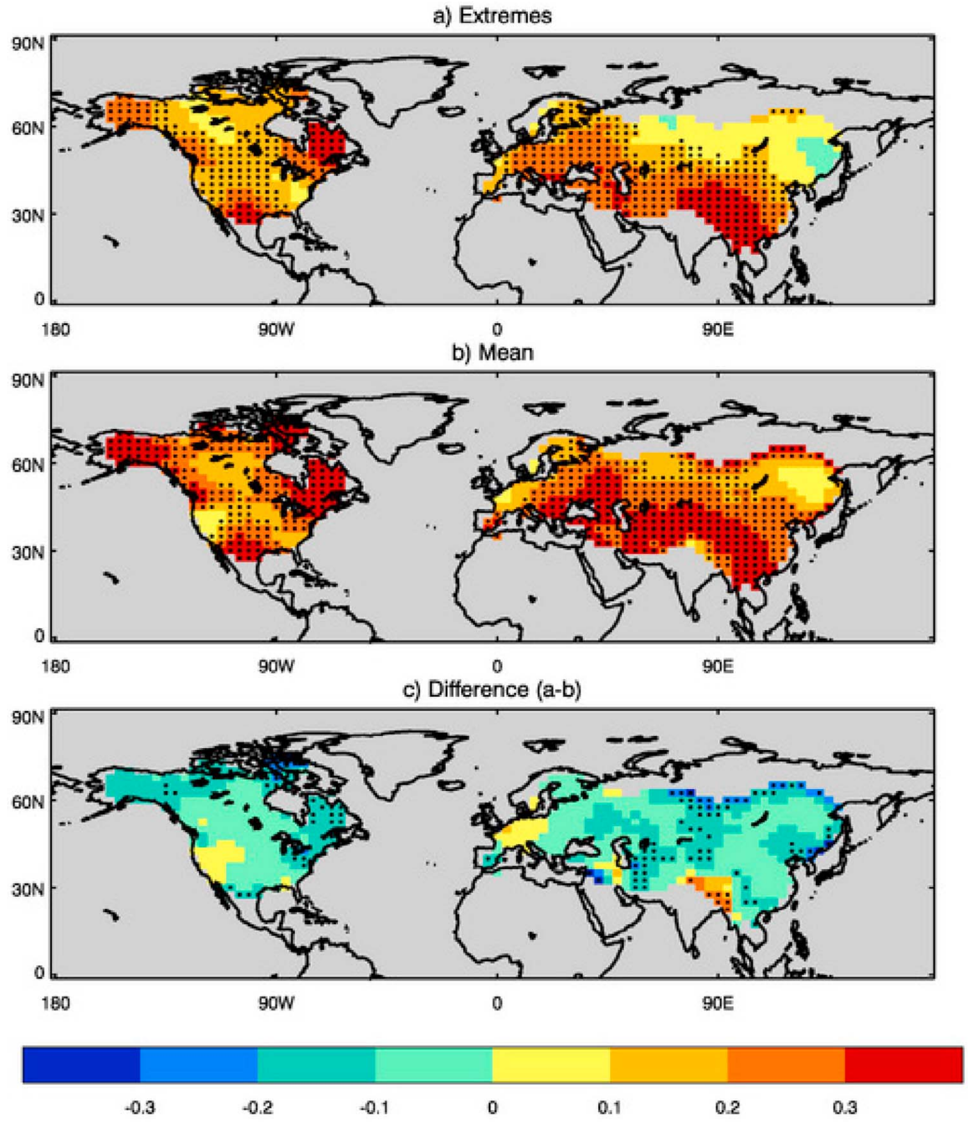

Figure 1. Contours of grid point skill for GloSea4's predictions of (a) the number of extreme days in a season and (b) the seasonal means, both averaged over all seasons and combinations. (c) The difference between the grid point skills shown in Figures 1a–1b. Black dots show areas with local significant difference from zero (at 5% level). Gray indicates missing data. The percentages of significant grid points are 64%, 77%, and 20%, for Figures 1a, 1b, and 1c, respectively. The global average skills are 0.20, 0.27, and −0.07, respectively. Here all global average skills/differences are significantly different from zero at the 1% level. For all significance tests a one-tailed test was used for Figures 1a and 1b and a two-tailed test for Figure 1c.

prediction of the distribution of daily data about its mean and (2) a perfect prediction of the seasonal mean.

[31] We investigate the two idealized predictions:

[32] 1. The "perfect daily data hindcast" is obtained by removing the seasonal mean from the observed daily data of each season and adding the corresponding hindcast mean (i.e., creating an idealized forecast where the distribution of daily data is as observed, but the skill in predicting the seasonal mean is the same as in the model hindcast).

[33] 2. The "perfect mean hindcast" is obtained by removing the seasonal mean from the hindcast daily data of each season and adding the corresponding observed mean (i.e., creating an idealized forecast where the seasonal mean is as observed but the distribution of daily data is the same as in the model hindcast).

[34] First, consider the perfect daily data hindcast (Figure 4b, blue). For the extreme percentiles, there potentially is an increase of approximately 0.3–0.4 in global skill to be gained by perfectly predicting the shape of the daily data (but not improving the skill of the mean), whereas toward the central percentiles, the potential increase is much lower (approximately 0.1). The extra potential predictability in the tails is a result of errors in the mean being less important here because of the scarcity of data at the edge of the distribution.

Table 1. Global Area-Weighted Average Spearman's Correlation for Each Combination[a]

