Weather and Climate Extremes 9 (2015) 68–77

Contents lists available at ScienceDirect

Weather and Climate Extremes

journal homepage: www.elsevier.com/locate/wace

The ability of a multi-model seasonal forecasting ensemble to forecast the frequency of warm, cold and wet extremes

Acacia S. Pepler [a,*], Leandro B. Díaz [b], Chloé Prodhomme [c], Francisco J. Doblas-Reyes [c,d,e], Arun Kumar [f]

[a] ARC Centre of Excellence for Climate System Science and Climate Change Research Centre, University of New South Wales, Sydney, Australia
[b] Centro de Investigaciones del Mar y la Atmósfera/CONICET-UBA, DCAO/FCEN, UMI IFAECI/CNRS, Buenos Aires, Argentina
[c] Institut Català de Ciències del Clima (IC3), Barcelona, Spain
[d] Institució Catalana de Recerca i Estudis Avançats (ICREA), Spain
[e] Barcelona Supercomputing Center-Centro Nacional de Supercomputación (BSC-CNS), Spain
[f] NOAA Climate Prediction Center, College Park, ML, USA

ARTICLE INFO

Article history:
Received 3 December 2014
Received in revised form
17 June 2015
Accepted 23 June 2015
Available online 24 June 2015

Keywords:
Extremes
Seasonal forecasting
ENSO
Climate model
Ensemble

ABSTRACT

Dynamical models are now widely used to provide forecasts of above or below average seasonal mean temperatures and precipitation, with growing interest in their ability to forecast climate extremes on a seasonal time scale. This study assesses the skill of the ENSEMBLES multi-model ensemble to forecast the 90th and 10th percentiles of both seasonal temperature and precipitation, using a number of metrics of 'extremeness'. Skill is generally similar or slightly lower to that for seasonal means, with skill strongly influenced by the El Niño-Southern Oscillation. As documented in previous studies, much of the skill in forecasting extremes can be related to skill in forecasting the seasonal mean value, with skill for extremes generally lower although still significant. Despite this, little relationship is found between the skill of forecasting the upper and lower tails of the distribution of daily values.

1. Introduction

Forecasts of seasonal mean temperature and precipitation are now an area of significant study and focus of ongoing improvement (e.g. Doblas-Reyes et al., 2013 and references therein). Statistical forecasts based on relationships between variability of seasonal mean climate and slowly-varying climate drivers such as the El Niño-Southern Oscillation (ENSO; e.g. Barnston, 1994; Drosdowsky and Chambers, 2001) have now been used for several decades. In addition, a growing number of global dynamical climate models are now used for seasonal forecasts in recent years (e.g. Alessandri et al., 2011; Barnston et al., 2003; Cottrill et al., 2013; Graham et al., 2011). Substantial skill has been achieved in forecasting seasonal mean values, with skill consistently higher in the tropics and in regions with strong teleconnections with ENSO and notably decreased skill during 'neutral' ENSO conditions (e.g. Phelps et al., 2004; Peng et al., 2011; Landman and Beraki, 2012).

While forecasts of above or below average seasonal mean conditions can be useful for industries such as agriculture (e.g. Wang et al., 2009), over recent years there has been growing interest in the ability of models to forecast extreme events such as prolonged heatwaves (Dole et al., 2013; Katsafados et al., 2014; Luo and Zhang, 2012). Several organisations now produce forecasts of the likelihood of seasonal temperature or rainfall to be in the highest 15% or 20% of the distribution of seasonal means, with forecasts generally more skillful than climatology, particularly for extreme temperature and at short lead times (Barnston and Mason, 2011; Becker et al., 2012; Marshall et al., 2013). However, skill is generally lower than the skill found for seasonal means, with the forecast model under-forecasting the frequency of extreme seasons (Barnston and Mason, 2011). This is another aspect of the complex characterisation of the model systematic error.

An alternative approach was taken by Hamilton et al. (2012) and Eade et al. (2012). Rather than forecasting the likelihood of an extreme season, these studies assessed the ability of the UK Met Office seasonal and decadal forecasting models to forecast the number of hot or cold *daily* extremes within a season, with extremes defined on the 90th or 10th percentile of the model and the observations, respectively. Focusing on the Northern Hemisphere, Hamilton et al. (2012) found that forecast skill was generally greater than that of a naïve climatological forecast, but lower

* Correspondence to: Climate Change Research Centre, University of New South Wales, Kensington 2052, Australia.
E-mail address: a.pepler@student.unsw.edu.au (A.S. Pepler).

http://dx.doi.org/10.1016/j.wace.2015.06.005

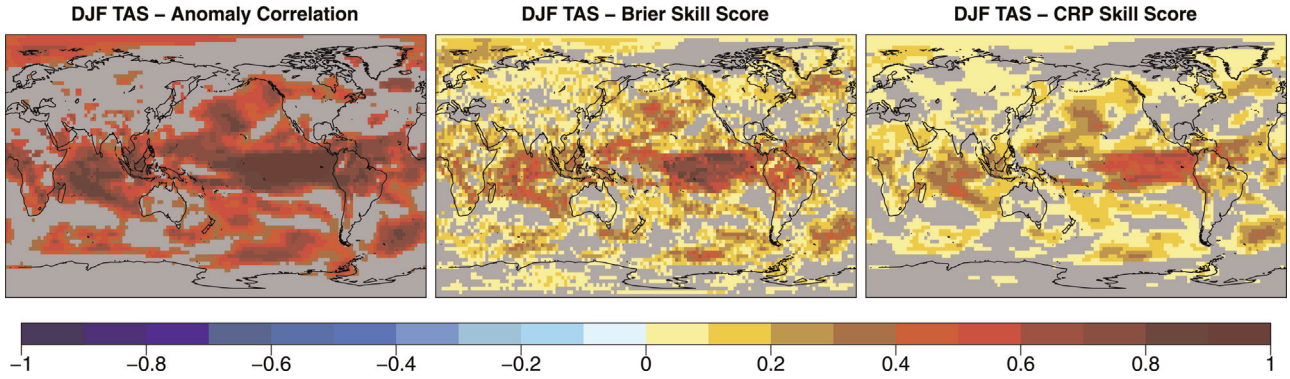

Fig. 1. Skill for the multi-model ensemble in forecasting seasonal mean temperature during DJF using three different measures of skill, with only skillful correlations shown. Skill measures are the anomaly correlation (left), Brier skill score (center) and continuous rank probability skill score (right).

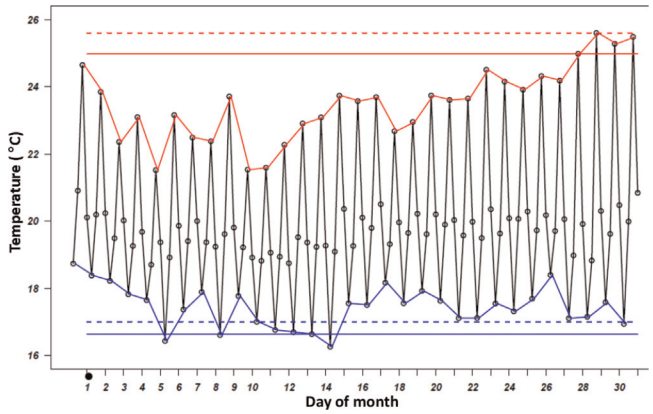

Fig. 2. An example of how the extremes are calculated for an arbitrary month of ERAI 6-hourly temperature data from a sample point. Red lines indicate the daily maximum temperature and blue lines the daily minimum temperature. The dashed horizontal lines indicate the climatological 90th and 10th percentiles of daily variability, while the solid horizontal lines indicate the 90th and 10th percentiles calculated for the month of interest. This month has 7 'cold extreme' days and one 'warm extreme' day, while the monthly 90th and 10th percentiles are both lower than the long-term mean. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

than that for seasonal mean temperature, concluding that the use of daily data from the forecasts to estimate high percentiles instead of a simple seasonal average adds no additional information. The largest contribution to the skill in predicting both mean and extreme temperatures arose from a combination of long-term climate trends and variability associated with ENSO. However, in their analysis, the use of the HadGHCND observational extremes data (Caesar et al., 2006) prevented them from carrying out a full analysis of areas where skill in extreme temperatures may be higher than that for means, such as in a small area northeast of India and in the tropics where ENSO teleconnections are strong.

