Reliability of African climate prediction and attribution across timescales

Fraser C Lott, Margaret Gordon, Richard J Graham, Adam A Scaife and Michael Vellinga

Met Office Hadley Centre, Fitzroy Road, Exeter EX1 3PB, UK
E-mail: fraser.lott@metoffice.gov.uk

Received 4 June 2014, revised 22 August 2014
Accepted for publication 22 September 2014
Published 20 October 2014

Abstract
This study investigates the reliability of seasonal to multi-decadal climate simulations of the wet seasons of several key African regions. It is found that reliability varies across regions and seasons, and that simulations of precipitation are universally less reliable than simulations of temperature. Similar levels of reliability are found across all the timescales considered for most (but not all) region/season combinations. Reliability for temperatures increases on longer timescales, both due to the differences in the modelling systems for each timescale and, in part, due to the contribution from systematic climate warming. Though the use of reliability is well-established for forecasting, its meaning for attribution is less clear, and further work is underway to further clarify this.

Keywords: climate, Africa, reliability, skill, seasonal, decadal, simulation

Introduction
Reliability is a measure of the ability of a forecast system to predict events with probabilities which match the observed frequency of occurrence [1]. If similar processes are present, the values deduced from retrospective climate predictions and simulations can be used to infer the reliability of current predictions. It has even been suggested [2] that the reliability of seasonal hindcasts can be used to discount future climate projections. However, it is not yet clear whether processes across the two timescales are sufficiently similar for this suggestion to be valid [3]. For some studies attributing climate change [4, 5], reliability has also been used as validation for the underpinning simulations. It is therefore desirable to determine how reliability varies with timescale, particularly in a rapidly developing continent such as Africa, where longer term forecasts and attribution of extreme events can have an influence on which developments are prioritized.

In this study, a shared formulation of forecast probability is used in reliability diagrams to evaluate model performance across a range of timescales in a consistent way, rather than using a definition which varies with spatial resolution and number of model ensemble members. In particular, it is applied to seasonal and decadal forecasts, and to the multi-decadal climate simulations which underpin attribution studies. The reliability and the Brier Skill Score [1, 6] are used to evaluate the performance of these climate simulations on a regional scale within Africa. For this study, they are applied to key African regions as detailed in table 1, focusing on the rainy season and including assessment of extreme seasonal events.

Modelling systems and observations
Three different sets of climate simulations were evaluated. On a seasonal scale, hindcasts from the GloSea 5 [7] seasonal forecast system were analysed for 1996–2009, initialized around the start of February, May, August and November (on days 1 and 9 of the month in question, and day 25 of the previous month). Each hindcast year had 15 members (five members from each initialization day). For the Sahel, the seasons of interest corresponded to months 3–5 of the model runs, so as to correspond with the most intense rainfall period. For all other regions, months 2–4 were used. The GloSea5 system uses an atmospheric model (HadGEM3 with a 0.833
The Met Office Decadal Prediction System 2 [8] was evaluated using retrospective forecasts between 1960 and 2005. Each hindcast year has four members, all initialized on 1st November. Members are run out for five years and results are averaged over years 1–5. The decadal prediction system uses an atmosphere model (HadGEM3 at N96 grid scale, i.e. 1.875 by 1.25 degrees) coupled to a one degree ocean and sea ice model. Greenhouse gas and aerosol concentrations are prescribed using CMIP5 historical forcing. Further details (e.g. the initialization of ocean and atmosphere) can be found in [8]. Both the seasonal and decadal hindcasts used here show skilful prediction of El Niño Southern Oscillation (ENSO) [7, 8].

