A New HadGEM3-A-Based System for Attribution of Weather- and Climate-Related Extreme Events

NIKOLAOS CHRISTIDIS, PETER A. STOTT, ADAM A. SCAIFE, ALBERTO ARRIBAS, GARETH S. JONES, DAN COPSEY, JEFF R. KNIGHT, AND WARREN J. TENNANT

Met Office Hadley Centre, Exeter, United Kingdom

(Manuscript received 31 January 2012, in final form 8 October 2012)

ABSTRACT

A new system for attribution of weather and climate extreme events has been developed based on the atmospheric component of the latest Hadley Centre model. The model is run with either observational data of sea surface temperature and sea ice or estimates of what their values would be without the effect of anthropogenic climatic forcings. In that way, ensembles of simulations are produced that represent the climate with and without the effect of human influences. A comparison between the ensembles provides estimates of the change in the frequency of extremes due to anthropogenic forcings. To evaluate the new system, reliability diagrams are constructed, which compare the model-derived probability of extreme events with their observed frequency. The ability of the model to reproduce realistic distributions of relevant climatic variables is another key aspect of the system evaluation. Results are then presented from analyses of three recent high-impact events: the 2009/10 cold winter in the United Kingdom, the heat wave in Moscow in July 2010, and floods in Pakistan in July 2010. An evaluation assessment indicates the model can provide reliable results for the U.K. and Moscow events but not for Pakistan. It is found that without anthropogenic forcings winters in the United Kingdom colder than 2009/10 would be 7–10 times (best estimate) more common. Although anthropogenic forcings increase the likelihood of heat waves in Moscow, the 2010 event is found to be very uncommon and associated with a return time of several hundred years. No reliable attribution assessment can be made for high-precipitation events in Pakistan.

1. Introduction

Weather and climate extreme events, often accompanied with adverse socioeconomic impacts, are of great interest to both the public and policy makers. In the aftermath of a catastrophic event, scientists are invariably challenged to explain possible links between the event and human influences on the climate. While most of the observed warming on global scales has been attributed to emissions of greenhouse gases with a likelihood of more than 90% (Hegerl et al. 2007) according to the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4), the attribution of changes in extremes to possible causes is a more difficult task. First, the rarity of the events requires long observational records with good spatial coverage that are not always available. Second, information most relevant to adaptation planning comes from regional rather than global studies and smaller spatial scales are associated with a lower signal-to-noise ratio (Stott et al. 2011a).

Despite these challenges, a lot of progress has been made in recent years on the attribution of observed changes in extreme indicators. A number of studies that analyze information from observations and general circulation model (GCM) experiments in a formal statistical framework have already detected the prominent role of anthropogenic forcings in the observed post-1950s warming in extreme daily temperatures (Christidis et al. 2005, 2011; Zwiers et al. 2011) and the intensification of extreme precipitation in the same period (Min et al. 2011) on quasi-global scales. Analyses of regional changes in temperature extremes in recent decades are also indicative of the prominence of the anthropogenic fingerprint in continental and subcontinental-scale regions (Meehl et al. 2007; Stott et al. 2011b; Morak et al. 2011). The extremes in these studies are defined using simple indicators like the warmest or wettest day of the year, threshold exceedances, etc.
While attribution of long-term changes in regional indicators that describe extremes provides valuable information, it is attribution assessments of individual high-impact events that are of greatest interest to the public and various stakeholders. Examples of such recent events are numerous. Two of the most severe natural hazards in living memory occurred in summer 2010: the heat wave in the Moscow area, marked by daily temperature anomalies of over 10°C in some of the affected regions (Darmenov et al. 2010; Dole et al. 2011), and the devastating floods in Pakistan, the worst ever known in the region (Houze et al. 2011; Webster et al. 2011). High-impact events continued into 2011, with three-quarters of the state of Queensland in Australia declared a disaster zone due to flooding early in the year (Sweet 2011) and the summer drought in East Africa that triggered a food crisis in Somalia, Ethiopia, and Kenya (Funk 2011). While a reliable forecasting system can help the community prepare against impending catastrophic events, a reliable attribution system can help the community adapt to changes in the frequency of such events. Here, we concentrate only on the attribution aspect, bearing in mind that timely and robust information from the scientific community is essential to decision makers who enquire whether high-impact events may become more or less frequent in a changing climate.

Attribution of weather and climate extreme events aims to estimate how a specific cause of climate change has altered the probability, or frequency, of these events (Allen 2003; Stott et al. 2012). High-impact extreme events may of course occur solely as a result of unforced climate variability. For example, a severe heat wave in a certain region would lie in the warm tail of the distribution of the temperatures expected in that region because of internal climatic variations and would be associated with a long return period. However, if the entire distribution shifted to higher temperatures as a result of a climatic forcing like greenhouse gas emissions, then the same event would become more common in the altered climate (it would be nearer the central sector of the new temperature distribution) and its return period would be reduced. The goal of the new system described in this paper is to quantify how the interplay between natural variability and anthropogenic forcings might have altered the frequency of various types of extreme events.

Stott et al. (2004) is the first example of formal attribution with reference to a specific event, the European heat wave in summer 2003, which cost the lives of tens of thousands of people (Robine et al. 2008). The authors estimated how the summer temperatures respond to anthropogenic and natural forcings in a region affected by the heat wave using optimal fingerprinting (Hasselmann 1993; Allen and Stott 2003) and combined the decadal change in temperature with model estimates of year-to-year internal variability. The result was a pair of distributions of the annual mean summer temperature in the region, one for the actual climate (with both natural and anthropogenic forcings) and one for a hypothetical natural climate without human influences. They found that human influences have at least doubled the odds of having a summer as hot as 2003. The methodology has been recently extended to examine temperature extremes in 22 more subcontinental regions (Christidis et al. 2012).

Following a different approach, Pall et al. (2011) carried out an attribution analysis on the autumn 2000 floods in the United Kingdom. They also compared two different types of climate, one with and one without the influence of greenhouse gas emissions. However, the distributions of precipitation and other climatic variables came directly from two experiments with an atmospheric GCM that, unlike previous work, use prescribed sea surface temperatures (SSTs) and sea ice for the two climatic regimes. Each experiment consisted of a large ensemble of simulations (typically a few thousand members). The output of these model runs was then fed into a precipitation-runoff model to provide the distributions of the daily river runoff, the measure of flooding used in the study. They found that greenhouse gas emissions very likely (probability > 90%) increased the flooding risk in the reference region in autumn 2000 by more than 20%. However, Pall et al. (2011) did not establish whether similar events are reliably reproduced in past cases with the model they employed.