In spite of the increasing interest in estimating the ability of forecast systems to forecast extreme events, recent work has only explored this ability for individual seasonal forecast models. As was shown in several previous works (e.g. Hagedorn et al., 2005; Weigel et al., 2008; Batté and Déqué, 2011), multi-model ensembles can, on average, outperform the best single forecast system. As a result, higher skill and possibly different conclusions to those obtained by Hamilton et al. (2012) and Eade et al. (2012) could be expected when exploring forecasts of extremes based on an ensemble of models.

In this paper, we use the ENSEMBLES multi-model ensemble (Weisheimer et al., 2009) to assess the skill of seasonal outlooks of both extreme *daily* temperature and precipitation across the globe. In section 2 we will discuss the ENSEMBLES and verification datasets, with Section 3 discussing a number of methods of assessing seasonal forecasts of extremes. This is followed by a discussion of the spatial variation of skill in both the seasonal mean value and the frequency of daily extremes both for temperature (Section 4) and precipitation (Section 5), including the extent to which skill in extremes is similar to the skill in forecasting the seasonal means alone. We also identify the influence of ENSO on skill in both seasonal means and daily extremes. Finally, we identify areas and variables for which skill in extremes may be higher than that for seasonal means, and discuss possible causes.

2. Data and methodology

The ENSEMBLES multi-model data (Weisheimer et al., 2009) comprises the dynamical seasonal forecasts from the global coupled models of five major international modelling centers: the UK Met Office (UKMO), Météo-France (MF), the European Centre for Medium-Range Weather Forecasts (ECMWF), the Leibniz Institute of Marine Sciences at Kiel University (IFM-GEOMAR) and the Euro-Mediterranean Centre for Climate Change (CMCC-INGV) in Bologna. A 9-member ensemble of daily forecasts for each model is available for the seven months following the forecasts initialised on the 1st of February, May, August and November for each year between 1960 and 2005. While the models used are several years

Table 1
Proportion of grid points over the globe with statistically significantly positive anomaly correlation for four measures of seasonal extremes and three different climate extremes during JJA. The significance level is estimated for each grid point with a Fisher transformation and taking into account the actual number of independent data.

	Hot days (maximum temperature Q90) (%)	Cold nights (minimum temperature Q10)	Wet days (daily precipitation Q90)
Seasonal mean	64	64	28
Percentile value calculated from seasonal daily data	41	36	17
Percentile value calculated from daily data for each month (seasonally averaged)	57	58	25
Seasonal count of extreme days	39	35	18
Seasonal mean of monthly count of extreme days	51	52	27

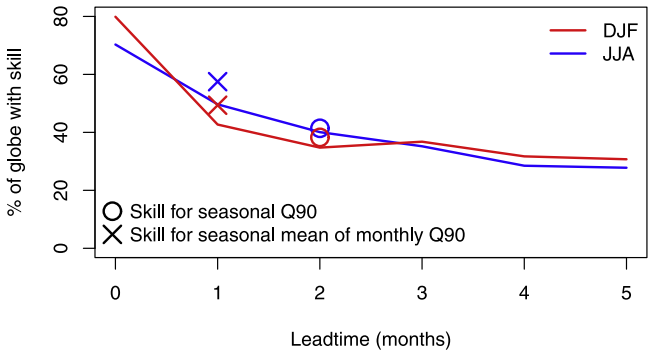

Fig. 3. Proportion of the globe with statistically significantly positive anomaly correlation for the 90th percentile of daily maximum temperature for both DJF (initialised in November) and JJA (initialised in May) as a function of forecast month, noting that the seasonal forecast is for months 1–3. For comparison, the proportion of the globe with skill at forecasting the seasonal 90th percentile is shown in comparison to skill for month 2, while the skill at forecasting the seasonal mean of the monthly 90th percentiles is shown in comparison to skill for month 1, indicating that the skill for the seasonal mean of the monthly percentiles is higher than for any individual month.

older than the most recent suite of seasonal forecast models, the use of an ensemble of climate models with a very long hindcast period (46 years) gives a good indication of areas of widespread or consistent skill, and also allows us to address questions posed in Section 1.

Forecasts from the ENSEMBLES dataset are validated using the ERA-Interim (ERAI) six-hourly surface data (Dee et al., 2011), with daily maximum and minimum temperature derived from values for the four time steps comprising each UTC day. While ERAI is available at a resolution of 0.75°, for consistency of analysis both the observations and all model outputs were interpolated to a consistent 2.5° resolution. Although reanalyses are known to underestimate the values of climate extremes, the variability in extreme values is consistent with observational datasets in areas where in situ observations are available (Donat et al., 2014), while the global spatial coverage of reanalysis allows us to assess skill in areas where in-situ observational data is limited. However, there is significant variation in the observed trends in extremes between reanalyses in areas where data is poorly constrained by observations, so caution is warranted.

The ERAI reanalysis dataset is available from 1979 to present; consequently, the analyses in this paper are restricted to the shorter period 1979–2005. We focus on three-month periods, namely June-to-August (JJA) and December-to-February (DJF), with outlooks assessed for a 1-month lead time. While extremes by our definition occur with equal frequency at all times of the years, these two seasons were chosen as the periods of the year when the values of warm and cold extremes reach their largest magnitudes, and therefore, have largest impacts on economy, agriculture, human health and ecosystems. We also note that DJF is also the season when ENSO teleconnections in the Northern Hemisphere have their largest amplitude and ENSO is also at its peak (e.g. Trenberth et al., 1998). The majority of analyses will focus on the multi-model mean only.

El Niño and La Niña years are defined as per the Climate Prediction Centre of the National Oceanic and Atmospheric Administration, available at http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.shtml.

2.1. Assessment of skill

There are a number of ways to assess the skill of seasonal forecasts, each of which gives slightly different information. To simplify the discussion, results in this paper will focus on a single method, the anomaly correlation, which is the correlation between the multi-model mean forecast anomaly and the observed anomaly. The model is deemed 'skillful' when the correlation between the multi-model mean anomaly and the observation is positive and statistically significant at the 5% level using a two-tailed test, consistent with similar studies (Becker et al., 2012; Hamilton et al., 2012). Prior to calculating skill, both the observations and models were linearly detrended using a least-squares regression to remove any signal from long-term trends; this had little impact on the results. Forecasts of mean and extreme precipitation are only shown for regions where the climatological mean precipitation is higher than 10 mm in all three months of the season – this has almost no effect on results, as anomaly correlations were statistically insignificant in almost all of these regions.

We use Student's t distribution with N degrees of freedom to estimate the significance level of correlations, N being the effective number of independent data calculated following the method of von Storch and Zwiers (2001), with significance assessed at the 5% level. The significance of the difference between two correlations is estimated using a Fisher z-transformation, following the Olkin and Finn (1995) methodology. These take into account the independent number of data, which is necessary given the serial correlation typical of the time series considered.