On multi-decadal timescales, this study evaluated 20th century climate simulations [4]. These were five members of HadGEM3-A model [9] at N96 scale (i.e. the same atmospheric model used for the seasonal and decadal predictions, including all historical forcings) run between 1960 and 2010 with prescribed sea surface temperatures (SSTs) and sea ice. These are set to the observed SSTs for the corresponding time in the HadISST [10] dataset. These runs were originally produced in order to validate and bias-correct event attribution experiments [4], where the fixed SSTs enable these simulations to be as close to observed events as possible. However, they have an additional useful benefit to this study, since they also provide a measure of potential predictability [11]. As seasonal forecasts are heavily dependent on their simulation of SST, using observed SST should bring them close to the theoretical limit of the prediction system (not including the modelling of two-way coupled systems, such as the Madden-Julian oscillation [12]).

All observations (temperature and rainfall) used for evaluating the models in this paper come from the CRU TS 2.1 dataset [13], which covers the entire time period used by the modelling systems. The results of reliability diagrams will depend upon the limitations of the observation dataset which has been used, but it is beyond the scope of this paper to examine this effect using different datasets.

### Calculating reliability

A typical reliability diagram [1] shows observed frequency against modelled/forecast probability. Points along the 1–1 line will indicate perfect reliability, such that the event in question occurs with a frequency which matches the forecast probability.

To actually plot such a diagram, a decision must be made as to what constitutes an ‘event’. The easiest way is to set a criterion, such as an above-average temperature over a season, or precipitation below some threshold in that region. Such criteria (detailed further later) were chosen in this study during the rainy season of each region, corresponding to heat waves and droughts. This then enables the calculation of the observed frequency and forecast probability of achieving that threshold. These figures can be arrived at in a number of ways and, for this study, a shared method was devised to measure regional reliability across the three sets of simulations.

Conventionally, predicted probabilities are calculated purely using the fraction achieving the threshold across the model ensemble. This is done for single model grid boxes and then aggregated, or occasionally variables averaged over the region are used. The single grid box method is a standardized WMO technique [14], but this requires a large sample size, as well as there being an implicit assumption that the probability in each box is independent. In the case of this study, a new technique is devised to increase the sample size and reduce sensitivity to the grid box size and the ensemble size, thereby enabling it to be used for any of the variously-sized model

| Region               | Longitude   | Latitude   | Season(s)          |
|----------------------|-------------|------------|--------------------|
| Sahel                | 20°W–30°E   | 10°N–15°N  | July–September (JAS) |
| Southern Africa      | 20°E–60°E   | 35°S–10°S  | December–February (DJF) |
| Greater Horn of Africa | 30°E–55°E  | 10°S–10°N  | September–November (SON) March–May (MAM) |
| Gulf of Guinea       | 20°W–15°E  | 0°–10°N   | September–November (SON) March–May (MAM) |

Table 1. Regions and season combinations used in this study.

**Figure 1.** An unbinned reliability diagram created using the area-fraction technique, in this case for above-median Southern African wet-season (December–February) temperatures between 1960 and 2010, from multidecadal climate simulations.
ensembles used at the different timescales. The set of grid boxes within each study region and season described above is used as a data pool, as if each originated from a different ensemble member. (This will lead to a regionally-averaged probability, so will be closest to results of a conventional reliability diagram where the region of study is relatively homogeneous. To this end, only the land points in each region are used.) The method is then as follows:

1. For the observations and for each model ensemble member, calculate the proportion of the region for which the event criterion is met.
2. Calculate the mean proportion for all model ensemble members for that event. This is taken to represent the forecast probability of an event, such as a drought, for locations in that region.
3. Plot the observed fraction of the region meeting the criterion (representing its regional average frequency) against the modelled probability. This produces a figure with each point representing the studied season in each year. An example is shown in figure 1.

The data points in figure 1 are then binned according to their modelled probability. Within each fixed-width bin (10% in figure 2), the values are averaged to a mean modelled and mean observed probability. To indicate the weight that should be given to that binned value, it is plotted as a circle whose area is proportional to the number of contributing samples. Once binned, the skill of the forecasts contributing to the plot can be easily summarized as the Brier Skill Score, the scaled difference between resolution (the difference of the observed frequency of each bin from the mean observed climatology) and reliability (the distance from the ideal diagonal, effectively the slope). Thus a plot with greater reliability than resolution will have a negative Brier Skill Score. This is detailed in equation (1), where $i$ is the bin number, $I$ the total number of bins, $N_i$ the number of events in bin $i$, $o$ is the observed probability and $y$ is the modelled probability. An
The area fraction in the lower quintile of rainfall, plotted by year, for two example region/season combinations. Black-outlined bars show observed fractions, while blue shows the ensemble mean area fraction for multi-decadal model runs.

