THEME SPA.2013.1.1-05

EUCLEIA

(Grant Agreement 607085)

EUropean CLimate and weather Events: Interpretation and Attribution

Deliverable D6.*5*

Description of reliability assessment methods

Deliverable Title	*Description of reliability assessment methods*	
Brief Description	*Discussion on model evaluation methods suitable for event attribution studies. Comparison between observed and simulated statistical characteristics of extremes and conditions pertinent to extreme events. Reliability diagrams. The effect of forecast reliability on the changing risk of extremes. Estimating the uncertainty in the Fraction of Attributable Risk (FAR).*	
WP number	6	
Lead Beneficiary	*Nikolaos Christidis, Met Office*	
Contributors	*Nikolaos Christidis, Met Office; Fraser Lott, Met Office; Omar Bellprat, IC3; Francisco Doblas-Reyes, IC3*	
Creation Date Version Number Version Date		
Deliverable Due Date Actual Delivery Date	June 30, 2015	
Nature of the Deliverable	×	*R - Report*
		P - Prototype
		D - Demonstrator
		O - Other
Dissemination Level/ Audience	×	*PU - Public*
		PP - Restricted to other programme participants, including the Commission services
		RE - Restricted to a group specified by the consortium, including the Commission services
		CO - Confidential, only for members of the consortium, including the Commission services

Version	**Date**	**Modified by**	**Comments**
0.1	06/05/2015	NC	Rough outline
0.2	22/06/2015	NC, FL, OB	First complete draft
0.3	29/06/2015	Internal Review	Final draft
0.4	30/06/2015	Project Manager	Final version

Table of Contents

List of Figures

References

1. Executive Summary

A range of methodologies are presented here for the evaluation of models used in event attribution studies. Evaluation assessments are essential to help understand the strengths and limitations of models, quantify the uncertainty in attribution statements and ensure the information provided by an attribution system is robust and trustworthy. This report provides details on the development of methodologies which assess different aspects of the model skill and illustrates their application.

- Atmospheric circulation patterns and modes of variability have in many cases been identified as a key factor that favours the occurrence of extreme events. While oceanic variability may be accurately represented in atmospheric model simulations through the prescribed boundary conditions, atmospheric variability and circulation patterns need to be examined independently. We illustrate how pressure fields from reanalysis datasets can be employed to test the model representation of these patterns and how the atmospheric flow may be used as a constraint in event studies.
- Multi-decadal simulations of the recent climate are compared against observational or reanalyses datasets to examine whether the statistical characteristics of relevant climatic variables are well captured by the model. A model that yields a good representation of the climatological distribution of these variables would be suitable for attribution, even if it had little skill in forecasting extreme events. A suite of tools can be employed to facilitate this kind of assessments, including formal statistical tests, extreme value theory, timeseries and power spectra analyses.
- The predictive skill of the models is often inferred from reliability diagrams which compare the modelled probability of events against their observed frequency. Using an idealised framework we show that forecast reliability systematically affects attribution studies. Unreliable systems are prone to overestimate the fraction of attributable risk (FAR) and further found to increase the uncertainty in FAR.
- A new way of constructing reliability diagrams has been developed within EUCLEIA that helps increase the sample from which the modelled probabilities are derived and reduce sensitivity to the grid box size.
- A systematic investigation of the error associated with the estimated change in the frequency of extremes due to human influence has been undertaken. Estimates of the changing frequency obtained from later and earlier periods of the observational record as a proxy for the present-day and the pre-industrial climate are found to provide a reliable basis for the assessment of the error.

In addition to the evaluation assessments mentioned so far, it should also be ensured that models can accurately represent key physical processes that influence the chances of extreme events. Such processes are investigated by several tasks of WP6 and will be discussed in future deliverables. Application of methodologies discussed in this deliverable will be employed for the evaluation of the new high-resolution HadGEM3-A model developed in WP8 of EUCLEIA.