	Percentile[b]	Method 1	Method 2	Persistence (ST)	MT	Persistence (MT)	Mean	Persistence (Mean)
				DJF				
Tmin	10	0.11	0.12	**0.21**	0.15	**0.14**	0.12	**0.18**
Tmin	90	0.15	*0.11*	0.08	*0.09*	0.13		
Tmax	10	0.13	*0.13*	**0.18**	0.14	**0.12**	0.19	**0.16**
Tmax	90	0.20	0.21	0.13	0.20	**0.16**		
				MAM				
Tmin	10	**0.28**	**0.32**	0.11	**0.22**	0.07	**0.28**	**0.14**
Tmin	90	**0.14**	0.11	0.02	**0.29**	−0.02		
Tmax	10	**0.26**	**0.32**	0.10	**0.29**	0.09	**0.33**	**0.18**
Tmax	90	**0.24**	**0.24**	0.07	**0.34**	0.09		
				JJA				
Tmin	10	**0.20**	**0.20**	**0.06**	**0.23**	**0.08**	**0.24**	0.08
Tmin	90	**0.28**	**0.30**	**0.17**	**0.32**	0.08		
Tmax	10	**0.26**	**0.25**	0.02	**0.30**	**0.18**	**0.34**	**0.13**
Tmax	90	**0.30**	**0.33**	**0.21**	**0.33**	**0.15**		
				SON				
Tmin	10	0.05	0.09	**0.09**	0.16	**0.14**	**0.29**	**0.15**
Tmin	90	**0.26**	**0.26**	**0.13**	**0.33**	**0.14**		
Tmax	10	0.08	0.13	**0.11**	**0.30**	**0.12**	**0.39**	**0.21**
Tmax	90	0.25	0.33	**0.28**	0.34	**0.27**		
Average		**0.20**	**0.22**	**0.12**	**0.25**	**0.12**	**0.27**	**0.15**

[a] For persistence, scores in bold font are significantly greater than zero. Otherwise, bold indicates that the score is significantly better than the relevant persistence forecast. Significance is assessed at the 5% level using a bootstrapping technique. For GloSea4 forecasts, scores shown in italic are not significantly greater than zero at the 5% level (assessed using a bootstrapping technique). Note that all GloSea4 forecasts are significantly greater than zero at the 10% level. DJF, December–February; MAM, March–May; JJA, June–August; SON, September–November; ST, static-threshold extremes; MT, moving-threshold extremes.

[b] The percentile column refers only to skill in predicting extremes, not skill in predicting the mean.

[35] Now consider the perfect mean hindcast (Figure 4b, green). The possible gain in skill is much greater than for the perfect daily data hindcast (around 0.5 for 10th and 90th percentiles compared to approximately 0.3). There is less to be gained for more extreme percentile thresholds (approximately 0.3) as the relationship between the extremes and the mean weakens. This is as a result of noise caused by the lack of data points in the tail of the distribution.

4.3. What Are the Physical Sources of the Prediction Skill?

[36] ENSO and the climate change signal are well-known sources of skill in seasonal forecast systems [*Doblas-Reyes et al.*, 2006; *Kiladis and Diaz*, 1989; *Liniger et al.*, 2007]. To assess the extent to which these factors contribute to the skill in extreme prediction in GloSea4, the contributions played by the representation of ENSO and climate change are removed from the hindcast. To do this, seasonal mean ENSO and climate change indices were first calculated for every member of the hindcast set. The ENSO index was taken to be the area-weighted average of sea surface temperatures (SSTs) in the Niño 3.4 region (bounded by 120°W–170°W, 5°N–5°S). Observed changes in global SSTs were taken to characterize the climate change index. Specifically, the following procedure was used: First, the global seasonal area-weighted average of SST was calculated for each member of the hindcast. The dependence of this index on ENSO was then removed through linear regression with the ENSO index previously defined. The two indices are therefore independent by construction. An example of the ENSO and climate change indices for DJF can be seen in Figure 5a. These have been calculated for a single member of the hindcast.

[37] Once the indices had been calculated for each season and hindcast member, the linear relationship between each of the indices and the seasonal mean temperature at each grid point was ascertained through leave-one-out cross-validated linear regression. This allowed seasonal mean temperature predictions to be made at each grid point using either one of the indices. These predictions were taken to be the part of the hindcast seasonal mean temperature that the forecast ENSO or climate change index is responsible for, or in other words, the response of seasonal mean temperature at each grid point to the ENSO or climate change index. To remove the contribution of ENSO or climate change from the predictions of the numbers of extremes, this response was subtracted from the daily temperatures in the hindcast for every day of each season. This analysis, therefore, has similarities with the "perfect mean" analysis of section 4.2, in that entire seasons of daily data are moved so that they have a new mean, in this case determined by the relevant index. The 10th and 90th temperature thresholds were recalculated from this adjusted daily data, and the number of extremes counted using method 1. The average skill over all types of extreme events was then calculated as before, but for the hindcast with the separate removal of ENSO (Figure 5b) and the climate change response (Figure 5c).

[38] The global average skill in forecasting extremes was reduced from 0.2 to 0.18 and 0.15, respectively. These reductions are significant at the 5% and 1% levels, respectively. The results vary widely spatially. For the removal of ENSO, the greatest reductions in skill occur when the teleconnections in the model world coincide with those in the

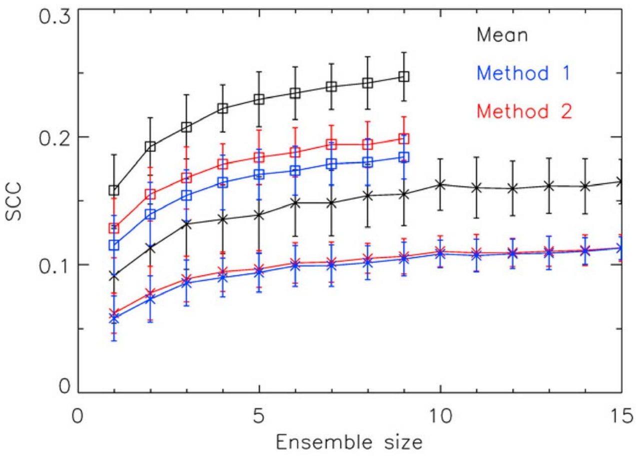

Figure 2. Global skill for different numbers of ensemble members. Black lines show how global average skill in predicting the mean varies. Red and blue lines show how the global skill in predicting the number of extreme events varies with ensemble size, using methods 1 and 2, respectively. The global skill is for GloSea4 (maximum 9 members, square symbols) and GloSea3 (maximum 15 members, crosses).