The ideal Skill Score is $1$, where every point is on the leading diagonal, while a horizontal line along observed climatology, indicating not reliability or resolution, would give $-1$. The dashed no-skill line in subsequent figures indicates the change of sign.

$$\text{Brier Skill Score} = \frac{\text{Reliability} - \frac{1}{n} \sum_i N_i (\sigma_i - \sigma)^2 - \frac{1}{n} \sum_i N_i (y_i - \sigma_i)^2}{\sigma(1 - \sigma)}.$$  \hspace{1cm} (1)

Points in the reliability diagram which lie along a horizontal line centred on the observed climatological probability of that event show no resolution. This means that the model is unable to distinguish between events of high and low probability, making any forecasts from this simulation over-confident. If points follow a line of positive gradient, this will have resolution, but will only make a positive contribution to the model skill according to the Brier Skill Score beyond a line of $1$ in $2$ gradient (i.e, where the probability seen in the model increases at twice the rate of that observed, or where the reliability and resolution in equation (1) are equal) which passes through the point of observed and model climatology.

**Comparison of temperature and precipitation reliability**

Figure 2 shows the reliability for the above-median temperature and below-median rainfall in climate simulations of the Southern Africa summer (DJF) across all three timescales, and is chosen because it is typical of many of the other reliability plots in this study. All three timescales have near-perfect reliability for temperature, which improves as the timescale increases. Part of this is likely to be due to a greater sample size from the longer simulations involved conveying an improvement in signal-to-noise ratio, but a longer timescale also enables the reliability to be influenced by the temperature trend over the last few decades. This can give additional reliability, since the presence of the trend in both the simulations and observations gives the same shift in the probability of extremes $[3]$. Precipitation is more localized in nature, and more dependent on large-scale circulation $[2]$, than temperature, making it much harder to predict. The pooling of events into larger regions, as in this study, alleviates this to some degree. Nevertheless, it is clear from figure 2 that predictions for precipitation are less skilful than for temperature (this can also be seen for the later analysis shown in table 2). Previous studies have often found this to be the case $[2, 15]$, including previous work on African regions $[5]$. Predictions of precipitation can be improved in certain regions and seasons where there are known teleconnections to large-scale SST phenomena, such as ENSO, which drive rainfall $[15]$.

Comparing the reliability of southern African summer temperatures and rainfall in figure 2 gives a good illustration of the reliability of model predictions for temperature and rainfall. Most notably on the seasonal and multi-decadal timescales, the model populates little more than the two bins around the climatology point of the rainfall reliability diagram. Thus, it is effectively only able to offer a climatological prediction, resulting in reliability but no resolution or predictive ability. Since the multi-decadal simulations have observed SSTs, and the other timescales have similar skill, this is unlikely to be an issue with their ocean modelling. Instead, further teleconnections would need to be understood in order to improve the skill. Southern Africa is the most extreme example; Sahel summer (JAS) and Greater Horn of Africa short rains (SON) offer better reliability. In contrast, the long rains in the Greater Horn (MAM) show little reliability, resolution or skill. This will be discussed further in the following section, and can be seen in the upper panels of figure 4.