To test the usefulness of this measure of skill, we compared the

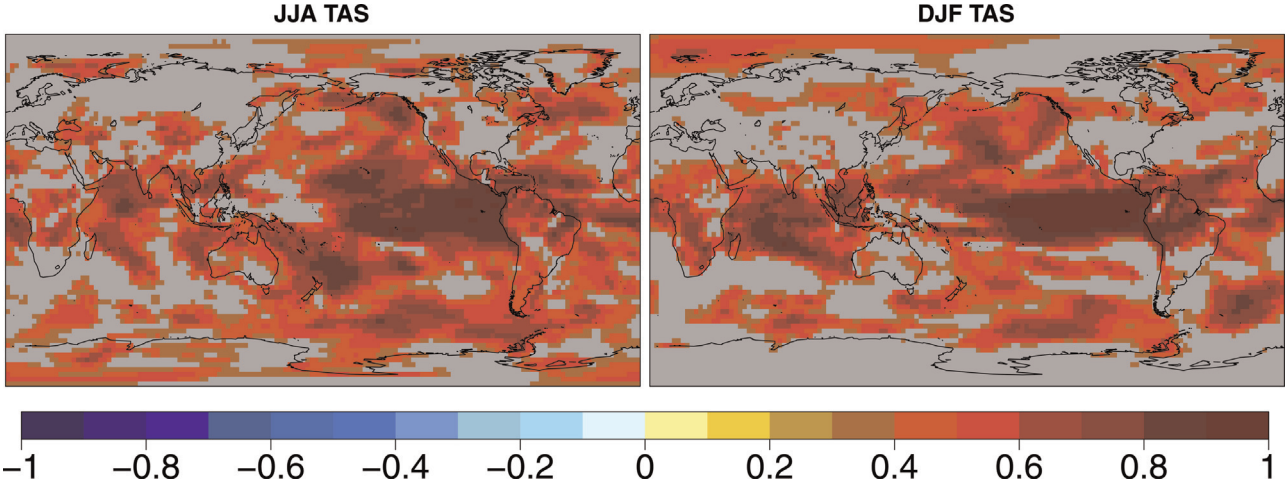

Fig. 4. Anomaly correlations between the ENSEMBLES ensemble mean temperature forecast and ERA-Interim reanalysis for JJA (left) and DJF (right), 1979–2005. Correlations are only shown where they are significant at the $p=0.05$ level using a Fisher test that takes into account the effective number of independent data.

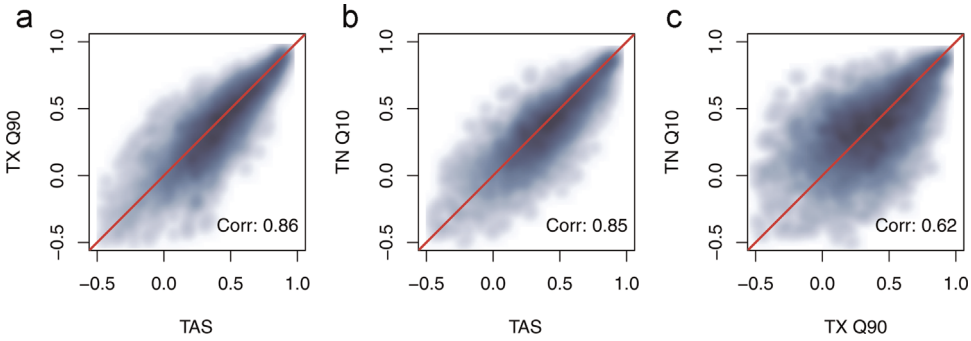

Fig. 5. Scatter plot of the skill in terms of anomaly correlation for the mean daily temperature and the 90th and 10th percentiles over the entire globe during JJA over 1979–2005, with the correlation between the two samples indicated in the upper left corner of each panel. The red line indicates a perfect relationship. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

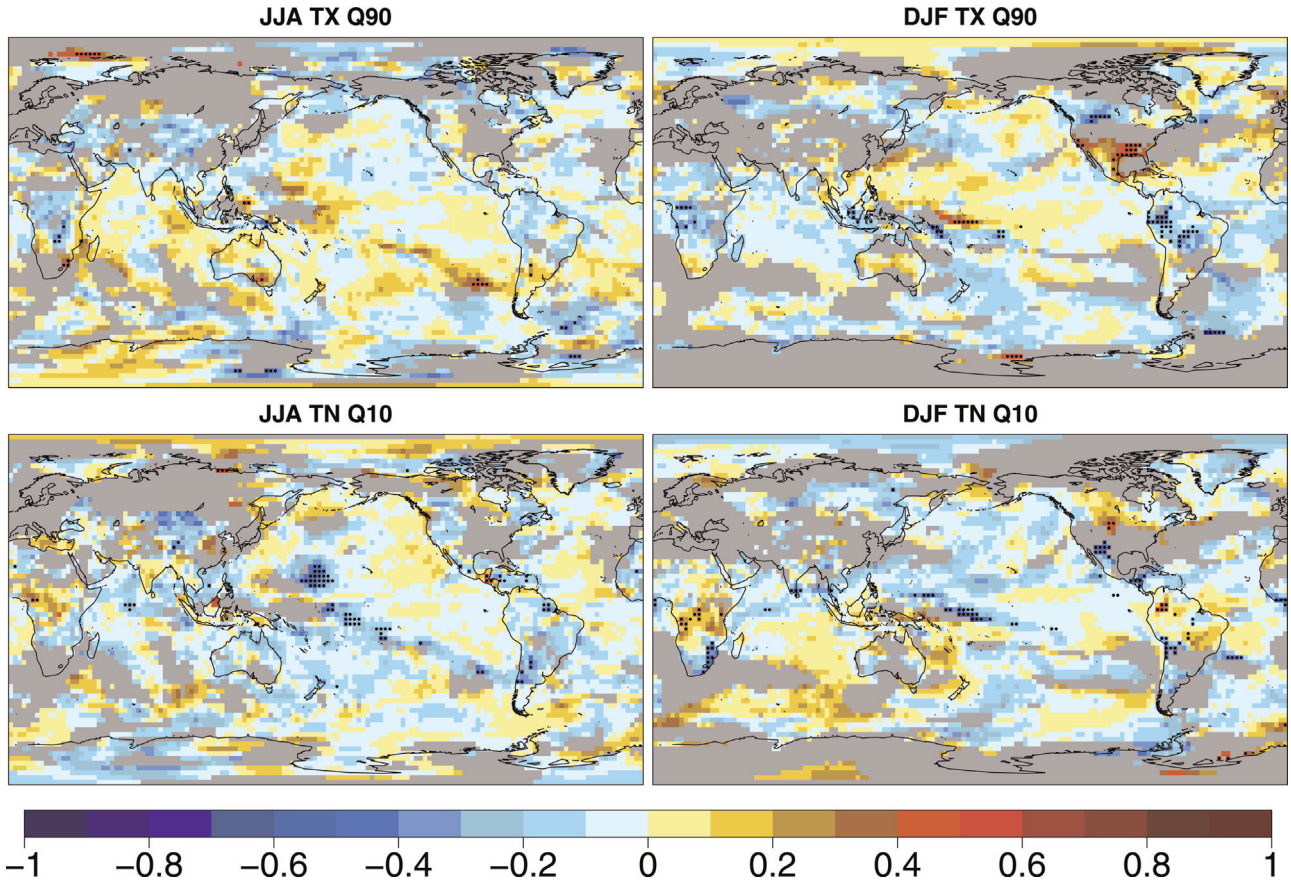

Fig. 6. The difference between the anomaly correlation between the ERAI observations and the ENSEMBLES multi-model forecast mean as obtained for indices of seasonal extremes, and the same correlation for the seasonal mean temperature. This is an indication of areas where the skill of forecasts of extreme temperature is higher or lower than forecasts of the seasonal average. Differences are only shown for regions where either the correlation for the seasonal mean temperature or the temperature extreme of interest are statistically significant, while black dots indicate areas where the difference between correlations is statistically significant. Correlations are calculated over the period 1979–2005 for the JJA (left) and DJF (right).

calculated anomaly correlations with two other measures of skill. The Brier skill score (Brier, 1950) is one way of assessing a probabilistic forecast, where the full multi-model ensemble is used to forecast the likelihood of a season exceeding a certain threshold – in this case, the likelihood of above-median seasonal values. The continuous ranked probability (CRP) skill score (Epstein, 1969) is a measure of the skill of a model ensemble forecast which employs the full forecast ensemble, compared to skill for forecasts of the observed climatology (e.g. Bröcker, 2012). Despite the different approaches to both forecast generation and verification across these methods, the patterns of skill are very similar (Fig. 1), with spatial correlations of 0.75–0.85. This is to be expected as various measures of skill are related and high (low) values for one skill measure generally correspond for the same for the other (Kumar, 2009). As the anomaly correlation is a more accessible method of measuring skill, we show results based on this with the expectation that the results shown will be broadly consistent with results for other methods.