2. Project Objectives

With this deliverable, the project has contributed to the achievement of the following objectives (DOW, Section B1.1):

No.	Objective	Yes	No
1	Derive the requirements that targeted user groups (including regional stakeholders, re-insurance Companies, general public/media) have from attribution products and demonstrate the value to these users of the attribution products developed under EUCLEIA.	Indirectly, by providing evidence for the robustness of the attribution products	
2	Develop experimental designs and clear ways of framing attribution studies in such a way that attribution products provide a fair reflection of current evidence on attributable risk.		×
3	Develop the methodology for representing the level of confidence in attribution results so that attribution products can be trusted to inform decision making.	A range of methodologies has been developed and are discussed in detail in this deliverable.	
4	Demonstrate the utility of the attribution system on a set of test cases of European weather extremes.	Contributes to this objective by providing ways to assess the strengths and limitations of the attribution system	
5	Produce traceable and consistent attribution assessments on European climate and weather extremes on a range of timescales; on a fast-track basis in the immediate aftermath of extreme events, on a seasonal basis to our stakeholder groups, and annually to the BAMS attribution supplement.		×

3. Detailed Report

3.1 Evaluation of event attribution systems

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., 2013). Attribution systems developed to output regular assessments, e.g. in an operational framework, have been designed on the methodological approach introduced by Pall et al. (2011) in his study of the autumn 2000 floods in the United Kingdom. The approach is based on two distinct ensembles of simulations generated by an atmosphere-only model, one ensemble that represents the real world and includes all external forcings that act on the climate system and one that represents a counterfactual natural world without the effect of human influence on the climate. The two ensembles yield estimates of the probabilities P_1 and P_0 of the occurrence of an extreme event with and without the effect of human influence respectively, which in turn are used to calculate the Fraction of the Attributable Risk (FAR) defined as $1-(P_0/P_1)$. The development of an attribution system based on the HadGEM3-A model (Hewitt et al., 2011) is a large component of the attribution service that EUCLEIA aims to deliver, built on a pre-existing system (Christidis et al., 2013) that has been successfully employed in several studies of high-impact extreme events.

As attribution of extreme events relies largely on climate model simulations, providing evidence that the models employed in the study are fit for purpose would be essential in order to demonstrate the degree of confidence one can have in the results. A model is likely to have different skill in reproducing different types of extremes in different regions and therefore its evaluation needs to be tailored to the event under consideration. Model evaluation encompasses a range of different aspects:

a) Physical processes. Land-surface related processes may play a key role in the occurrence of extremes. For example, a large precipitation deficit together with strong positive radiative anomalies in the months preceding the extreme European summer heatwave of 2003 contributed to a rapid loss of soil moisture that provided favourable conditions for the heatwave event (Fischer et al., 2007). Similarly, a decrease in soil moisture and surface evapotranspiration under stable atmospheric conditions were crucial during the US summer heatwave of 2012 (Diffenbaugh and Scherer, 2013).There are three tasks within Work Package 6 of EUCLEIA that aim to identify key processes driving extremes, collect necessary observational data and use them in model evaluation assessments (tasks 6.1-6.3). This work is ongoing and will be presented in future deliverables (D6.1 - D6.3).
b) Atmospheric circulation and modes of variability. It is often the case that synoptic conditions are critical in the development of extreme events. The exceptionally high temperature anomalies during the 2003 European heatwave, for example, were initiated by a persistent anticyclonic circulation. Modes of climate variability like the El Niño Southern Oscillation (ENSO) may be similarly important, as they influence the chances of extremes. For example, La Niña conditions increase the likelihood of extreme precipitation in southeastern Australia (King et al., 2013), droughts during the short-rains season in East Africa (Lott et al., 2013) and flooding over Southern Africa (Bellprat et al., 2015). Models used in event attribution often employ prescribed

sea surface temperatures (SSTs) and therefore represent the actual observed modes of oceanic variability. It is essential that circulation patterns and atmospheric variability modes that influence the development of extremes are also reproduced with a realistic frequency in model simulations.

c) Statistical characteristics. Attribution statements are inferred from statistical analyses that typically involve estimates of the changing probabilities of extremes because of human influence, as expressed, for example, by the FAR. Model evaluation at its most basic level should therefore examine whether the simulated distributions from which the probabilities are derived compare well with the observations. The rare nature of extremes makes this a challenging task, due to the lack of long observational records, which may also lack good spatial coverage. Scientists often resort to considering more moderate rather than extreme events and/or employ reanalyses data in their evaluation assessments (Christidis et al., 2013).