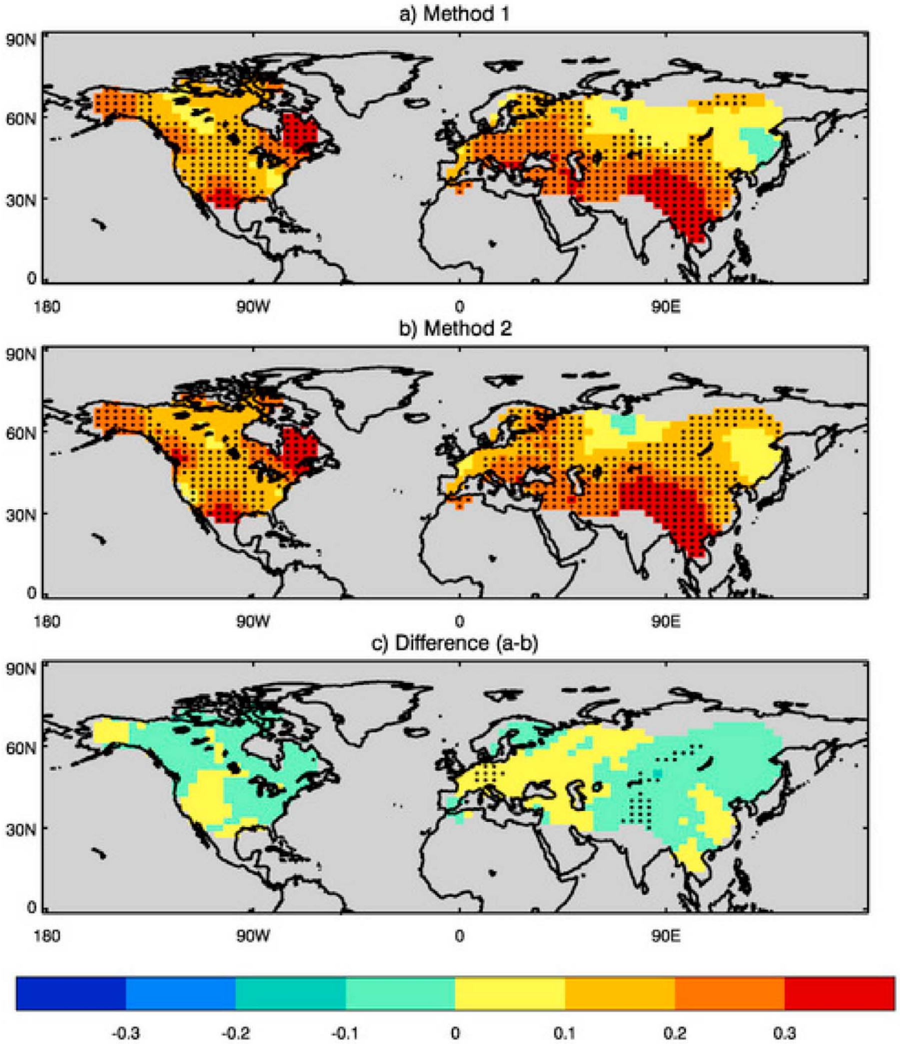

Figure 3. Figure 3a is identical to Figure 1a; it is repeated for ease of comparison. Figure 3b is as Figure 1b but for grid point skill in predicting extreme days using the mean only (method 2). Figure 3c is as Figure 1c except for the difference between Figures 3a and 3b. The percentages of significant grid points are 64%, 72%, and 6%, for Figures 3a, 3b, and 3c, respectively. The global average skills/differences are 0.20, 0.22, and −0.01, respectively. The global average skills are significant at the 1, 1, and 20% levels, respectively. Significance is calculated as Figure 1.

real world, and the magnitude of the signal is not too small in comparison to the interannual variability. Similarly with climate change, they occur where the model response to climate change is similar to the observed response and both responses take a relatively large range of values. For ENSO, this occurs in Alaska, the southern states of the United States and Southeast Asia (Figure 5b); these are known regions with strong ENSO teleconnections. For the removal of the climate change response, it is northeast Canada and southern Eurasia (Figure 5c). Note that these were areas where relatively high skill was seen in the original hindcast (Figure 1).

[39] Given the importance of ENSO and climate change signals of temperatures on the seasonal timescales, one might expect the reduction in skill from removing ENSO and climate change to be greater than that found previously. However, there are several factors, which have reduced the influence of climate change and ENSO on the hindcast skill. First, analysis is restricted spatially and does not include many of the areas known to be most influenced by ENSO (such as the tropics). Second, the 21 year hindcast period is relatively short; this means that the amplitude of the climate change signal is small in comparison to interannual variability (in observations it typically can explain around 10% of the variance). Finally, the indices and their teleconnections are imperfectly predicted, this reduces the amount of variance that can be explained by the responses in the hindcast compared with that seen in similar studies of observations alone.

[40] In addition to the ENSO and climate change factors, the seasonal predictability in the frequency of daily temperature extremes in our model arises partly from interannual fluctuations in soil moisture. Soil moisture is initialized from ERA-Interim using anomaly initialization at the start of each hindcast period [*Arribas et al.*, 2011] and then evolves through the forecast using the Joint U. K. Land Environment Simulator interactive land surface scheme [*Walters et al.*,