3. Defining the seasonal extremes

For the objectives of this paper, extremes are defined by the 90th percentiles of daily precipitation and maximum temperature

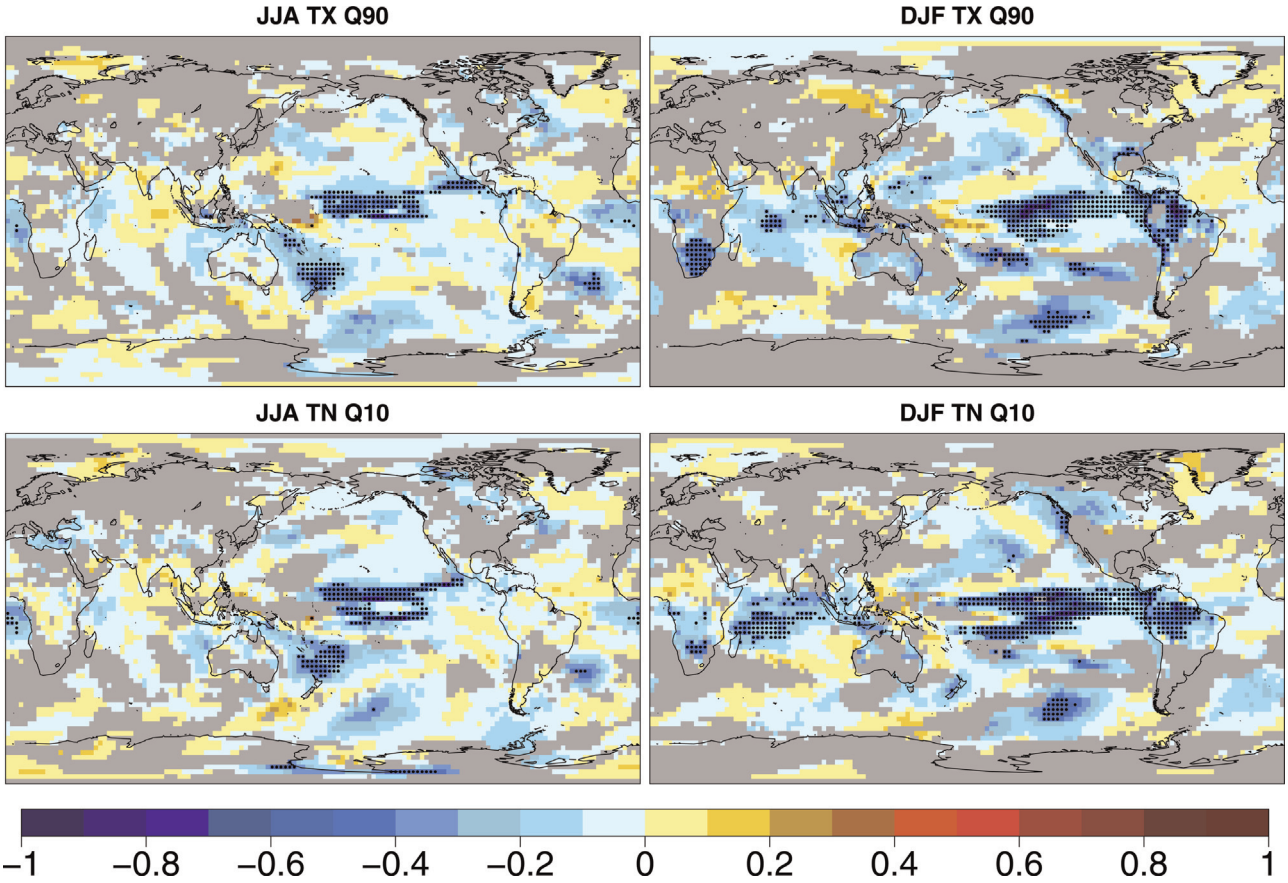

Fig. 7. Change in forecast skill for seasonal extreme temperatures after the linear regression with the seasonal NINO3.4 is removed. Differences are only shown for regions where either the initial or residual correlations are statistically significant, while black dots indicate areas where the difference between correlations is statistically significant. Correlations are calculated over the period 1979–2005 for JJA (left) and DJF (right).

(TX), and the 10th percentile of daily minimum temperature (TN), within a month or season. These are 'mild' extremes, such that they are expected to occur on 10% of days during a given season. For the small samples available in seasonal forecasting, they are more statistically robust and easily verifiable than other extreme and rare events, and are expected to have higher forecast skill than more astringent definitions for extremes (Becker et al., 2012; Hamilton et al., 2012).

We assess two different methods of calculating the daily extremes for a given period. For the first method, we first calculate the climatological 10th percentile (Q10) or 90th percentile (Q90) over all 27 years of daily data for both the observation and models during a given month or season. This climatological percentile is then used to calculate the number of days exceeding the Q90 threshold (or below the Q10 threshold) in each year.

In the second method, we calculate the value of Q90 (or Q10) individually for each year; that is, the magnitude of the third-largest (or third-smallest) daily value during that month in °C or mm. The anomaly in the quantile value relative to the long-term mean calculated from this data can be considered an indication of, for example, how much warmer the warmest days in that month were compared to an average month.

In both cases, the number of extreme days (or quantile value) can be calculated for the season as a whole (i.e. from daily data for all three months), or the values can be calculated separately for each month and then averaged across the season. While strictly-speaking averaging a quantile value across three months is poorly defined, particularly for rainfall, it can be considered the average anomaly of the warmest days across all three months. This can be a useful measure of whether the tail of the distribution is warmer or cooler than average. Fig. 2 demonstrates the calculation of these measures of extremeness for an arbitrary set of monthly temperature data.

In order to determine which methods provide the most skillful forecasts for further analysis, we assess the proportion of the globe where the anomaly correlations are statistically significant. These are shown for JJA forecasts of three different climate extremes in Table 1, with similar results for DJF (results not shown).

In all cases, forecasts of the seasonal average of the monthly forecasts have statistically significant correlations across substantially larger areas of the globe than when extremes were calculated across the season as a whole, consistent with results of Hamilton et al. (2012), although the proportion of the globe with skillful forecasts of the seasonal mean value is higher than the proportion obtained with any method of assessing extremes. The difference in skill between the two approaches has been compared with the variation in forecast skill across the season, where there are generally more skillful forecasts at shorter lead times (Fig. 3). While using the seasonal mean of the monthly forecast of the extremes allows the prediction method to detect the extremes along the season, a constant seasonal threshold means that the majority of extreme temperature values in both DJF and JJA would be expected to occur during the second month of the forecast, January and July respectively. However, the skill for the season as a whole remains higher than the skill for month two alone, with the skill at forecasting the seasonal mean of the monthly values being higher than that for any individual month.