Event attribution often borrows methodological tools from seasonal forecasting, developed for the evaluation of ensemble-based forecasts (Palmer et al., 2008), like the popular reliability diagram (Wilks, 2011). Reliability diagrams indicate whether the model is able to capture the predictable features of the event under consideration and plot its model frequency against the observed frequency of occurrence. The forecast probabilities are estimated from an ensemble of long multi-decadal simulations and the proximity of the plotted curve to the diagonal is a measure of the model skill. Good reliability implies that the model is able to forecast the event well, which is essential for seasonal forecasting. In cases with little, or no predictive skill, causal factors cannot be identified. Although this is a major hindrance in seasonal forecasting, it is not necessarily an impediment to attribution studies. The reason is that a model with no predictive skill may still robustly reproduce the statistical characteristics of extreme events similar to the one under consideration (i.e. their climatological probability of occurrence) and thus provide reliable estimates of the change in their frequency under the influence of anthropogenic forcings. In this more general case, an attribution analysis would still be possible with reference to events similar to the one of interest. Thus, the simulated probability of an extreme event would be a general estimate of the chance of occurrence that is not constrained by any predictive factors like, for example, the prevalent SSTs etc.

In this deliverable we summarise methodologies employed in model evaluation that can subsequently be applied to event attribution studies with the HadGEM3-A system developed in EUCLEIA. In section 3.2 we consider the model skill in reproducing circulation patterns important for extreme events and present an example of how this assessment can be made. The simulated accuracy of statistical characteristics of extremes is discussed in section 3.3. This evaluation can be done in advance over different regions and types of extremes, contributing to a better understanding of the strengths and limitations of the model. Section 3.4 explores the links between forecast reliability and attribution. An introduction to reliability diagrams and a novel way of constructing them is presented in section 3.5. Finally, in section 3.6 we develop a method of evaluating the FAR using physical models.

Deliverable 6.5 feeds directly into project objective 3, which is centred on the development of methodologies for the evaluation of attribution results. Additional work outlined in tasks 6.1 - 6.3 regarding key physical process for extremes and their representation by models will be the second main contributor to this project objective. Scientifically robust attribution products with well-quantified confidence levels become a trustworthy basis for decision-making.

Hence, the work discussed here helps ensure stakeholders will be provided with high quality products, which partly falls within the remit of project objective 1. Finally, model evaluation constitutes an integral part of any event attribution study, and the development of evaluation methodologies discussed in this report lays the groundwork for future analyses of European extremes delivered by project objective 4.