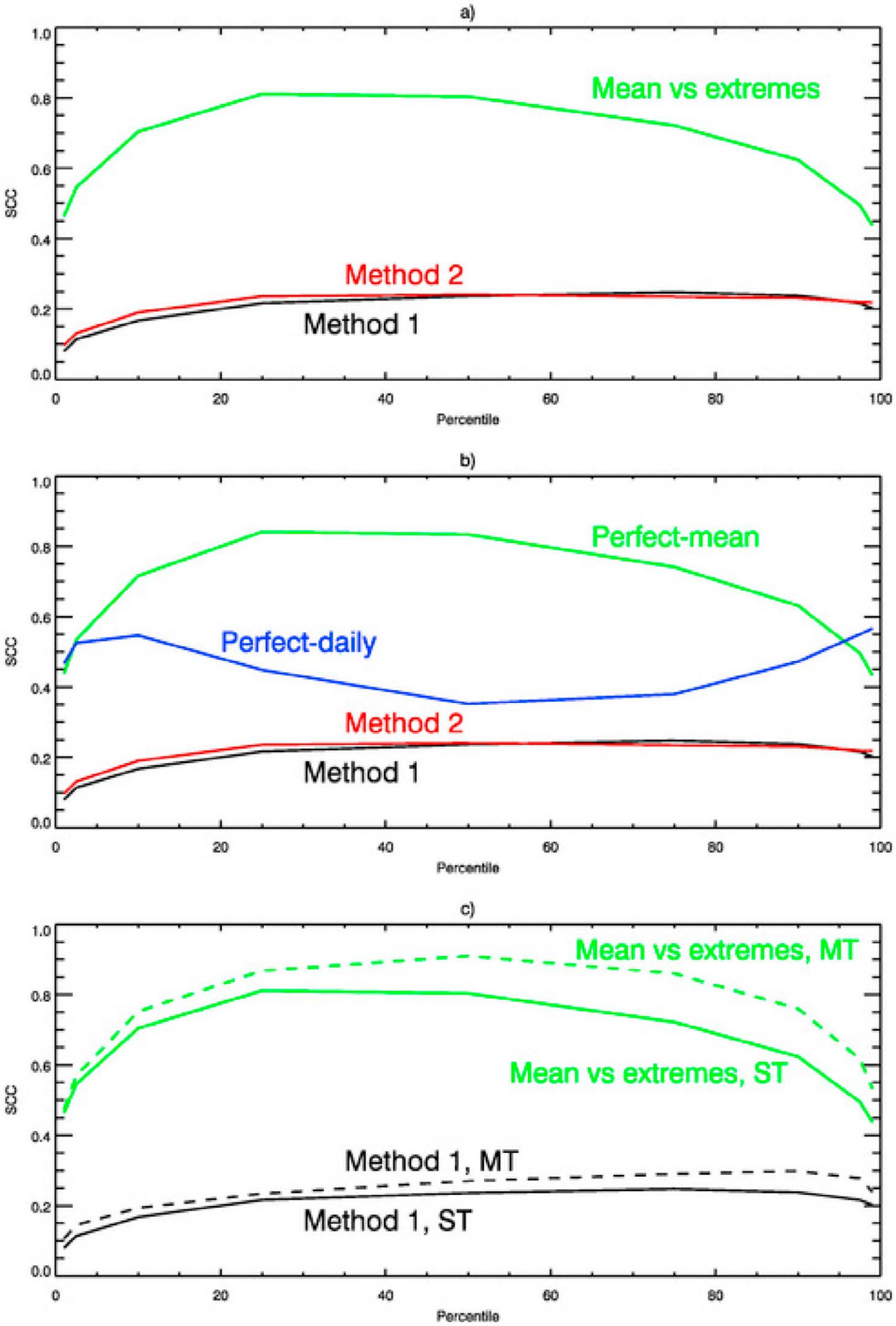

Figure 4. (a) Global skill of method 1 (black) and method 2 (red) against the threshold chosen (in percentiles of 1989–2009 climatology). Also plotted is the global average correlation between observed seasonal mean anomaly and number of static-threshold extremes (green), where the average is taken over all eight combinations of Tmin and Tmax with the four standard seasons. (b) Red and black lines as shown in Figure 4a. Global skill of perfect mean hindcast (green) and perfect daily data hindcast (blue). (c) Black and green lines as shown in Figure 4a. Also plotted is the grid point skill in predictions of moving-threshold extremes (dashed black) and the global average correlation between observed seasonal mean anomaly and number of moving-threshold extremes (dashed green). All the thresholds take the values of the 1st, 2.5th, 10th, 25th, 50th, 75th, 90th, 97.5th, and 99th percentiles.

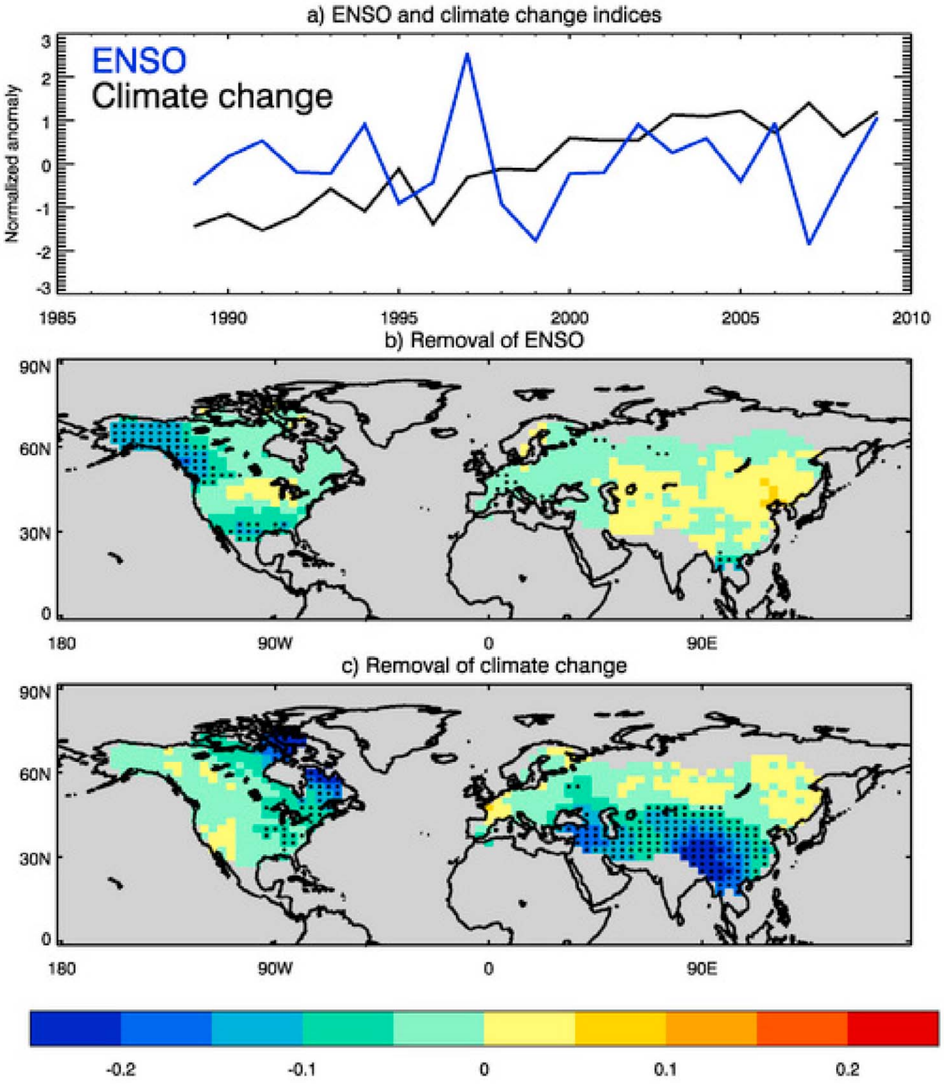

Figure 5. (a) The time series of ENSO and climate change indices over December–February (DJF) for a single ensemble member. Difference in skill in predicting extremes temperatures using original hindcast and hindcast (b) with ENSO response removed and (c) with climate change response removed. Average is over all 16 types of extreme temperature. Negative scores indicate that the removal of the signal has degraded the forecast. Points showing significance at the 5% level are stippled (one-tailed test).