The new attribution system discussed in this paper adopts the Pall et al. (2011) approach; that is, it is based on ensembles generated by an atmosphere-only model. In this study, we examine the overall effect of all anthropogenic forcings on the frequency of extreme events, though the system can also be used to single out the effect of other forcings or forcing combinations. Unlike previous studies, we focus more extensively on the evaluation of the new system. It is imperative to provide information that is robust and reliable, as poor attribution assessments could lead to poor adaptation decisions. The evaluation of the new system is therefore deemed as of great importance and we propose here a number of tests to assess the skill of the system in representing different types of extremes in different regions. While attribution analyses can be made for any given event, it is only when the model can reliably predict the event or the frequency of events of the same type that the output information has real value to the user.
The remainder of the paper is structured as follows: Section 2 describes how the new attribution system was set up and summarizes the experimental design for the representation of the actual and natural climate. Section 3 provides details on the evaluation of the system and presents assessments of the model skill with regard to regional temperature and precipitation extremes. The first experiment with the new system examines the climate in year 2010 and is discussed in section 4. Results for three high-impact extreme events are shown in the same section. The main conclusions and some discussion are given in section 5.

2. The development of the new system

The new attribution system is based on Hadley Centre Global Environmental Model version 3-A (HadGEM3-A), the atmosphere-only component of the HadGEM3 (Hewitt et al. 2011), run at a N96 horizontal resolution (1.25° longitude by 1.875° latitude) with 38 vertical levels. As part of the HadGEM family of models, HadGEM3-A has a nonhydrostatic dynamical core (Davies et al. 2005), employing a semi-implicit, semi-Lagrangian time integration scheme. The improved dynamics of the model offer an advantage relative to Hadley Centre Atmospheric Climate Model version 3 (HadAM3; Pope and Stratton 2002), used in Pall et al. (2011), though the simulations here are produced at a lower horizontal resolution (N96 as opposed to N144) to reduce the computational cost. The methodology requires large ensemble simulations to represent the climate with and without human influences. Close links with research groups in the Met Office working on ensemble applications benefit the project, as technical facilities and expert advice on the evaluation of ensembles (section 3) are readily available. Examples of such ensemble applications include short-range, global, and regional forecasting (Bowler et al. 2007); quantifying model uncertainty in coupled model runs (Murphy et al. 2004); and seasonal and decadal forecasting (Arribas et al. 2011; Smith et al. 2007, 2010).

The ensembles for the attribution experiments are generated by systematic sampling of the modeling uncertainty. This is done in two different ways. First, we employ the “perturbed physics” approach, whereby random perturbations are introduced to represent the uncertainty in a number of parameters of some major components of atmospheric and surface physics (e.g., large-scale clouds, convection, radiation, boundary layer, land surface processes, etc.). Examples of ensembles generated by introducing perturbations to 29 parameters can be found in the studies of Murphy et al. (2004), Collins et al. (2006), and Webb et al. (2006). In addition to the perturbed physics approach, we also make use of a stochastic kinetic energy backscatter (SKEB) scheme (Tennant et al. 2011). The scheme adds wind increments at each model time step to counteract energy loss due to numerical smoothing and account for unrepresented grid-scale convective sources of kinetic energy. The same two schemes are employed by the Met Office short-range, medium-range, and seasonal prediction systems.

The two experiments for the climate with and without human influences consist of ensemble simulations with different climatic forcings, and different SSTs and sea ice specifications. Initial conditions come from the European Centre for Medium-Range Weather Forecasts (ECMWF) Interim Re-Analysis (ERA-Interim; Dee et al. 2011). The simulations for the actual climate include anthropogenic forcing factors that comprise changes in well-mixed greenhouse gases, aerosols (sulfate, black carbon, and biomass burning), tropospheric and stratospheric ozone, and land-use changes, as well as natural forcings that comprise changes in volcanic aerosols and solar irradiance. The simulations for the natural climate include the two natural forcing factors only. The same forcings were used in simulations with the HadGEM1 (Martin et al. 2006) that provided data for the IPCC AR4. Full details on the forcings are given in Stott et al. (2006). The boundary conditions used in the simulations of the two climate scenarios is a key difference between the two experiments. The SST and sea ice data prescribed in the simulations of the actual world come from the Hadley Centre Sea Ice and Sea Surface Temperature dataset (HadISST; Rayner et al. 2003). For the natural world simulations, a model estimate of the anthropogenic change in the SSTs is subtracted from the HadISST data. Section 4 provides details of how these estimates were calculated for the first study with the new system. The change in the sea ice coverage under the effect of anthropogenic forcings is estimated based on an empirical relationship as in Pall et al. (2011). For each hemisphere we applied a linear fit to HadISST gridpoint data to model the relationship between SST and sea ice. We then applied this simple linear model to the estimated anthropogenic change in the SSTs to compute the change in the sea ice, making sure we limit the sea ice fraction in the natural climate to vary between 0 and 1. As the linear relationship may not accurately represent the sea ice change, it is important that we employ a more sophisticated approach in the future. However, in these early stages of the system development, the linear approach is deemed sufficient, at least for the study of events that are not directly influenced by changes in the sea ice.
Attribution studies of extreme events with the new system use the ensembles for the two types of climate to estimate how a climate driver (in this paper anthropogenic forcings) has changed the probability of an event occurring (Allen 2003). There are several studies that report such estimates of changing probabilities obtained with coupled GCMs (e.g., Stott et al. 2004; Christidis et al. 2012) or with the same approach as in this paper (Pall et al. 2011). First, two distributions of a climatic variable (temperature, precipitation, drought index, etc.) are constructed from the model ensemble runs for the climate with and without human influences and a threshold that describes the event is defined. The probabilities of exceeding the threshold (or going below it, when, e.g., looking at a cold event) is then estimated with \( P_1 \) and without \( P_0 \) human influences and the fraction of attributable risk (FAR; Allen 2003) is calculated as \( FAR = 1 - \frac{P_0}{P_1} \). Positive values of FAR imply the event has become more common because of the anthropogenic effect on the climate, whereas negative values suggest it would have been more common in the alternative natural world. If the chosen threshold lies at the tails of the distributions, a statistical extreme distribution needs to be employed to provide estimates of \( P_0 \) and \( P_1 \). Moreover, the model output can be fed into regional models to study local events or even socioeconomic models to study a range of climate change impacts. In Pall et al. (2011) the HadAM3 output was used to drive a precipitation-runoff model from which the change in the flood risk was estimated.