The differences in skill between the two metrics of daily extremeness are smaller than that for the two approaches used for seasonal estimates. In the case of both warm (Q90) and cold (Q10)

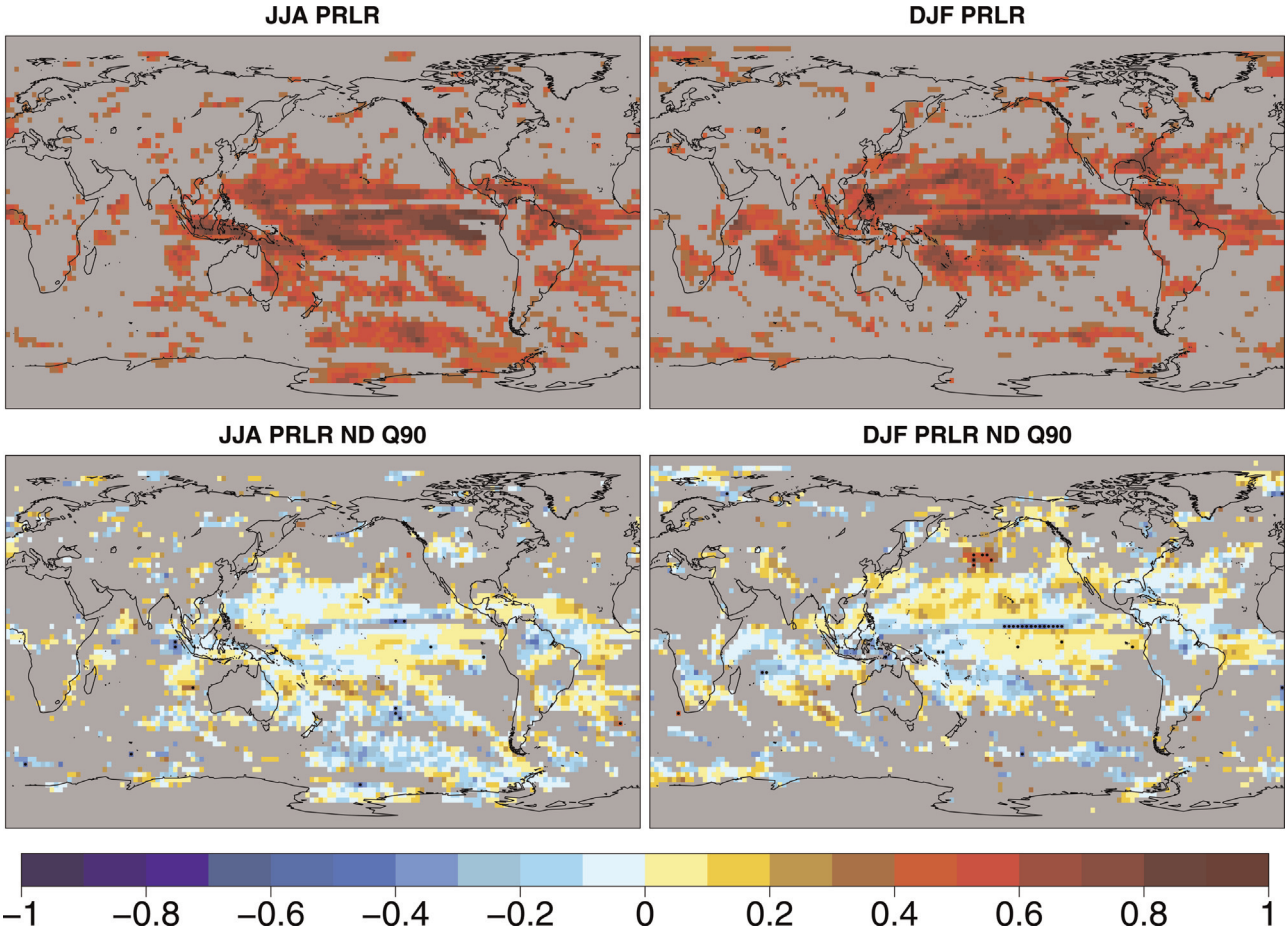

Fig. 8. (top) Anomaly correlation between the ENSEMBLES ensemble mean precipitation forecast and ERA-Interim reanalysis for JJA (left) and DJF (right) over 1979–2005. (bottom) Differences in anomaly correlations (relative to the correlation for the seasonal precipitation) for the number of days with precipitation above the climatological Q90. Differences are only shown for regions where either the correlation for mean precipitation or the number of extreme precipitation days is statistically significant and where monthly mean precipitation is higher than 10 mm, while black dots indicate areas where the difference between correlations is statistically significant.

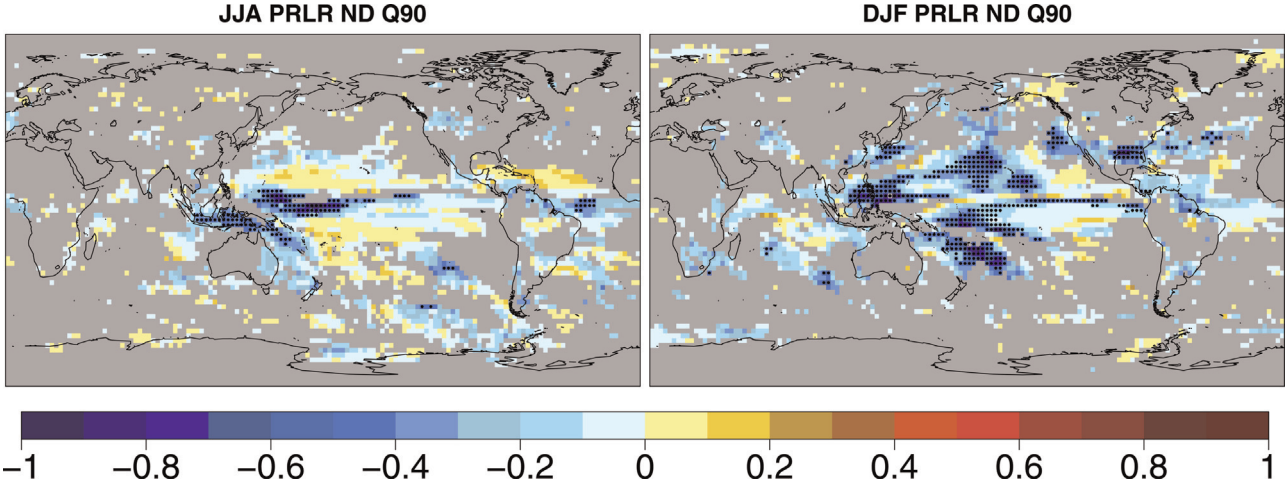

Fig. 9. As in Fig. 7, but for changes in anomaly correlations for the number of days with precipitation above the climatological Q90. Differences are only shown for regions where the correlation is statistically significant and where monthly mean precipitation is higher than 10 mm.

daily temperature extremes forecasts of changes in the value of the seasonal 10th or 90th percentile have skill across a larger proportion of the globe than forecasts of the number of warm/cold days relative to a static climatological threshold. In contrast, forecasts of the number of days with precipitation exceeding the climatological 90th percentile are skillful over larger proportions of the globe than forecasts of changes in the value of the 90th percentile. While the cause of these differences has not been assessed, it may be related to the very different distributions of daily data for these two variables.

While all extreme forecast methods have better skill than climatology, using the most skillful metrics allows us to best assess the sources of forecast skill, particularly in comparison to forecasts for seasonal mean values. For this reason, the remainder of the