2011]. Previous studies show that the impact of soil moisture anomalies on surface temperature is greatest in summer and in continental regions where large seasonal and interannual fluctuations in soil moisture occur [*Koster and Suarez*, 2003; *Seneviratne et al.*, 2006]. Figure 6a shows the correlation between regional surface temperature in Tmax over summer (JJA) and the corresponding soil moisture initialization (May). Here, soil moisture is taken to be the total moisture per square meter in the top four levels of the GloSea4 soil moisture initialization (a depth of 2 m). Areas with small interannual variation in soil moisture have been masked out, where small is taken to be an interannual range of less than 100 kg/m^2 in the soil moisture. This masking creates a region for analysis similar to the affected regions found in previous studies. As expected, the correlations are negative over large regions of the continents. Furthermore, similar regional patterns emerge to those found in other studies [e.g., *Koster and Suarez*, 2003]. This supports the known mechanism: for positive soil moisture anomalies, a larger-than-normal fraction of the solar irradiance goes into driving evaporation of the soil moisture, and so a smaller fraction heats the surface. Conversely, for negative soil moisture anomalies, a greater fraction of solar irradiance goes toward surface heating, and a smaller fraction drives evaporation. In addition, positive soil moisture anomalies mean that cloud and precipitation is more likely to occur than when the soil is dry. This not only has a cooling effect, but also acts as a positive feedback sustaining the positive soil moisture anomaly.

[41] Figure 6b shows the difference in skill between the original GloSea4 hindcast and an adjusted version of the GloSea4 hindcast where the linear response to soil moisture has been removed. Here, a method similar to that for the removal of ENSO and climate change was used, but now the

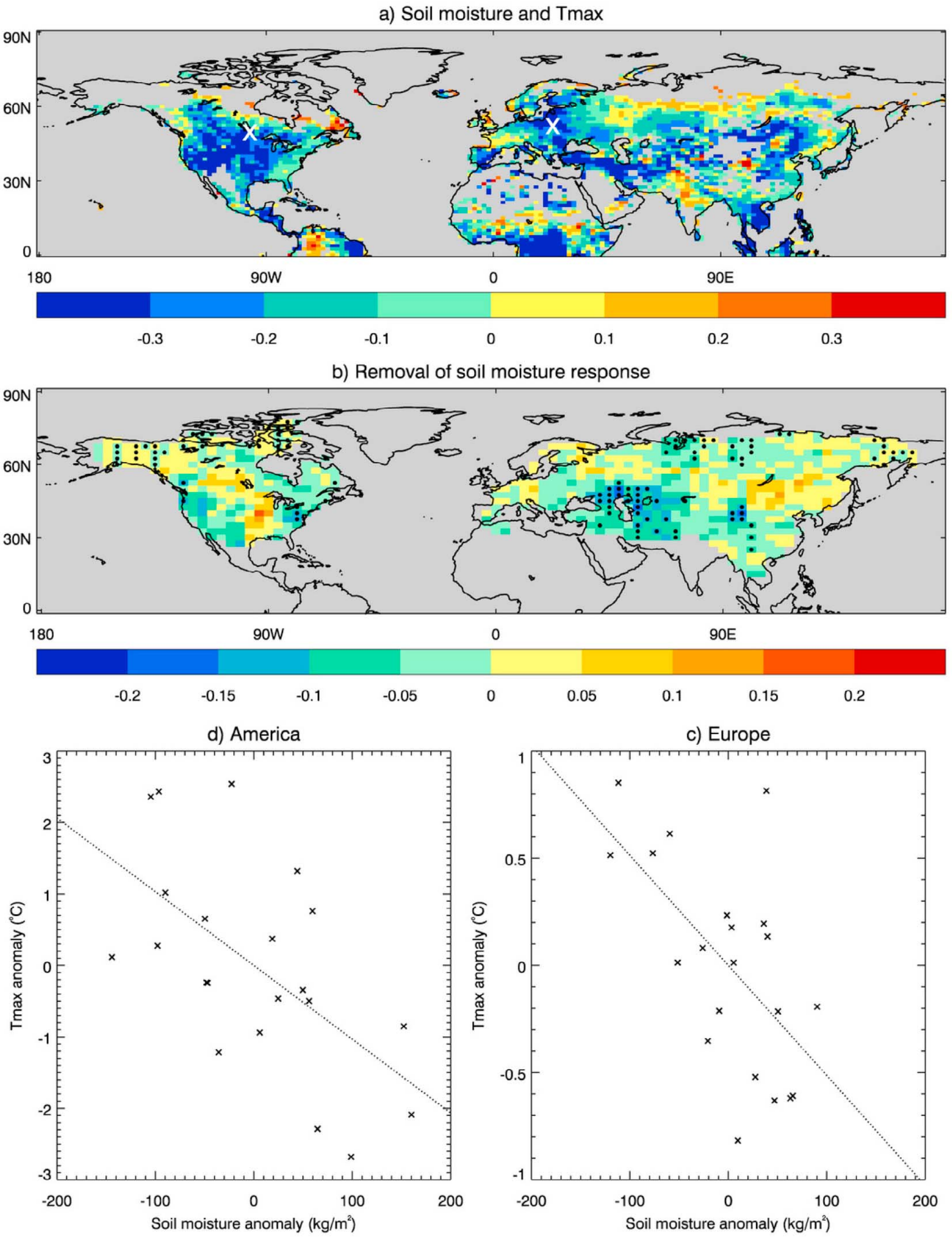

Figure 6. (a) Average Pearson's correlation coefficient between soil moisture initialized in May and hindcast seasonal mean Tmax during June–August (JJA). Areas where the interannual range of soil moisture is less than 100 kg/m^2 are masked in gray. White crosses indicate the location of the grid points used for Figures 6c and 6 d. (b) Average difference in skill in predicting extreme temperatures using original hindcast and hindcast with soil moisture response removed. Average is over both the 10th and 90th percentiles extremes of Tmax over JJA. Significance is as shown in Figures 6b and 6c. Grid points with missing data are masked in gray. (c and d) Variation of initialized soil moisture with ensemble mean seasonal mean Tmax over JJA for each of the 21 years in the hindcast.