### 3. Model evaluation

To attribute changes in the odds of a specific extreme event, we need to have confidence that the model has the right mechanisms to reproduce the event under consideration. Events with a high degree of predictability would be strongly linked to the SSTs prescribed in the simulations. For example, given the characteristic SST patterns related to the El Niño–Southern Oscillation (ENSO), Arribas et al. (2011) show that the skill of their ensemble prediction system in reproducing boreal winter precipitation in the Greater Horn of Africa increases markedly during El Niño and La Niña years.

Climate models used in attribution systems may not be able to forecast a specific extreme event but should nonetheless be able to reproduce events like the one under investigation with realistic frequency and climatological characteristics. With regard to events with a low predictive skill, where causal factors cannot be identified, the attribution system may still be able to provide useful information, if the statistics of the event in question are well represented. For example, if the modeled monthly temperature distribution in a certain region is reliable, then we can still get a reliable estimate of the probability of exceeding a threshold that describes a heat wave event in the month and region under consideration, even if the model has little skill in predicting the event itself. In this case we assess the changing frequency of events of the same kind as the actual one, rather than the changing frequency of the actual event under human influences on the climate.

Reliability is assessed here in two ways: First, we examine the predictive skill of the model (using reliability diagrams discussed below). If the skill is good, then we have confidence that the attribution analysis provides the changing odds of the actual event under consideration. We also compare the modeled distribution and variability of the relevant variable (e.g., temperature) against observations (or reanalysis data). If the model compares well with the observations but has a low predictive skill, it can still be used to estimate the changing frequency of the type of events like the one under consideration, though not of the actual event itself.

In this section we employ (i) reliability diagrams to assess how well the model reproduces various types of extreme events and (ii) some simple measures like probability distributions and power spectra to assess whether the model can reproduce the observed statistics for the same types of extremes. Our evaluation examines how well the model represents the actual climate and the changes that have taken place during the period covered by the runs used for the assessment. We assume that the model representation of the alternative natural climate is of the same quality as of the actual climate. Once we gain confidence that the attribution system provides reliable information, the probabilities of exceeding (or going below) a threshold can be estimated. It should be noted that these probabilities are conditioned by the main modes of variability represented by the prescribed SST. For example, if the event under investigation occurs during a La Niña year, the probabilities will be estimated given La Niña conditions. This is different to previous studies with coupled models (Stott et al. 2004; Christidis et al. 2012), which attempt to sample the full range of internal variability modes when estimating the probabilities.

The evaluation of HadGEM3-A presented here is based on a small ensemble of five multidecadal simulations that span from 1960 to 2010 and include all external forcings (anthropogenic and natural) and observed SSTs and sea ice data from HadISST. Despite the small ensemble size, the considerable length of the simulations ensures that we get enough data to test the model. In comparison, the evaluation of the Met Office seasonal prediction system is based on an ensemble of
TABLE 1. The 23 subcontinental regions used in model validation assessments, their abbreviated names, and coordinates (Giorgi and Francisco 2000).

<table>
<thead>
<tr>
<th>Region</th>
<th>Abbreviation</th>
<th>Coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northern Australia</td>
<td>NAU</td>
<td>28°–10°S, 110°–155°E</td>
</tr>
<tr>
<td>Amazon Basin</td>
<td>AMZ</td>
<td>20°–10°N, 80°–35°W</td>
</tr>
<tr>
<td>Western Africa</td>
<td>WAF</td>
<td>10°S–20°N, 20°W–20°E</td>
</tr>
<tr>
<td>Eastern Africa</td>
<td>EAF</td>
<td>10°S–20°N, 20°–50°E</td>
</tr>
<tr>
<td>Southeast Asia</td>
<td>SEA</td>
<td>10°S–20°N, 95°–155°E</td>
</tr>
<tr>
<td>South Asia</td>
<td>SAS</td>
<td>5°–30°N, 65°–100°E</td>
</tr>
<tr>
<td>Southern Australia</td>
<td>SAU</td>
<td>50°–28°S, 110°E–180°</td>
</tr>
<tr>
<td>Southern South America</td>
<td>SSA</td>
<td>55°–20°S, 75°–40°W</td>
</tr>
<tr>
<td>Central America</td>
<td>CAM</td>
<td>10°–30°N, 115°–85°W</td>
</tr>
<tr>
<td>Western North America</td>
<td>WNA</td>
<td>30°–60°N, 130°–105°W</td>
</tr>
<tr>
<td>Central North America</td>
<td>CAN</td>
<td>30°–50°N, 105°–85°W</td>
</tr>
<tr>
<td>Eastern North America</td>
<td>ENA</td>
<td>25°–50°N, 85°–60°W</td>
</tr>
<tr>
<td>Mediterranean Basin</td>
<td>MED</td>
<td>30°–50°N, 10°W–40°E</td>
</tr>
<tr>
<td>Northern Europe</td>
<td>NEU</td>
<td>50°–75°N, 10°W–40°E</td>
</tr>
<tr>
<td>Southern Africa</td>
<td>SAF</td>
<td>35°–10°S, 10°W–50°E</td>
</tr>
<tr>
<td>Sahara</td>
<td>SAH</td>
<td>20°–30°N, 20°W–65°E</td>
</tr>
<tr>
<td>East Asia</td>
<td>EAS</td>
<td>20°–50°N, 100°–145°E</td>
</tr>
<tr>
<td>Central Asia</td>
<td>CAS</td>
<td>30°–50°N, 40°–75°E</td>
</tr>
<tr>
<td>Tibet</td>
<td>TIB</td>
<td>30°–50°N, 75°–100°E</td>
</tr>
<tr>
<td>North Asia</td>
<td>NAS</td>
<td>50°–70°N, 40°E–180°</td>
</tr>
<tr>
<td>Antarctica</td>
<td>ANT</td>
<td>90°–60°S, 0°–180°</td>
</tr>
<tr>
<td>Alaska</td>
<td>ALA</td>
<td>60°–70°N, 170°–105°W</td>
</tr>
<tr>
<td>Greenland</td>
<td>GRN</td>
<td>50°–85°N, 105°W–10°E</td>
</tr>
</tbody>
</table>