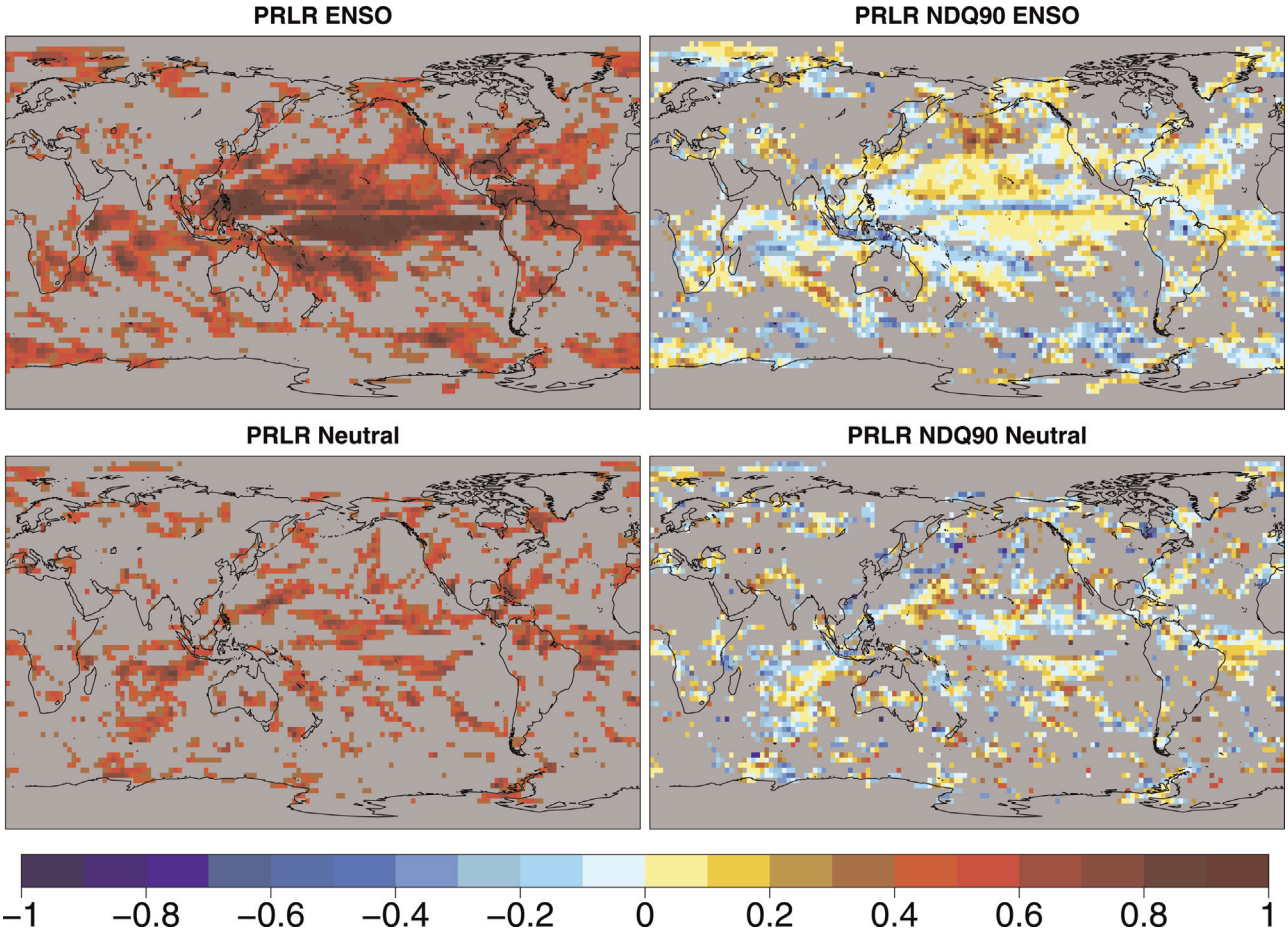

Fig. 10. (left) DJF Anomaly correlations between the ENSEMBLES ensemble mean precipitation forecast and ERA-Interim reanalysis, 1979–2005, with years separated into ENSO (top) and neutral (bottom) based on the NOAA CPC classifications. (right) Differences in anomaly correlations (relative to the correlation for the seasonal precipitation) for the number of days with precipitation above the climatological Q90. Differences are only shown for regions where either the correlation for mean precipitation or the number of extreme precipitation days is greater than 0.32, which is the statistical significance threshold for the full 27 year database; statistical significance thresholds for ENSO years (15 years) and neutral years (12 years) are 0.52 and 0.58 respectively.

paper will first assess the skill in forecasting the 10th percentile of minimum temperature (TN Q10) and the 90th percentile of maximum temperature (TX Q90) for each month, averaged across the season, in comparison to the skill for the monthly mean temperature (TAS). This is followed by an assessment of the skill in forecasting the number of days with precipitation exceeding the climatological 90th percentile (PRLR Q90ND) for each month of the season, in comparison to skill for monthly total precipitation (PRLR).

4. Skill for extreme temperatures

As noted in Alessandri et al. (2011), the ENSEMBLES multi-model mean has skill for mean temperatures, particularly in the tropics. More than half of the globe has statistically significant correlations between the multi-model forecasts and the ERAI observations (Fig. 4), with the skill for the multi-model mean in both seasons exceeding that for any individual model (results not shown). The majority of the skill is concentrated over the oceans, where more than 60% of areas have skillful forecasts during both seasons, increasing to over 80% of tropical oceans.

Fig. 4 shows that skill is higher in the Southern than in the Northern Hemisphere, with very few areas of the Northern Hemisphere midlatitudes having statistically significant correlations. During JJA, this is largely a result of lower skill over land areas, with skillful forecasts in 64% of land areas in the Southern Hemisphere midlatitudes (23.5–66°S), compared to just 36% of the Northern Hemisphere midlatitude land areas. In comparison, 48% of land areas in Southern Hemisphere mid-latitudes have skillful forecasts during DJF, with large areas of skill in the Indian Ocean. It is interesting to note that, if a constant correlation value was considered to detect significance was used across the globe, the proportion of the northern midlatitudes with skill would increase substantially in both seasons. This is related to higher levels of autocorrelation in temperatures in the Northern Hemisphere, so that the required significance levels taking into account the number of independent data is correspondingly higher in these regions.

The proportion of the globe with skillful forecasts for temperature extremes is high, particularly in the tropics, although slightly lower than for forecasts of the seasonal mean temperature. There is also a strong relationship between skill of seasonal averages and that in the extremes (Fig. 5). We can quantify this in terms of the spatial correlation between the skill for seasonal averages and for extreme temperatures at each grid point, which vary between 0.85 and 0.87. However, the relationship between skill in predicting the two tails is substantially lower, with a global spatial correlation of just 0.62 in JJA (Fig. 5c). This suggests that the skill in predicting one extreme is not well correlated with skill in predicting the opposite extreme. Indeed, while half of the globe has higher skill in predicting the skill in one extreme than in the seasonal averages, less than 10% of the globe has skill higher than that of the seasonal averages for predicting both tails.

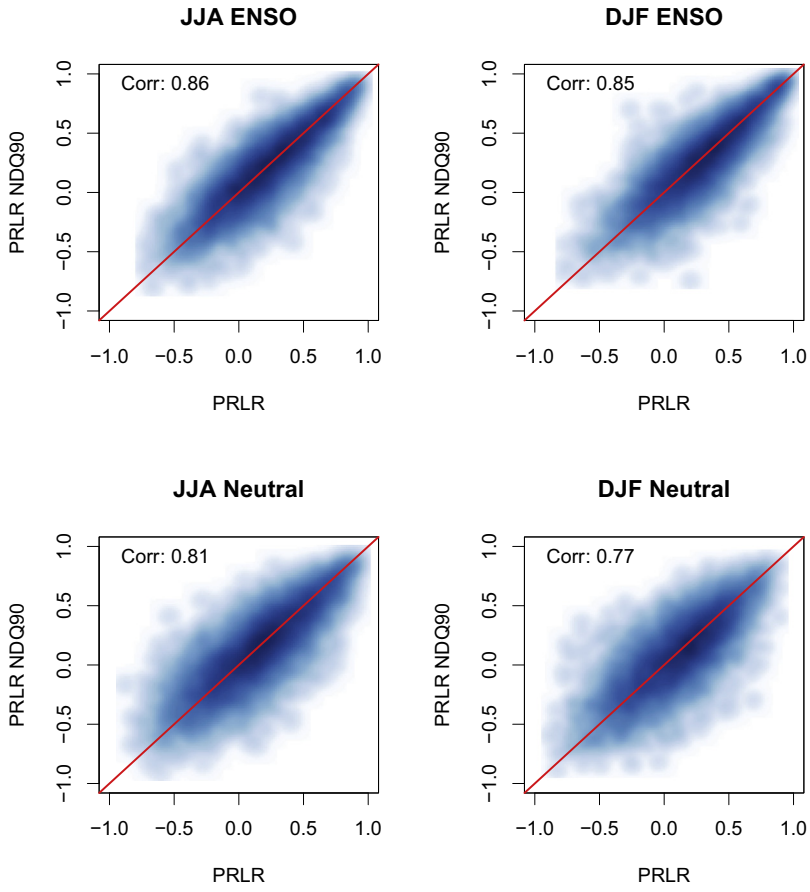

Fig. 11. Scatter plot of the correlation between the skill in forecasting mean precipitation and the skill in forecasting the number of days above the climatological Q90 for JJA (left) and DJF (right). Years are separated into ENSO (top) and neutral (bottom) based on the NOAA CPC classifications.