index (soil moisture) is defined at each grid point rather than globally. Negative (blue) regions of Figure 6b show where the removal of the soil moisture response has decreased the skill. Areas of negative correlation between soil moisture and hindcast Tmax (Figure 6a, blue) mostly correspond to areas where there is reduction in skill (Figure 6b, blue). A notable exception is over the central Unites States. The reason for this disparity is that, over this region, the May soil

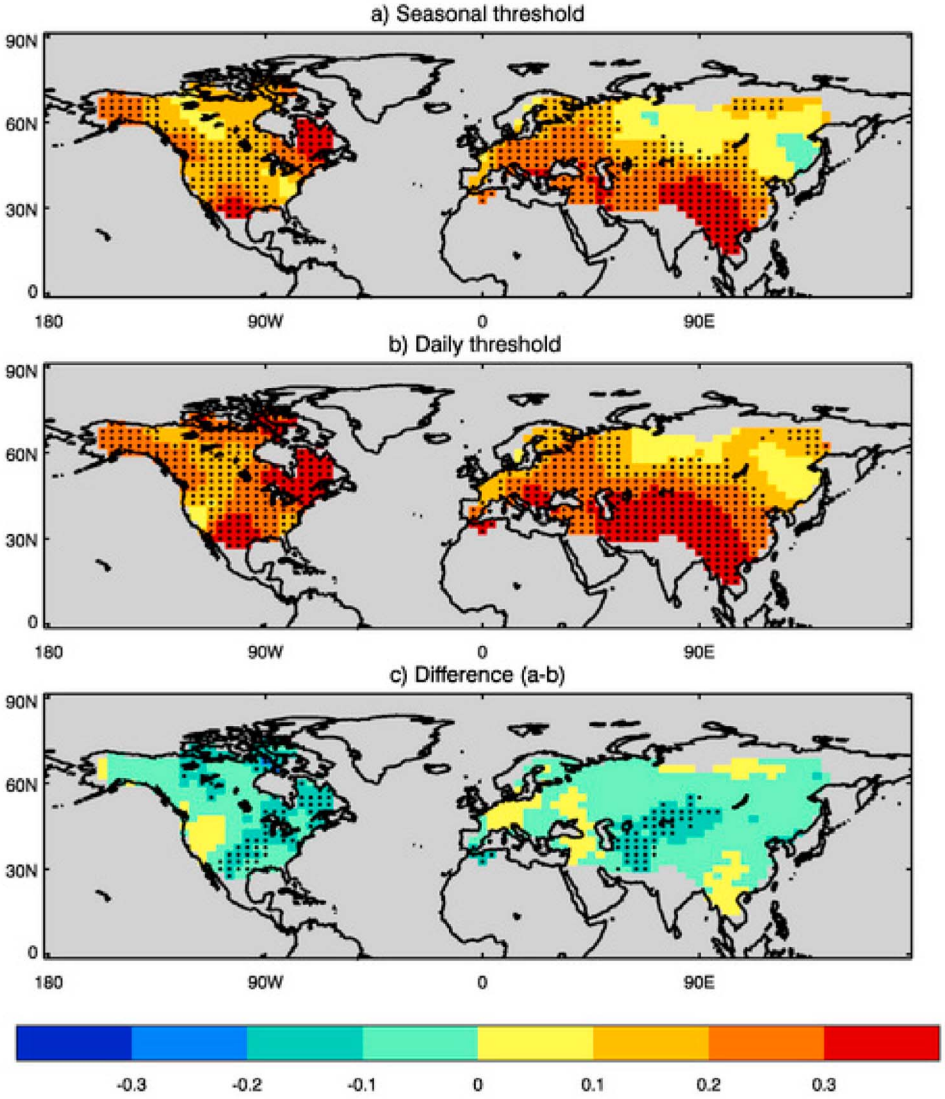

Figure 7. Same as Figure 1 except for (a) grid point skill in predicting extreme days using method 1 and the static-threshold definition and (b) grid point skill in predicting extreme days using method 1 but considering moving-threshold extremes. The percentages of significant grid points are 64%, 77%, and 14% for Figures 7a, 7b, and 7c, respectively. The global skills are 0.20, 0.25, and −0.05, respectively. The global average skills/differences are all significant at the 1% level. Significance is calculated as Figure 1.

moisture used for initialization is not negatively correlated with the observed seasonal mean Tmax over JJA (not shown). The difference in correlation may be because the model does not have the correct response to soil moisture anomalies in this region (i.e., there should not be negative correlation in Figure 6a over the central United States), but it is also likely to be a result of well-known imperfections in soil moisture initialization (i.e., the observed seasonal mean Tmax over JJA do not have negative correlation with the soil moisture initialization as the initialization is wrong). Overall, the removal of the soil moisture response significantly (1% level) reduces the global average skill in JJA from 0.28 to 0.26. Figures 6c and 6d quantify the relationship between soil moisture anomalies and near-surface temperature for key affected regions of the United States and Europe. From Figures 6c and 6d it can be seen that a typical interannual fluctuation of the soil moisture of around 100–200 kg/m^2 in the hindcasts can be responsible for temperature variations on the order of 1 or 2 K. Interestingly, to an order of magnitude, these numbers make quantitative physical sense when we treat the energy required for evaporation of these quantities of soil water as a climate forcing or as a perturbation to the annual cycle of insolation. When considering soil moisture as a climate forcing we may write

$$\Delta T = \frac{MLs}{\tau} \sim 2\,\mathrm{K},$$

where M is a typical variation of soil moisture (100 kg/m^2), L is the latent heat of evaporation of water (2×10^6 J kg^{-1} K^{-1}), s is a transient climate sensitivity (~0.1) [e.g., *Held et al.*, 2010, Figure 1] and τ is the seasonal time scale (10^7 s). In the case where soil moisture is regarded as a perturbation to the annual cycle of insolation, we may write

$$\Delta T = \frac{AML}{I\tau} \sim 2\,\mathrm{K},$$

where I is the change in solar irradiance over the seasonal cycle (~100 W/m^2) and A is the amplitude of the seasonal cycle in midlatitude continental regions (~20 K).

[42] Although these two estimates are only accurate to an order of magnitude, they support the findings in the model of sensitivity on the order of degrees Kelvin to typical year-to-year fluctuations in the soil moisture at the start of summer. This also helps to explain the finding that seasonal skill in summer extremes is larger than that in other seasons when soil moisture has little effect.