9–12 simulations that cover a period of 14 yr. We examine extremes that fall under three categories, characterized by high temperatures, low temperatures, and high precipitation. The model evaluation is performed for the 23 subcontinental regions listed in Table 1 introduced by Giorgi and Francisco (2000). Here we examine extremes over regional scales that the model can adequately treat. For more local events the model output would first need to be downsampled, as was the case in the study of the U.K. flood by Pall et al. (2011). In this section, we apply our evaluation methodologies to all the 23 regions and employ some general definitions of warm, cold, and wet events to assess model reliability over a wide range of climatic regimes and different types of events. We decided not to illustrate results only for a subset of the regions, given that the model performance may vary between regions, even if they belong to the same climate zone, and to avoid excluding regions that some readers may have a particular interest in. For the attribution of specific extreme events, like the case studies presented in section 4, the model evaluation will focus on the region and type of event under investigation. The model is validated against data from the NCEP–NCAR reanalysis (Kalnay et al. 1996). Reanalysis data have the advantage of full spatial and high temporal resolution that allows us to easily assess any region around the world, which would otherwise be difficult, especially for precipitation. Although the use of real observations would be preferable, at this stage we aim to establish the broad aspects of the model skill and for that purpose the reanalysis data would suffice. It is recommended that future studies on specific events should compare the model against high-quality observational sets relevant to the events under consideration. For the model evaluation related to temperature extremes we will also employ the Climate Research Unit (CRU) temperature product version 3 (CRUTEM3) dataset (Brohan et al. 2006) as an additional independent test on the model results.

a. Reliability diagrams

Reliability diagrams are widely used in the evaluation of ensembles of probabilistic forecasts (Wilks 1995). They are constructed by plotting the observed frequency of an event against the forecast probability. The model skill is assessed by the proximity of the resulting curve to the diagonal: the higher the model skill, the closer the curve is to the diagonal. Here we investigate how well HadGEM3-A can reproduce high and low seasonal mean temperatures and high seasonal mean precipitation rates in the 23 Giorgi regions detailed in Table 1. The thresholds that define the events are set to the upper or lower terciles of the 1971–2000 climatology estimated using the NCEP–NCAR reanalysis.

Details on how the reliability diagrams are constructed are given next. Each of the five 51-yr-long (1960–2010) runs used for model evaluation provides 51 seasonal mean temperature and precipitation values for each season: that is, 204 realizations to verify for all the seasons in each region. We compute the forecast probability by checking how many of the five estimates (from the five ensemble runs) of each hindcast fall above the threshold (or below in the case of cold temperatures). If, for example, two of the five ensemble simulations exceed the threshold for a given year and season, then the forecast probability is 20%–40%. In this way we group the 204 hindcasts for each region into five probability bins of equal size (0%–20%, 20%–40%, ..., 80%–100%). We then check whether the threshold for each year and each season was actually exceeded or not according to the reanalysis and calculate the number of events that are observed and the number of events that are not observed (nonevents) for the hindcasts in each probability bin. The observed frequency is estimated by the number of events divided by the sum of events and nonevents. Here we assume the model skill does not depend on the time of the year when the event occurs and therefore construct diagrams based on all the seasons in a year, which has the advantage of increasing the hindcast sample from 51 to 204. It would also be possible
to discriminate between seasons, though with smaller samples it is more likely than some probability bins may be underrepresented or even unpopulated. In section 4 we will investigate specific extreme events and will then construct reliability diagrams relevant to the period in which the events developed.

Figures 1–3 illustrate the reliability diagrams for high and low temperatures and high-precipitation events, respectively. The panels in each figure correspond to different Giorgi regions and the inset histogram in each panel is the “sharpness diagram,” which shows how many hindcasts go into each probability bin. Estimates of the observed frequency in probability bins with a small number of hindcasts are less reliable, while there may also be some bins that are not populated. The reliability diagrams for high-temperature events (Fig. 1) indicate that in a large number of regions the model has good skill in reproducing the event. There are, however, some regions like SSA associated with a low reliability. For such regions we need to test whether the model reproduces the distribution of regional seasonal temperatures well to decide whether an attribution study related to hot extremes in the region would be reliable. There are also several regions for which the diagrams for cold temperatures indicate good model reliability (Fig. 2), though these are fewer compared to warm events. The predictive skill for high precipitation rates (Fig. 3) is notably lower than for warm and cold temperatures.

Regions in the tropics (plotted in red) are generally associated with a higher reliability than regions in the extratropics. We also constructed the reliability diagrams with CRUTEM3 for temperature-related events. Figure 4 compares the results obtained with reanalysis and CRUTEM3 for the case of high temperatures. Despite some differences, the results are generally consistent (the same is the case for cold-temperature events) and the majority of regions with high reliability that can be identified with the two validation datasets are the same.

b. Evaluation of model statistics

We next carry out some simple tests to assess whether the model statistics of warm, cold, and wet daily extremes are well modeled. We use three simple indices that describe these three types of events: namely, the maximum and minimum daily mean temperature and the maximum daily mean precipitation rate of each year during the period covered by the runs used for model evaluation (1960–2010). Time series of the warm day index from the five runs and the reanalysis in the 23 Giorgi regions are illustrated in Fig. 5. Note that the temperatures are reported as anomalies relative to period mean. While the reanalysis time series are in most cases within the range of the model time series, there are regions, most notably NAU, where the model interannual variability is much higher. The probability density function (PDF) of the warm, cold, and wet extreme indices from the model and the reanalysis over the reference period are shown in Figs. 6–8, respectively, and the power spectra (Stouffer et al. 2000; Gillett et al. 2000; Christidis et al. 2012) of the index time series are shown in Figs. 9–11. The NAU panels in Figs. 6 and 9 corresponding to the warm day extremes illustrate in a different way that the model may have little ability to reproduce the statistics of warm events in the region, as it produces a much wider index distribution and much higher variability across different time scales compared to the reanalysis. The model, however, seems to have better skill with regard to cold and wet events in the region. Investigating the reasons why the model skill differs in different regions and for different types of events is beyond the scope of this work, as we here only try to formalize and give general examples of a model evaluation approach that should be an integral part of any attribution analysis. Considering the bulk of the regions examined here, the model in the majority of cases seems to adequately represent the statistics of warm, cold, and wet events.