The difference in skill between predicting seasonal mean temperature and predicting the tails of the seasonal temperature distribution is generally small, and the short length of the record means that very few areas of the globe have statistically significant difference in skill between mean and extreme temperatures (Fig. 6). However, there are several regions where an increase in the skill for one tail appears to be associated with a decrease in skill in the other. This is notably the case for parts of southeast Australia, where increased skill for extreme high temperatures is found during both seasons, with generally lower skill for extreme cold nights. Parts of the southern US also have higher skill in forecasting warm extremes during the boreal winter (DJF) than forecasts of the seasonal mean, but lower skill in forecasting extreme cold nights. Tropical South America and parts of tropical Africa show increased (decreased) skill for extreme low (high) temperatures during DJF.

Though ENSO is a known major driver of skill in mean temperatures and precipitation (e.g. Manzanas et al., 2014), its role in the extremes skill remains to be assessed for the ENSEMBLES multi-model ensemble. We separate the influence of ENSO by first calculating the linear regression between the temperature at each grid point and that in the Niño 3.4 region (SST averaged between 5°S and 5°N, 120° and 170°W) for both the multi-model mean and the observations. We then estimate the skill of the residual values. With the ENSO signal removed, the magnitude of correlations decreases across much of the globe, particularly in the equatorial Pacific (Fig. 7). However, the proportion of the globe with skillful correlations for both TX Q90 and TN Q10 decreases by only 4–5% during JJA when compared to the initial forecast, with some remnant skill in most areas regardless of ENSO (Fig. 7). The contribution of ENSO to the skill is larger during DJF, when ENSO has strong teleconnections and is a large contributor to skill in forecasting both high maximum and low minimum temperatures across much of northern South America and southern Africa. Similar results are found when the skill is estimated separately for the years categorised as either phase of ENSO (both El Niño and La Niña) or neutral years by the Climate Prediction Centre of the National Oceanic and Atmospheric Administration[1] (not shown).

5. Skill for extreme precipitation

Skill for forecasting precipitation is substantially lower than that for temperatures, with forecasts with significant skill covering less than 30% of the globe. The skill is concentrated in the tropics and over the oceans (Fig. 8), with skillful forecasts over 15% of land areas in both seasons, primarily in areas where ENSO teleconnections with precipitation are strong. The proportion of the globe with skillful forecasts of the number of days with precipitation above the climatological 90th percentile has a similar level of skill to that for mean precipitation, with no clear pattern of differences in skill between mean and extreme precipitation.

Removing linearly the part of the skill associated with ENSO decreases the proportion of the globe with skillful forecasts of extreme precipitation to 19% in JJA and 16% in DJF, with similar declines in skill for the seasonally averaged precipitation. Almost all areas of the globe exhibit decreased skill in DJF after the

[1] http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.shtml.

relationship with ENSO is removed, with large and statistically significant differences in skill across much of the tropical Pacific and maritime continent (Fig. 9).

The importance of ENSO can also be tested by splitting the database into either neutral years or "ENSO years" (both El Niño and La Niña years) based on classifications from the NOAA CPC (Fig. 10). Consistent with the results for Fig. 9, we see substantially higher skill in years with a strong ENSO signal, with skillful forecasts over only 12% of the globe during the JJA of neutral years, and just 7% of the globe during DJF.

The global relationship between skill in the seasonal mean and extreme precipitation is also slightly higher during ENSO years (Fig. 11). This is not solely a consequence of the higher skill in those years, with a larger difference in the correlation of skill in mean and extreme precipitation between ENSO and neutral years when correlations are only performed across gridpoints where forecasts are skillful. While the short period (27 years) used for the analysis prevents extracting robust conclusions, there are also some areas such as northern India or the southern US where the skill during seasons with a strong ENSO driver is slightly higher for extreme precipitation forecasts than for seasonal average precipitation, which may be of particular relevance for forecasting the impacts of major ENSO events (Fig. 10).

6. Conclusion

The ENSEMBLES collection of seasonal climate hindcasts has skill for forecasting a number of indices for temperature and rainfall extremes, particularly in the tropics, with skill for the multi-model ensemble consistently better than for any individual model. While skill for extreme temperatures is slightly lower than that for the seasonal average of daily-mean temperature, there are some areas of the globe where skill for extreme temperatures can be higher than that for seasonal means. This includes forecasts for warm winter days in Australia and the southern US, as well as forecasts of cold summer nights in parts of central Africa. The skill is generally lower for both mean and extreme rainfall, particularly in land areas, but some skill is observed in forecasts for extreme DJF rainfall in places like northern India.

As observed in previous studies the El Niño-Southern Oscillation is an important source of skill across the globe, particularly during DJF; however, forecasts of mean and extreme temperatures remain skillful in many regions even when the relationship with ENSO is removed. Several areas of the globe have little to no skill for rainfall forecasts when skill is assessed using ENSO-neutral years only. Australia and much of Southern America are among these regions.

One interesting result is a lack of relationship between skill in forecasting the two tails of the temperature distribution; indeed, areas with increased skill for forecasting warm daytime temperatures frequently show a corresponding decrease in skill for forecasts of cold nights. This is an area in need of further research, especially when it is used to analyse how the individual models represent changes in the probability density function (PDF) of seasonal temperatures.

This study has focused on the use of anomaly correlation as a measure of skill, using the ERA-Interim reanalysis dataset as a proxy to the observations. This choice may have impacts on the reliability of our results for extremes in areas such as Africa where reanalyses are poorly constrained by observations. Further research could address the ability of the model to forecast seasons with an unusually high number of extreme days using more adequate measures for extreme events such as the Symmetric Extremal Dependence Index (SEDI; Ferro and Stephenson, 2011).

Acknowledgments

We acknowledge the World Climate Research Programme (WCRP) and the International Centre for Theoretical Physics (ICTP), which have supported this research through the WCRP-ICTP Summer School on Extremes (2014). The authors would like to acknowledge summer school participants Karthik Kashinath, Sarah Abelan and Ulrich Diasso who assisted with the original study development as part of the summer school. The authors would also like to thank four anonymous reviewers, whose comments substantially improved this paper. This project has also received funding from the Australian Research Council Centre of Excellence for Climate System Science, grant CE110001028, the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreements 607085 (EUCLEIA) and 308378 (SPECS), and Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) PIP 112-201201-00626-CO. Leandro B. Díaz was supported by a Ph.D. grant from CONICET, Argentina.

References

Alessandri, A., Borrelli, A., Navarra, A., Arribas, A., Déqué, M., Rogel, P., Weisheimer, A., 2011. Evaluation of probabilistic quality and value of the ENSEMBLES multimodel seasonal forecasts: comparison with DEMETER. Mon. Weather Rev. 139, 581–607. http://dx.doi.org/10.1175/2010MWR3417.1.

Barnston, A.G., 1994. Linear statistical short-term climate predictive skill in the northern hemisphere. J. Clim. 7, 1513–1564. http://dx.doi.org/10.1175/1520-0442(1994)007 < 1513:LSSTCP > 2.0.CO;2.

Barnston, A.G., Mason, S.J., 2011. Evaluation of IRI's seasonal climate forecasts for the extreme 15% tails. Weather Forecast. 26, 545–554. http://dx.doi.org/10.1175/WAF-D-10-05009.1.