4.4. Sensitivity to the Use of Static or Moving Thresholds

[43] It was seen earlier that due, in part, to the strong correlation between the observed seasonal mean and the observed number of threshold exceedances, predicting the seasonal mean accurately contributes more to the skill in the prediction of the number of extremes than predicting the shape of the daily temperature distribution accurately. This would imply that, assuming that seasonal mean skill remains constant, the stronger the observed relationship between the seasonal mean and the number of exceedances, the greater skill the hindcast has in predicting the number of threshold exceedances. Figure 4c shows average global average correlation between the observed seasonal mean temperature and the number of static-threshold exceedances (solid green) and moving-threshold exceedances (dashed green). The relationship is stronger at all percentile thresholds for the moving-threshold exceedances than for the static-threshold exceedances. Figure 4c also shows the skill in predicting the number of static-threshold and moving-threshold exceedances. The stronger relationship between the extremes and the mean for the moving-threshold exceedances may have contributed to greater skill at all percentiles for the moving-threshold extremes. This difference is significant at the 1% level. However, Figures 7a and 7b show that the geographical distribution of skill of moving-threshold extremes is similar to that observed for static-threshold extremes. This demonstrates that our conclusions are not very sensitive to the use of static or moving thresholds, implying that predictions of extremes could potentially be made for a variety of user applications.

5. Summary and Concluding Remarks

[44] We have shown that, in general, it is possible to forecast the number of extreme daily temperatures in a season with skill that is significantly better than persistence. This is the first demonstration of skill at seasonal lead times in the frequency of daily temperature extremes. The skill is low, especially in the extra-tropics and is less than that for forecasting the seasonal mean temperature, but globally, the magnitudes of the mean correlation scores for forecasts of the seasonal mean and number of extremes is similar.

[45] Skill varies considerably between seasons. When broken down by season and type of extreme forecasts, GloSea4 predictions are significantly better than persistence in the Northern Hemisphere spring and summer but not in autumn or winter. The most skillful season was shown to be summer. Initialization of soil moisture was found to be a contributing factor in this result. Other sources of skill were the correct prediction of ENSO and the climate change signal. However, even when the signals from all these phenomena were removed, the resulting hindcast still exhibited significant skill so other factors are also at work.

[46] We have shown that a simple method in which the number of extreme days is implied from the hindcast seasonal mean temperature has almost identical skill to directly analyzing the daily hindcast data. This suggests that on seasonal timescales, skill in predicting the number of daily extreme temperature events arises largely from the skill in predicting the seasonal mean temperature anomaly, and little or no additional information is achieved by considering daily forecast data. This hypothesis is further supported by the fact that areas with greater skill in the prediction of the mean coincide with areas of greater skill in the prediction of the number of extreme days. Similar results are found in other climate modeling contexts, for example, shifts in large scale climate modes and, hence, regionally averaged temperature also explain shifts in regional extremes [*Kenyon and Hegerl*, 2008; *Scaife et al.*, 2008].

[47] In crop modeling, where the timing and persistence of extreme days is important, *Cantelaube and Terres* [2005] found that a weather generator could be applied to the seasonal mean of various variables to create daily data, which was useful in the seasonal prediction of crop yields over Europe. Further work might test whether the daily hindcast data may allow skillful prediction of the timing of extreme days during the season or the presence of consecutive extreme days that would produce better predictions of crop yield.

[48] The results in this article do not include the forecasting of precipitation extremes on seasonal timescales, as over regions with adequate observational data little evidence of skill was found. Seasonal skill for precipitation extremes is investigated by Eade et al. (manuscript in preparation, 2012).

[49] We found that, in general, prediction skill decreases as the temperature threshold considered becomes more extreme. However, the level of skill is remarkably constant between 25th and 99th percentiles and only drops slightly for predicting the 10th and 1st percentiles. Similar asymmetry in the prediction skill of cool temperatures compared with warm temperatures has been noted in earlier studies. The drop in skill at outer percentiles is because of the sampling noise caused by the sparseness of data in the tails.

[50] Idealized analysis showed that there is greater potential to improve skill in predictions of the number of extreme days by improving predictions of the mean than by improving the distribution of daily data about the seasonal mean (except for very extreme percentiles).

[51] Finally, we considered a definition where the threshold that defines “extreme” varies on a daily basis (rather than being static over each season). We find that this increases the correlation between the extremes and the mean in observations and therefore results in greater skill in predicting these extremes. The final choice of methodology between static- and moving-threshold extremes would, of course, depend on the application.

[52] **Acknowledgment.** This study was supported by the Joint UK DECC/Defra Met Office Hadley Centre Climate Programme (GA01101).

References

Arribas, A., et al. (2011), The GloSea4 ensemble prediction system for seasonal forecasting, *Mon. Weather Rev.*, *139*(6), 1891–1910, doi:10.1175/2010MWR3615.1.

Bowler, N. E., A. Arribas, K. R. Mylne, K. B. Robertson, and S. E. Beare (2008), The MOGREPS short-range ensemble prediction system, *Q. J. R. Meteorol. Soc.*, *134*(632), 703–722, doi:10.1002/qj.234.

British Standards Institute (2001), Power transformers. Part 3: Insulation levels, dielectric tests and external clearances in air, *BS EN 60076–3:2001*, London.

Caesar, J., L. Alexander, and R. Vose (2006), Large-scale changes in observed daily maximum and minimum temperatures: Creation and analysis of a new gridded data set, *J. Geophys. Res.*, *111*, D05101, doi:10.1029/2005JD006280.

Cantelaube, P., and J. M. Terres (2005), Seasonal weather forecasts for crop yield modelling in Europe, *Tellus. Ser. A*, *57*(3), 476–487, doi:10.1111/j.1600-0870.2005.00125.x.

Curriero, F. C., K. S. Heiner, J. M. Samet, S. L. Zeger, L. Strug, and J. A. Patz (2002), Temperature and mortality in 11 cities of the eastern United States, *Am. J. Epidemiol.*, *155*(1), 80–87, doi:10.1093/aje/155.1.80.

Dee, D., P. Berrisford, P. Poli, and M. Fuentes (2009), ERA-Interim for climate monitoring, *ECMWF Newsl.*, *119*, 5–6.