Simple model evaluation tests are valuable guides that need to precede any attribution assessment to establish the credentials of the model employed. These become particularly useful in regions where the model has low skill in reproducing the event in question, as indicated, for example, by the reliability diagram. Provided that the model distribution of the relevant climatic indicator is consistent with the observed distribution, useful inferences can still be made for events in the tails. Here we have examined the overall distribution over several decades, though in a changing climate the distribution is changing with time. We therefore also checked the model PDFs of the last decade (2001–10) and compared them against the index range from the reanalysis in the same period. The model PDFs and the reanalysis range (not shown here) are again found to be consistent in most regions but, because of the small sample, this test provides only a rough indication of the model ability to represent the statistics of extreme events in recent years. Finally, considering the model evaluation plots for NEU shown in this section, our model appears to be suitable for an attribution assessment of the 2000 autumn floods in the United Kingdom examined in Pall et al. (2011). Although the reliability diagram for high-precipitation events (Fig. 3) indicates the event is not reproducible, the simple tests illustrated in Figs. 8 and 11 suggest that the model statistics for high-precipitation events in the northern European region are consistent with the
FIG. 1. Reliability diagrams for high-temperature events in 23 Giorgi regions. Tropical, midlatitude, and polar regions are plotted in red, green, and blue, respectively. The events are defined as seasonal mean temperatures above the upper tercile of the 1971–2000 climatology computed with the NCEP–NCAR reanalysis. The diagrams are constructed using 204 hindcasts provided by each of five ensemble simulations. The hindcasts are classified into five probability bins that define the forecast probability. The inset histograms show the number of hindcasts in each probability bin.
FIG. 2. As in Fig. 1, but for low temperature and with events defined as seasonal mean temperatures below the lower tercile.
Fig. 3. As in Fig. 1, but for high precipitation and with events defined as seasonal mean precipitation rates above the upper tercile.
Fig. 4. As in Fig. 1, but with the addition of the CRUTEM3 results (dashed lines).
Fig. 5. Time series of the temperature of the warmest day in the year during 1960–2010 in 23 Giorgi regions. The temperatures are given as anomalies relative to the mean of the entire period. Time series from five model simulations are plotted in gray and time series from the NCEP–NCAR reanalysis are plotted in black.
Fig. 6. Distributions of the temperature of the warmest day in the year during 1960–2010 in 23 Giorgi regions. The temperatures are given as anomalies and are relative to the mean of the entire period. The distributions plotted in blue come from five model simulations and the distributions plotted in red come from the NCEP–NCAR reanalysis.
FIG. 7. As in Fig. 6, but for the coldest day.
FIG. 8. As in Fig. 6, but for the precipitation rate of the wettest day.
FIG. 9. Power spectra corresponding to the time series of the temperature of the warmest day of the year during 1960–2010 in 23 Giorgi regions (plotted in Fig. 5). The spectra plotted in gray come from five model simulations and the spectra plotted in black come from the NCEP–NCAR reanalysis.
FIG. 10. As in Fig. 9, but for the coldest day.
FIG. 11. As in Fig. 9, but for the precipitation rate of the wettest day.
reanalysis. Provided the quality of the reanalysis in the region is good, we might then expect an assessment of the changing frequency of a U.K. autumn flood like the one in 2000 to provide useful information. It should be stressed that this is an indication based on the model evaluation relating to northern European high-precipitation events. A more focused investigation would consider the precise region affected and the period in which the event developed. In the next section we will present such attribution assessments for high-impact events in year 2010.

Given the evaluation assessments presented in this section, we can now set criteria for the model adequacy for reliable attribution analyses. First, the model reliability for the type of event under consideration in the region of interest is inferred using the reliability diagram. If the reliability turns out to be high, then attribution can be carried out. If, however, the model persistently reproduces climatological probabilities for the event in question, then attribution may only be carried out provided a second assessment of the model ability to reproduce the observed statistics of the same type of extremes in the region of interest is successful. This is done using the probability distributions and power spectra as illustrated in this section. Attribution is possible in this case because the model can still reliably reproduce the frequency of extremes, but it does this by producing climatology, which is by definition reliable. However, if the model fails both criteria, then an attribution analysis should be avoided as all the evaluation assessments indicate the model is an inadequate tool for that purpose.

4. The ACE 2010 experiment

a. Experimental setup

The first application of the new system focuses on extreme events in year 2010. This is part of a common international experimental activity agreed among a number of research groups working on the Attribution of Climate-Related Events (ACE) that recognize the need of enhanced coordination and collaboration in order to provide the best attribution systems. Here we present some first results obtained with the HaGEM3-A-based system and refer to this study as the ACE 2010 experiment. ACE 2010 encompasses 16-month-long simulations from September 2009 to December 2010 and aims to estimate changes in the frequency of extreme events that occurred during this period under the influence of anthropogenic forcings. The experiment comprises an ensemble of 100 simulations that represents the actual climate and three 100-member ensembles that provide three different realizations of the alternative climate without human influences. An important challenge in ACE 2010 is the representation of this alternative climate and, as in Pall et al. (2011), we use different estimates of the prescribed SSTs and sea ice in the simulations of the natural world to examine how sensitive the results are to the boundary conditions employed in the natural climate experiment.