**Reliability and skill of quintiles**

This study is particularly concerned with the reliability of forecasting and attribution systems for the extreme events to which African regions are vulnerable, specifically droughts and heat waves. It could be argued, therefore, that the reliability of events relative to the median is not sufficiently extreme to represent the statistics of the real events of interest. Conversely, there is a limit whereby the events being studied are sufficiently rare that it is not possible to provide a large enough sample to evaluate. It is therefore optimal to investigate upper-quintile temperature (i.e. in the top $20\%$ of the record) which represents heat waves, and lower-quintile (bottom $20\%$) rainfall representing drought conditions. As an example of the relationship between lower-quintile rainfall and drought conditions, figure 3 shows the fraction of the
Figure 4. A comparison of reliability between the short rains (left, September–November) and the long rains (right, March–May) in the Greater Horn of Africa region, and between below-median rainfall (top diagrams) and lower-quintile rainfall (lower diagrams). These examples are for the multi-decadal timescale.

Table 2. The mean Brier Skill Score across timescales in upper-quintile temperature (T) and lower-quintile rainfall (pr) for the six region/season combinations studied. Values around zero denote skill no better than climatology. Values greater than 0.25 are shown in bold, approximately corresponding to the Brier Skill Score falling in the 80th percentile of a Monte–Carlo distribution (calculated by comparing the model to observations from random year, rather than its own).

|                      | Sahel | Southern Africa | Greater Horn of Africa | Gulf of Guinea |
|----------------------|-------|-----------------|------------------------|---------------|
|                      | No. of members | JAS | DJF | SON | MAM | SON | MAM |
| Seasonal T           | 15    | 0.34            | 0.05                   | 0.20          | −0.13 | −0.18 | 0.39 |
| Decadal T            | 4     | 0.18            | 0.20                   | 0.41          | 0.51  | 0.25  | 0.14 |
| Multi-decadal T      | 5     | 0.25            | 0.19                   | 0.34          | 0.16  | 0.28  | 0.30 |
| Seasonal pr          | 15    | 0.05            | −0.01                  | 0.06          | −0.08 | −0.02 | −0.03 |
| Decadal pr           | 4     | 0.18            | 0.02                   | −0.03         | −0.03 | 0.07  | −0.10 |
| Multi-decadal pr     | 5     | 0.07            | 0.01                   | 0.00          | 0.00  | −0.01 | 0.00 |
Sahel and the Greater Horn of Africa which experienced the lower quintile of rainfall between July and September, and March and May (the ‘long rains’), respectively.

The proportion of the regions meeting these quintile-related criteria in simulations and observations can once again be compared using reliability diagrams. As can be seen in figure 4, the binned quintiles tend to cluster at 20% probability, and this clustering is greater than seen with the median threshold.

To effectively summarize this study without including an excessive number of plots in this study, the Brier Skill Score was calculated for simulated upper-quintile temperature and lower-quintile rainfall across all the regions and seasons. These are shown in table 2, where it is clear that there is variation in skill between regions, and that temperature is generally more skillful than precipitation, as discussed earlier. Sahel summer appears to be the most skillful region overall for temperature and rainfall. The March–May precipitation has particularly low skill for both the Gulf of Guinea and Greater Horn of Africa, whilst temperature appears least predictable for the Gulf of Guinea between September and November. Together with Greater Horn long rains, this also sees the highest variation in skill across timescales. It is notable for Sahel JAS and Gulf of Guinea MAM temperature, skill is highest for the seasonal prediction system. It would seem that, in these cases, the higher resolution of this system is more influential on the skill than increased timescale on the other systems.

Discussion

It is clear from this study that care should be taken when interpreting the results of African climate simulations, for regions and variables that do not show good reliability or skill for the recent past. This is already common practice in seasonal forecasting, but it is important to reiterate this for decadal prediction and event attribution in vulnerable regions, such as those African regions discussed here.

The results of this study suggest caution should be taken in both climate prediction and attribution of extreme events, when seasonal and regional reliability is low. It is particularly important that the reliability be examined for precipitation in regions such as the Greater Horn of Africa and the Gulf of Guinea. Though literature is extensive [e.g. 1, 16] in describing the importance of reliability and Brier Skill Score to forecasts, and how this may be used to recalibrate forecasts, there has been little study of what these figures mean for attribution, particularly the new field of event attribution [4, 17, 18]. The link between reliability and attribution statistics, such as the fraction of attributable risk, needs to be established, particularly as reliability does not look at the effect of the presence or absence of anthropogenic climate forcings. Evaluation of attribution models is one aim of EUCLÉIA [19], a European project examining the many aspects of event attribution. It may prove necessary to develop a new diagnostic more suited to these studies. Once it is clearer how model skill and reliability relate to the presence or absence of external climate forcings, attribution studies may also be recalibrated.