Doblas-Reyes, F. J., R. Hagedorn, T. N. Palmer, and J. J. Morcrette (2006), Impact of increasing greenhouse gas concentrations in seasonal ensemble forecasts, *Geophys. Res. Lett.*, *33*, L07708, doi:10.1029/2005GL025061.

Gershunov, A., and T. P. Barnett (1998), ENSO influence on intraseasonal extreme rainfall and temperature frequencies in the contiguous United States: Observations and model results, *J. Clim.*, *11*(11), 3062–3065.

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, *Clim. Dyn.*, *16*(2–3), 147–168, doi:10.1007/s003820050010.

Gosling, S. N., G. R. McGregor, and A. Paldy (2007), Climate change and heat-related mortality in six cities, Part 1: Model construction and validation, *Int. J. Biometeorol.*, *51*(6), 525–540, doi:10.1007/s00484-007-0092-9.

Held, I. M., M. Winton, K. Takahashi, T. Delworth, F. R. Zeng, and G. K. Vallis (2010), Probing the fast and slow components of global warming by returning abruptly to preindustrial forcing, *J. Clim.*, *23*(9), 2418–2427, doi:10.1175/2009JCLI3466.1.

Intergovernmental Panel on Climate Change (2001), *Climate Change 2001: The Scientific Basis: Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, edited by J. T. Houghton et al., 881 pp., Cambridge Univ. Press, New York.

Kenyon, J., and G. C. Hegerl (2008), Influence of modes of climate variability on global temperature extremes, *J. Clim.*, *21*(15), 3872–3889, doi:10.1175/2008JCLI2125.1.

Kiktev, D., D. M. H. Sexton, L. Alexander, and C. K. Folland (2003), Comparison of modeled and observed trends in indices of daily climate extremes, *J. Clim.*, *16*(22), 3560–3571, doi:10.1175/1520-0442(2003)016<3560:COMAOT>2.0.CO;2.

Kiktev, D., J. Caesar, L. V. Alexander, H. Shiogama, and M. Collier (2007), Comparison of observed and multimodeled trends in annual extremes of temperature and precipitation, *Geophys. Res. Lett.*, *34*, L10702, doi:10.1029/2007GL029539.

Kiladis, G. N., and H. F. Diaz (1989), Global climatic anomalies associated with extremes in the Southern Oscillation, *J. Clim.*, *2*(9), 1069–1090, doi:10.1175/1520-0442(1989)002<1069:GCAAWE>2.0.CO;2.

Koster, R. D., and M. J. Suarez (2003), Impact of land surface initialization on seasonal precipitation and temperature prediction, *J. Hydrometeorol.*, *4*(2), 408–423, doi:10.1175/1525-7541(2003)4<408:IOLSIO>2.0.CO;2.

Liniger, M. A., H. Mathis, C. Appenzeller, and F. J. Doblas-Reyes (2007), Realistic greenhouse gas forcing and seasonal forecasts, *Geophys. Res. Lett.*, *34*, L04705, doi:10.1029/2006GL028335.

Martin, A. J., A. Hines, and M. J. Bell (2007), Data assimilation in the FOAM operational short-range ocean forecasting system: A description of the scheme and its impact, *Q. J. R. Meteorol. Soc.*, *133*(625), 981–995, doi:10.1002/qj.74.

Philander, S. G. (1990), *El Niño, La Niña, and the Southern Oscillation*, 1st ed., 293 pp., Academic, San Diego, Calif.

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, *Clim. Dyn.*, *16*(2–3), 123–146, doi:10.1007/s003820050009.

Robertson, A. W., V. Moron, and Y. Swarinoto (2009), Seasonal predictability of daily rainfall statistics over Indramayu district, Indonesia, *Int. J. Climatol.*, *29*(10), 1449–1462, doi:10.1002/joc.1816.

Rogers, J. C., and R. V. Rohli (1991), Florida citrus freezes and polar anticyclones in the Great-Plains, *J. Clim.*, *4*(11), 1103–1113, doi:10.1175/1520-0442(1991)004<1103:FCFAPA>2.0.CO;2.

Scaife, A. A., C. K. Folland, L. V. Alexander, A. Moberg, and J. R. Knight (2008), European climate extremes and the North Atlantic Oscillation, *J. Clim.*, *21*(1), 72–83, doi:10.1175/2007JCLI1631.1.

Seneviratne, S. I., D. Luthi, M. Litschi, and C. Schar (2006), Land-atmosphere coupling and climate change in Europe, *Nature*, *443*(7108), 205–209, doi:10.1038/nature05095.

Smith, D. M., S. Cusack, A. W. Colman, C. K. Folland, G. R. Harris, and J. M. Murphy (2007), Improved surface temperature prediction for the coming decade from a global climate model, *Science*, *317*(5839), 796–799, doi:10.1126/science.1139540.

Taubenböck, H., J. Post, A. Roth, K. Zosseder, G. Strunz, and S. Dech (2008), A conceptual vulnerability and risk framework as outline to identify capabilities of remote sensing, *Nat. Hazards Earth Syst. Sci.*, *8*(3), 409–420, doi:10.5194/nhess-8-409-2008.

Taylor, N. A. S. (2006), Challenges to temperature regulation when working in hot environments, *Ind. Health*, *44*(3), 331–344, doi:10.2486/indhealth.44.331.

Walters, D. N., et al. (2011), The Met Office Unified Model Global Atmosphere 3.0/3.1 and JULES Global Land 3.0/3.1 configurations, *Geosci. Model Dev. Discuss.*, *4*(2), 1213–1271, doi:10.5194/gmdd-4-1213-2011.

Wilks, D. (1995), *Statistical Methods in the Atmospheric Sciences: An Introduction*, *Int. Geophys. Ser.*, vol. 59, 1st ed., Academic, San Diego, Calif.

Young, B. A. (1981), Cold stress as it affects animal production, *J. Anim. Sci.*, *52*(1), 154–163.

Zeng, Z., W. W. Hsieh, A. Shabbar, and W. R. Burrows (2010), Seasonal prediction of winter extreme precipitation over Canada by support vector regression, *Hydrol. Earth Syst. Sci. Discuss.*, *7*(3), 3521–3550, doi:10.5194/hessd-7-3521-2010.

R. Eade, R. J. Graham, E. Hamilton, C. MacLachlan, A. Maidens, A. A. Scaife, and D. M. Smith, Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK. (emily.hamilton@metoffice.gov.uk)