Barnston, A.G., Mason, S.J., Goddard, L., Dewitt, D.G., Zebiak, S.E., 2003. Multimodel ensembling in seasonal climate forecasting at IRI. Bull. Am. Meteorol. Soc. 84, 1783–1796. http://dx.doi.org/10.1175/BAMS-84-12-1783.

Batté, L., Déqué, M., 2011. Seasonal predictions of precipitation over Africa using coupled ocean-atmosphere general circulation models: skill of the ENSEMBLES project multimodel ensemble forecasts. Tellus A 63, 283–299. http://dx.doi.org/10.1111/j.1600-0870.2010.00493.x.

Becker, E.J., van den Dool, H., Peña, M., 2012. Short-term climate extremes: prediction skill and predictability. J. Clim. 26, 512–531. http://dx.doi.org/10.1175/JCLI-d-12-00177.1.

Brier, G.Q., 1950. Verification of forecasts expressed in terms of probability. Mon. Weather Rev. 78, 1–3. http://dx.doi.org/10.1175/1520-0493(1950)078 < 0001:VOFEIT > 2.0.CO;2.

Bröcker, J., 2012. Evaluating raw ensembles with the continuous ranked probability score. Q. J. R. Meteorol. Soc. 138, 1611–1617. http://dx.doi.org/10.1002/qj.1891.

Caesar, J., Alexander, L., Vose, R., 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. http://dx.doi.org/10.1029/2005JD006280.

Cottrill, A., Hendon, H.H., Lim, E.-P., Langford, S., Shelton, K., Charles, A., McClymont, D., Jones, D., Kuleshov, Y., 2013. Seasonal forecasting in the Pacific using the coupled model POAMA-2. Weather Forecast. 28, 668–680. http://dx.doi.org/10.1175/WAF-d-12-00072.1.

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., Vitart, F., 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q. J. R. Meteorol. Soc. 137, 553–597. http://dx.doi.org/10.1002/qj.828.

Doblas-Reyes, F.J., García-Serrano, J., Lienert, F., Biescas, A.P., Rodrigues, L.R.L., 2013. Seasonal climate predictability and forecasting: status and prospects. Wiley Interdiscip. Rev. Clim. Change 4, 245–268. http://dx.doi.org/10.1002/wcc.217.

Dole, R., Hoerling, M., Kumar, A., Eischeid, J., Perlwitz, J., Quan, X.-W., Kiladis, G., Webb, R., Murray, D., Chen, M., Wolter, K., Zhang, T., 2013. The making of an extreme event: putting the pieces together. Bull. Am. Meteorol. Soc. 95, 427–440. http://dx.doi.org/10.1175/BAMS-d-12-00069.1.

Donat, M.G., Sillmann, J., Wild, S., Alexander, L.V., Lippmann, T., Zwiers, F.W., 2014. Consistency of temperature and precipitation extremes across various global gridded in situ and reanalysis datasets. J. Clim. 27, 5019–5035. http://dx.doi.org/10.1175/JCLI-d-13-00405.1.

Drosdowsky, W., Chambers, L.E., 2001. Near-global sea surface temperature anomalies as predictors of Australian seasonal rainfall. J. Clim. 14, 1677–1687. http://dx.doi.org/10.1175/1520-0442(2001)014 < 1677:NACNGS > 2.0.CO;2.

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

Epstein, E.S., 1969. A scoring system for probability forecasts of ranked categories. J. Appl. Meteorol. 8, 985–987.
Ferro, C.A.T., Stephenson, D.B., 2011. Extremal dependence indices: improved verification measures for deterministic forecasts of rare binary events. Weather Forecast. 26, 699–713.
Graham, R.J., Yun, W.T., Kim, J., Kumar, A., Jones, D., Bettio, L., Gagnon, N., Kolli, R.K., Smith, D., 2011. Long-range forecasting and the Global Framework For Climate Services. Clim. Res. 47, 47–55. http://dx.doi.org/10.3354/cr00963.
Hagedorn, R., Doblas-Reyes, F.J., Palmer, T.N., 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting–I. Basic concept. Tellus A 57, 219–233. http://dx.doi.org/10.1111/j.1600-0870.2005.00103.x.
Hamilton, E., Eade, R., Graham, R.J., Scaife, A.A., Smith, D.M., Maidens, A., MacLachlan, C., 2012. Forecasting the number of extreme daily events on seasonal timescales. J. Geophys. Res. Atmos. 117, D03114. http://dx.doi.org/10.1029/2011JD016541.
Katsafados, P., Papadopoulos, A., Varlas, G., Papadopoulou, E., Mavromatidis, E., 2014. Seasonal predictability of the 2010 Russian heat wave. Nat. Hazards Earth Syst. Sci. 14, 1531–1542. http://dx.doi.org/10.5194/nhess-14-1531-2014.
Kumar, A., 2009. Finite samples and uncertainty estimates for skill measures for seasonal predictions. Mon. Weather Rev. **137** 2622-263.
Landman, W.A., Beraki, A., 2012. Multi-model forecast skill for mid-summer rainfall over southern Africa. Int. J. Climatol. 32, 303–314. http://dx.doi.org/10.1002/joc.2273.
Luo, L., Zhang, Y., 2012. Did we see the 2011 summer heat wave coming? Geophys. Res. Lett. 39, L09708. http://dx.doi.org/10.1029/2012GL051383.
Manzanas, R., Frías, M.D., Cofiño, A.S., Gutiérrez, J.M., 2014. Validation of 40 year multimodel seasonal precipitation forecasts: the role of ENSO on the global skill. J. Geophys. Res. Atmos. 119, 1708–1719. http://dx.doi.org/10.1002/2013JD020680.
Marshall, A.G., Hudson, D., Wheeler, M.C., Alves, O., Hendon, H.H., Pook, M.J., Risbey, J.S., 2013. Intra-seasonal drivers of extreme heat over Australia in observations and POAMA-2. Clim. Dyn., 1–23. http://dx.doi.org/10.1007/s00382-013-2016-1.
Olkin, I., Finn, J.D., 1995. Correlations redux. Psychological Bulletin 118 (1), 155–164. http://dx.doi.org/10.1037/0033-2909.118.1.155.
Peng, P., Kumar, A., Wang, W., 2011. An analysis of seasonal predictability in coupled model forecasts. Clim. Dyn. 36, 637–648.
Phelps, M.W., Kumar, A., O'Brien, J.J., 2004. Potential predictability in the NCEP CPC dynamical seasonal forecast system. J. Clim. 17, 3775–3785. http://dx.doi.org/10.1175/1520-0442(2004)017 < 3775:PPITNC > 2.0.CO;2.
Storch, H.v., Zwiers, F.W., 2001. Statistical analysis in climate research. Cambridge University Press, 484 pp.
Trenberth, E.K., Branstrator, G.W., Karoly, D., Kumar, A., Lau, N.-C., Ropelewski, C., 1998. Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. J. Geophys. Res. 103 (C7), 14291–14324.
Wang, E., Xu, J., Jiang, Q., Austin, J., 2009. Assessing the spatial impact of climate on wheat productivity and the potential value of climate forecasts at a regional level. Theor. Appl. Climatol. 95, 311–330. http://dx.doi.org/10.1007/s00704-008-0009-5.
Weigel, A.P., Liniger, M.A., Appenzeller, C., 2008. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? Q. J. R. Meteorol. Soc. 134, 241–260. http://dx.doi.org/10.1002/qj.210.
Weisheimer, A., Doblas-Reyes, F.J., Palmer, T.N., Alessandri, A., Arribas, A., Déqué, M., Keenlyside, N., MacVean, M., Navarra, A., Rogel, P., 2009. ENSEMBLES: a new multi-model ensemble for seasonal-to-annual predictions—skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. Geophys. Res. Lett. 36, L21711. http://dx.doi.org/10.1029/2009GL040896.