As mentioned in section 2, we use model estimates of the underlying SST change ($\Delta$SST) due to human influence, which we subtract from the HadISST data for the reference period to compute the SSTs in the natural world. Here we employ three estimates of $\Delta$SST from simulations performed with three different coupled GCMs: namely, HadGEM1 (Stott et al. 2006), the third climate configuration of the Met Office Unified Model (HadCM3; Johns et al. 2002), and the HadGEM2 earth system model (HadGEM2-ES; Jones et al. 2011). We use the mean of a small ensemble of simulations with historical anthropogenic forcings extended into the twenty-first century (typically 3–4 runs) to get estimates of the 2000–09 monthly mean SSTs and subtract the mean of several hundreds of years from a control simulation with the same model for each month. We use the resulting $\Delta$SST to estimate the monthly mean SSTs for each month in the analysis period and also to adjust the sea ice using the simple empirical relationships discussed in section 2. The outcome is monthly SST and sea ice values that we prescribe in the HadGEM3-A simulations for the natural climate. The model interpolates between the monthly values get estimates between the months. At the time of this analysis, HadGEM2-ES simulations with anthropogenic forcings only were not yet available and for this model the 2000–09 $\Delta$SST was derived from simulations with all forcings and natural forcings only (as well as the long control run), assuming the anthropogenic effect can be estimated by the difference between the all and natural forcings runs. The three versions of the $\Delta$SSTs are illustrated in Fig. 12 for four different months in the year. There is an overall warming of the ocean that is greater in northern polar regions during boreal winter and autumn months. It has been suggested that the lack of warming in the summer is due to the fact that the incoming solar radiation has the effect of melting the sea ice rather than increasing the temperature (Ingram 2006), whereas the winter warming is amplified by a near-surface inversion layer that directs the infrared radiation downward instead of allowing it to go out to space (Bintanja et al. 2011). The $\Delta$SST patterns from three models are broadly similar, but there are model-dependent features like the North Atlantic cooling off the west and east coasts of North America in the HadGEM2-ES patterns that is, for
example, absent in the HadCM3 patterns. It is important to investigate whether such differences in the prescribed SSTs have an impact on the attribution results.

b. The three case studies

To demonstrate the capabilities of the new event attribution system we will now examine three high-profile events and estimate the change in their frequency due to the effect of anthropogenic influences on the climate. The first event is the cold spell in the United Kingdom during December 2009–January 2010, characterized by temperature anomalies over the country that were 2–2.5 K below the 1971–2000 mean and significant snowfalls across the United Kingdom that caused airport closures and other major disruptions. A persistent pattern of northwesterly winds during December–January brought cold and moist air from the Arctic resulting in snowy weather. The mean 500-hPa geopotential height and winds during the two coldest months of the season are plotted in Fig. 13a using NCEP–NCAR reanalysis data. The second event we study is the heat wave in Moscow in July 2010 marked by record temperatures in the region and severe stress on human health, exacerbated by smog from widespread forest fires. The event is associated with a quasi-stationary anticyclonic circulation, illustrated in the July mean geopotential height map from the reanalysis (Fig. 13b). Finally, we also investigate the catastrophic floods in Pakistan during July 2010, which led to loss of life and destruction of property and set off an agricultural crisis. The event was associated with an active period of the summer monsoon (Hoyos and Webster 2007) with heavy rainfall extending from the Bay of Bengal across the plains of northern India into the affected region. The characteristic circulation pattern is illustrated in the July mean geopotential height map in Fig. 13c. The patterns for the three events that correspond to the ensemble mean from the 100 simulations of the actual climate are not similar to the reanalysis patterns but show a mainly zonal flow over the United Kingdom and the Moscow area (Figs. 13d,e) and also a shift of the low pressure center in the Asian monsoon–affected region to the northeast of India (Fig. 13f). There are, however, individual ensemble members that give patterns similar to the reanalysis, implying that the model is able to reproduce the synoptic conditions for each event. Examples of such ensemble members are shown in Figs. 13g–i. The three events examined here developed over several weeks and affected large regions and we therefore suggest there is no need to downscale the model output in our analyses. Regarding the floods in Pakistan, we here examine the broad characteristics of the event as described by the monthly rainfall in the
region. A more detailed study would require the use of a hydrological model, as in Pall et al. (2011).

c. Model evaluation

In section 3 we introduced different model evaluation methodologies and demonstrated their application over a number of regions using seasonal events or extreme indicators. Here we apply the same methodologies to the regions of the three case studies considering the relevant variable for each case (temperature or precipitation) in the period over which each event evolved. We first assess the model skill in reproducing (i) a cold event during December 2009–January 2010 in the United Kingdom, (ii) a hot event during July 2010 in the Moscow region, and (iii) a wet event during July 2010 in the flood-affected region of Pakistan. The three regions in our analysis are located at 49°–59°N, 8°W–3°E (United Kingdom); 45°–65°N, 30°–50°E (Moscow); and 25°–36°N, 66°–74°E (Pakistan). Cold events are again characterized by temperatures below the lower tercile of the 1971–2000 climatology, while warm or wet events are characterized by temperatures or rainfall rates above the upper climatological tercile. The reliability diagrams for the three events are shown in Fig. 14. For temperature-related events we show diagrams produced with both the reanalysis (orange line) and CRUTEM3 (red line). Note that, when using CRUTEM3, the model data are masked to have the same coverage as the observations. The reliability diagrams are based on 51 hindcasts per model run of the December–January mean U.K. temperature, the July mean temperature in Moscow, and the July mean precipitation in Pakistan during 1960–2010. The diagrams suggest that the model is not able to reliably reproduce the events in the United

![Fig. 13. Maps of the 500-hPa pressure (red contours, m) and wind (blue vectors, m s⁻¹) in (a), (d), (g) the U.K.–North Atlantic region in December 2009–January 2010, (b), (e), (h) the region of western Russia in July 2010, and (c), (f), (i) a region in South Asia in July 2010. Maps from (top) NCEP–NCAR reanalysis, (middle) the ensemble mean of the 100 simulations for the actual climate, and (bottom) individual runs that resemble the reanalysis.](image-url)
Kingdom and Pakistan but seems to have more reliability in reproducing warm events in the Moscow region, though the high probability bins are less populated and hence more uncertain. In other words, we have more confidence that the attribution system can estimate the changing frequency of the actual heat wave in Moscow under human influence, but we need to assess further whether it can also do so for events similar to the U.K. cold winter and the floods in Pakistan by looking at the model distributions and variability for this kind of events.