This report shows that forecasting systems show similar features across different timescales and in different regions, and that reliability can increase with timescale. However, it is clear that some regions are neither reliably nor skillfully simulated. Where possible, these should be recalibrated, while further research is required to establish how this can be achieved for attribution.

Acknowledgements

The authors would like to thank Nikos Christidis, Gareth S Jones, Dáithí A Stone and Peter A Stott for discussions during this study, and the reviewers of this paper for their input. This document is an output from a project funded by the UK Department for International Development (DFID) for the benefit of developing countries. The views expressed are not necessarily those of DFID. The research leading to these results has received funding from the European Union’s Seventh Framework Programme (FP7/2007–2013) under grant agreement n° 607085. This work was supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101). Reproduced with the permission of the Controller of Her Majesty’s Stationary Office.

References

[1] Wilks D S 2011 Statistical Methods in the Atmospheric Sciences 3rd edn (Amsterdam: Elsevier) chapter 7
[2] Palmer T N et al 2008 Toward seamless prediction calibration of climate change projections using seasonal forecasts Bull. Am. Meteorol. Soc. 89 459–70
[3] Scaife A A et al 2009 Comment on toward seamless prediction: calibration of climate change projections using seasonal forecasts Bull. Am. Meteorol. Soc. 90 1549–51
[4] Christidis N et al 2012 A new HadGEM3-A based system for attribution of weather and climate-related extreme events J. Clim. 26 2756–83
[5] Lott F C, Stott P A and Christidis N 2013 Can the 2011 East African drought be attributed to human-induced climate change? Geophys. Res. Lett. 40 1–5
[6] Brier G W 1950 Verification of forecasts expressed in terms of probabilities Mon. Weather Rev. 78 1–3
[7] MacLachlan C et al 2014 Global seasonal forecast system version 5 (GloSea5): a high resolution seasonal forecast system Q. J. R. Meteorol. Soc. doi:10.1002/qj.2396
[8] Knight J R et al 2014 Predictions of climate several years ahead using an improved decadal prediction system J. Clim. doi:10.1175/JCLI-D-14-00069.1 (in press)
[9] Walters D N et al 2011 The met net unified model global atmosphere 3.0/3.1 and JULES global land 3.0/3.1 configurations Geosci. Model Dev. 4 919–41
[10] Rayner N A et al 2003 Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century J. Geophys. Res. 108 4407
[11] Chervin R M 1986 Interannual variability and seasonal climate predictability J. Atmos. Sci. 43 233–51
[12] Madden R A and Julian P R 1971 Detection of a 40–50 day oscillation in the zonal wind in the tropical Pacific J. Atmos. Sci. 28 702–8
[13] Mitchell T D and Jones P D 2005 An improved method of constructing a database of monthly climate observations and associated high-resolution grids *Int. J. Climatol.* 25 693–712

[14] WMO 2002 Standardised verification system (SVS) for long-range forecasts (LRF) New Attachment II-9 to the Manual on the GDPS (WMO-No. 485) (Geneva: WMO)

[15] Arribas A et al 2011 The GloSea4 ensemble prediction system for seasonal forecasting *Mon. Weather Rev.* 139 1891–910

[16] Kharin V V and Zwiers F W 2003 Improved seasonal probability forecasts *J. Clim.* 16 1684–701

[17] Allen M 2003 Liability for climate change *Nature* 421 892

[18] Stott P A et al 2013 Attribution of weather and climate-related extreme events *Position paper for WCRP OSC*, in *Monograph: Climate Science for Serving Society: Research, Modelling and Prediction Priorities* (Berlin: Springer)

[19] EUCLEIA project website (http://eucleia.eu)