![Reliability diagrams](image-url)

**FIG. 14.** Reliability diagrams for (a) high December–January temperatures in the United Kingdom, (b) high July temperatures in the Moscow region, and (c) high precipitation rates in the region of Pakistan. The events are defined relative to terciles of the 1971–2000 climatology computed with the NCEP–NCAR reanalysis (black lines) and CRUTEM3 (gray lines). The diagrams are constructed using 51 hindcasts provided by each of five ensemble simulations. The hindcasts are classified into five probability bins that define the forecast probability. The inset histograms show the number of hindcasts in each probability bin.
A model comparison against CRUTEM3 temperature observations and NCEP–NCAR precipitation reanalysis data is illustrated in Fig. 15. The top panels present the time series of the mean December–January U.K. temperature (Fig. 15a), the mean July temperature in Moscow (Fig. 15b), and the mean July precipitation in Pakistan (Fig. 15c) from the five runs used for model evaluation and the observations or reanalysis data. The corresponding distributions are shown in Figs. 15d–f and the power spectra are shown in Figs. 15g–i. The last year in the time series shows that the 2-month winter period in 2009/10 is one of the five coldest in the United Kingdom since 1960, while July 2010 is the hottest in the Moscow region. The overall July 2010 precipitation in Pakistan does not seem to be exceptionally high, which, however, does not rule out short events of intense precipitation that can lead to flooding. The model gives a realistic representation of Moscow temperature distribution (Fig. 15e). The U.K. distribution is also consistent with the observations, though arguably narrower (Fig. 15b) and the power spectrum of the observed time series lies mostly within the range of the model spectra (Fig. 15c). However, HadGEM3-A produces a much narrower distribution of the rainfall in Pakistan than the reanalysis suggests (Fig. 15h). This is also evident in the corresponding power spectrum, which illustrates that the model variability is considerably lower than the reanalysis (Fig. 15i). We also examined the model distributions and variability for three simple indices of extreme events: namely, the warmest, the coldest, and the wettest day of the year. Comparison against the observations and reanalysis (not shown here) confirms that the statistical characteristics of the U.K. and Moscow events are well represented by the model, but this is not the case for extreme rainfall in Pakistan. In conclusion, we have confidence that an attribution
assessment of the Moscow heat wave with the new system would be robust and, given the predictive skill of the model for warm events in the region, the changing odds of the actual event under anthropogenic influences would be reliably estimated. In the case of the United Kingdom, the predictive skill is lower but, as the model reproduces well the winter temperature distribution and variability in the region, we can still get a useful estimate of the changing frequency of cold winters like the one in 2009/10, although this would not be an attribution of the event itself. Finally, the model evaluation results clearly indicate that the model can neither forecast well heavy rainfall events in the region of Pakistan nor accurately reproduce the rainfall distribution in the region and would therefore be an unreliable tool for an attribution assessment of the 2010 floods. This stresses the importance of model validation, which helps identify the strengths and limitations of the new system.

d. Attribution

We now investigate how anthropogenic influences on the climate have changed the likelihood of occurrence of the three events. With regard to the floods in Pakistan the model fails the criteria we set at the end of section 3 and an attribution analysis of this particular event should therefore be avoided. We nevertheless present results for this case study too, only to demonstrate the contradictory conclusions an attribution analysis may lead to, when rigorous model evaluation is not carried out, or its outcome is not taken into account. We construct the PDFs of the December 2009–January 2010 mean temperature in the United Kingdom (Figs. 16a–c), the July 2010 mean temperature in Moscow (Figs. 16d–f),
and the July 2010 mean precipitation in Pakistan (Figs. 16g–i) using the ensemble runs for the climate with and without the effect of anthropogenic forcings. For each event we have three different PDFs for the natural world, represented by the green histograms in separate panels, but a single PDF for the actual climate represented by the orange histograms. As the runs cover only the short period from September 2009 to December 2010, we cannot express the temperature and precipitation as anomalies relative to a climatological period like we did with the long runs used for model evaluation. Actual values of temperature and precipitation, unlike the anomalies, may have a model bias, which we correct following the simple approach employed in seasonal forecasting based on the difference between the model and reanalysis climatologies. For example, for the Moscow event, we calculate the 1960–2010 July mean actual temperature in the region using the reanalysis and the model evaluation runs. The correction factor comes from the difference between the mean of the five validation runs and the reanalysis and is then applied to the July temperature estimates from both the actual and natural world ensembles, assuming that the bias would remain unaltered without the effect of anthropogenic forcings. The reanalysis temperature and precipitation values for the events under consideration are also shown in Fig. 16. In our analysis we base our best estimate of the observed temperatures on reanalysis data, as the CRUTEM3 temperatures, available as anomalies, are not directly comparable with the actual temperatures from our short runs.

The distributions in Fig. 16 indicate a shift toward higher temperatures under the effect of human influences on the climate in both the U.K. and Moscow regions. Consequently, the cold winter months in the United Kingdom would have been more common in the hypothetical natural world, whereas the Moscow July heat wave would have been less common. However, as illustrated in Figs. 16d–f, the observed Moscow temperature lies in the far tail of both distributions, indicative of a very extreme and uncommon event with or without the anthropogenic effect. This is in agreement with the study of Dole et al. (2011), who demonstrated using both model and observational data that human influences do not seem to have contributed substantially to the magnitude of the heat wave. On the other hand, Rahmstorf and Coumou (2011) applied a simple statistical model to observational data and estimated that climate warming made the July 2010 record temperature in Moscow 80% more likely. The shift of the temperature distribution in Figs. 16d–f also suggests that the event is more likely in the actual climate but, since it is located in the far tail of the distribution, it remains extremely uncommon even in the warmer world. The distributions of rainfall in Pakistan (Figs. 16g–i) do not offer any conclusive results. The three representations of the natural climate give three contradictory indications: runs with ΔSSTs from HadGEM1 suggest July 2010 becomes slightly drier under human influences, whereas runs with ΔSSTs from HadGEM2 suggest the opposite and runs with ΔSSTs from HadCM3 show no notable change in the rainfall distribution. It should be reminded that the model validation discussed in the previous section indicates that we do not expect to get a reliable attribution assessment for this particular event.

We next calculate the probabilities $P_0$ and $P_1$ of exceeding (or, for the United Kingdom, going below) the threshold that defines each event in the natural and actual climate, respectively. The threshold is again estimated from the NCEP–NCAR reanalysis. The estimated probabilities are illustrated in Fig. 17. For each event, we get three estimates of $P_0$ for each of the three realizations of the natural world (colored bars), as well as an estimate that corresponds to the aggregate of all the temperature or precipitation estimates from the three ensembles (black dashed bar). As in Pall et al. (2011), we estimate the uncertainty in the probability estimates due to sampling using a simple Monte Carlo bootstrap procedure. More specifically, we randomly resample the 100 estimates from each ensemble 10 000 times and thus compute 10 000 estimates of the probabilities that give the best estimate (50th percentile) and the 5%–95% range shown in Fig. 17. A winter as cold as or colder than the one in 2010 in the United Kingdom (Fig. 17a) becomes an order of magnitude less likely under anthropogenic influences on the climate and its return time decreases from 4–10 yr to 15–450 yr. The Moscow heat wave is associated with very low probabilities (Fig. 17b), which are near zero in the climate without anthropogenic influences. In the actual world the return time of heat wave events in the region at least as severe as the one in 2010 is of the order of several hundred years. The probabilities associated with the flooding event in Pakistan (Fig. 17c) show disparity between the different realizations of the natural climate and provide no certain indication of whether this type of events has become more or less common because of human influences on the climate.

Using the estimated distributions of $P_0$ and $P_1$ we construct the distribution of the FAR shown in Fig. 18. We only plot results for the U.K. and Pakistan events, as the FAR for the Moscow heat wave is unity (the probability $P_0$ is almost zero as shown in Fig. 17b), indicating the event is unlikely to occur without the effect of anthropogenic forcings in agreement with Rahmstorf and
The FAR distributions for the United Kingdom (Fig. 18a) obtained with different realizations of the climate without human influences are in good agreement, which suggests the results are not very sensitive to the estimate of the SSTs prescribed in the simulations. In all cases the FAR is negative: that is, the event would be more common without the anthropogenic effect. Considering the 50th percentile of the FAR distribution, winters as extreme as 2010 (or more) would be 7–10 times more likely in the natural world, though of course there is uncertainty associated with this best estimate as shown in Fig. 18a. The FAR distributions for the floods in Pakistan (Fig. 18b) that we get from the three versions of the natural world are rather different, with one distribution lying on positive values, one on negative values, and one being centered at around zero.

In conclusion, our analysis suggests that (i) cold winters in the United Kingdom like the one in 2010 have become much rarer (the return time has increased by an order of magnitude) as a result of the anthropogenic effect on the climate, (ii) the Moscow heat wave of 2010 is a very extraordinary event, yet more common than in the absence of anthropogenic forcings, and (iii) it is not
possible to get a reliable attribution assessment of the 2010 floods in Pakistan using the new attribution system.

5. Discussion

We have developed a new system for the attribution of weather and climate extreme events and provided initial assessments for three high-impact events that occurred in the year 2010. While it is straightforward to estimate the changing frequency of events with the new system, it is essential that we first establish whether the model can adequately represent the event under consideration or its statistical properties. We showed, for example, that an attribution assessment of the floods in Pakistan would be unreliable as the model does not reliably reproduce this type of event in the region. For events associated with a better model reliability, the results did not show great sensitivity to the prescribed SSTs in the simulations of the natural climate, a key uncertainty in the methodology. In future assessments a wider range of estimates from different models can be used to explore this uncertainty in more detail. Better estimates of the SST change can also be constructed using optimal fingerprinting, which would scale model trends in the SST to match them best to the observations. The simple linear model that provides the estimates of the sea ice change also needs to be replaced by a more sophisticated and accurate scheme in future work. Moreover, it is important that in future ACE studies the model evaluation will be performed against actual observations rather than reanalyses. As observational data often come in form of anomalies relative to a climatological mean, we aim to generate long multi-decadal ensembles for the climate with and without human influences that will enable us to estimate anomalies for the model data and will remove the need of bias correction. Finally, while in this paper we concentrate on the cumulative anthropogenic effect on extreme events, it is possible to single out the effect of an individual forcing in future analyses and also feed the model output into impact models for studies of specific socioeconomic impacts.

The ACE 2010 experiment provides a valuable benchmark for the evaluation of attribution systems developed by different research centers. Comparison of the results obtained with different models is essential in better establishing the credentials of the methodology and the quality of the resulting attribution information. Such a comparison study will be conducted once the experiments completed. It is likely that this kind of multimodel studies may not become possible until the completion of a future coordinated experiment [Climate of the 20th Century (C20C)] that is currently at the planning stage and will supersede ACE 2010. It is envisaged that C20C will be a much longer experiment with simulations that cover several decades (e.g., the post-1950 period). Such an experiment, though computationally more expensive, would enable us to examine a wealth of events of different types in any region and to assess how changes in the frequency of extremes are evolving with time in a changing climate. Finally, an annual attribution study of main extreme events that have occurred in the year is also planned to be carried out, as part of a wider-scope project aimed to provide a comprehensive summary of the year’s climate. A new experiment with the attribution system presented in this paper has already contributed a study on the cold winter of 2010/11 in the United Kingdom to the summary of the 2011 climate [see “Lengthened odds of the cold U.K. winter of 2010/11 attributable to human influence” in Peterson et al. (2012)].

A quick response from the scientific community is strongly desired in the aftermath of a catastrophic event when there is a pressing demand for attribution-related information. Integrating the new system into an operational framework that runs in near–real time would provide an efficient way of delivering the required attribution information in a timely manner. Moreover, by taking advantage of the SST forecasts provided by the Hadley Centre decadal prediction system (Smith et al. 2007), it would also be possible to extend the model simulations into the future and thus proactively provide attribution information as soon as new events occur. The first system to issue weather risk attribution forecasts operationally was developed in the University of Cape Town based on the older Hadley Centre model HadAM3P (D. A. Stone 2011, personal communication). Given the long-term goal to provide robust operational systems, our work demonstrates the importance of supplementing attribution results with valuable validation assessments indicative of the model skill. Once confidence in the results has been established, the system becomes a valuable tool that can be used to inform policy and decision making.

Acknowledgments. We are grateful to the reviewers for their constructive comments. This work or was supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101).

REFERENCES


——, ——, and ——, 2012: The contribution of anthropogenic forcings to regional changes in temperature during the last decade. Climate Dyn., 39, 1259–1274.