3.2 Evaluating the effect of circulation patterns and large scale influences

The synoptic circulation prevalent at the time of an extreme event may play a key role in its development and models need to be able to reproduce the same kind of circulation patterns with a realistic frequency to be suitable for an attribution study. The severe heatwave in Moscow in July 2010, for example, was associated with a quasi-stationary anticyclonic circulation, a flow pattern that was found to be well-reproduced in simulations of the recent climate with HadGEM3-A (Christidis et al., 2013). Large-scale oceanic influences that may also favour the occurrence of extremes, like ENSO, are inherent in analyses with atmospheric models through the prescribed oceanic state. This, however, is not the case for atmospheric modes of variability that could be similarly important for the occurrence of extremes, and the model skill in representing them needs to be assessed. In their study of the cold spring of 2013 in the UK Christidis et al. (2014) demonstrated the significant role of the negative North Atlantic Oscillation (NAO) phase during the event and found that it is reproduced in about half of the 600 simulations of year 2014 conducted with HadGEM3-A. Apart from ensuring that dominant synoptic patterns and modes of variability are well-represented in models, it is also common to introduce them as constraints into the analysis and frame attribution questions in such a way that the FAR is estimated for events occurring under these characteristic conditions. In the example of the cold spring in the UK, the effect of human influence was derived solely from simulated years with a negative NAO phase and a 30-60% decrease in the chances of having a cold spring was reported, even though the negative NAO favours its occurrence. Similarly, King et al. (2013) analysed the extreme rainfall over southeastern Australia in summer 2011/2012 in the context of the observed La Niña conditions, but in their case they found no apparent influence from anthropogenic climate change on the event. Constraining attribution results on the atmospheric circulation also forms the basis of the flow analogue methodology (e.g. Yiou and Cattiaux 2013; 2014), whereby simulated events most correlated with the observed sea level pressure (SLP) pattern are selected before an inference of the anthropogenic effect is made.

Using the winter storms of 2013/2014 in the UK as an example (Christidis et al., 2015), we next illustrate the application of a simple model evaluation methodology based on a comparison between model and reanalysis data. The event was characterised by an exceptional clustering of vigorous storms driven by the North Atlantic jet stream, which triggered tidal surges across coastal parts of the country, widespread floodplain inundations and pronounced river flows (Huntingford et al., 2014). Synoptic scale atmospheric circulation is commonly described by the pressure field at the surface (SLP), or the 500 hPa geopotential height (Z500). The latter describes better the general flow over the lower part of the troposphere where weather systems develop and is used in this example. Model Z500 fields are compared against data from the NCEP/NCAP reanalysis (Kalnay et al., 1996), as observational datasets are not readily available. The sensitivity to reanalysis data can be assessed by employing alternative reanalyses, though for a relatively smooth spatial field

like Z500 the choice of the reanalysis is not expected to be a major caveat. Christidis et al. (2015) analysed historical climate simulations produced with seven models that contributed data to the Coupled Model Intercomparison Project phase 5 (CMIP5; Taylor et al., 2012). Application of the evaluation methodology helped establish the adequacy of the selected models for an attribution analysis of the extreme winter UK rainfall in 2013/2014. The analysis was subsequently carried out by comparison of modelled winters in recent years extracted from simulations with and without anthropogenic forcings, which display a characteristic atmospheric flow over a wider UK region similar to the one observed.

The winter mean 500 hPa geopotential height and wind fields in the North Atlantic averaged over the climatological period 1960-1990 and estimated with reanalysis data and the ensemble mean of 43 simulations with all external forcings are shown in Fig. 1a and 1b respectively. The models produce a westerly flow over the UK very similar to the reanalysis. Fig. 1c illustrates the 2013/2014 anomaly relative to the climatology, estimated from NCEP/NCAR data, which shows a large scale cyclonic circulation pattern to the west of the UK associated with persistent south-westerly winds over the country. Similar patterns can also be identified in the model simulations. By selecting modelled winters in the last 20 years that correlate well with the 2013/2014 circulation anomalies over the UK region (correlation coefficients greater than 0.6), an ensemble was constructed to represent near present-day UK winters with a characteristic south-westerly flow. The ensemble mean (Fig. 1d) displays the characteristic flow pattern of 2013/2014. Although it is often implicitly assumed that such synoptic patterns are expressions of unforced variability, it should be examined whether they might be affected by climate change and thus become more or less frequent due to external climatic forcings. Fig. 1e depicts timeseries of the frequency of the 2013/2014 pattern over the UK since 1900. The frequency was estimated using 5-year running means, as the number of winters in each 5-year window that resemble 2013/2014 per total number of winters in the same time window. Testing the hypothesis that the least-square fit has a zero slope indicates that the trend in the pattern frequency is significantly different from zero, consistent with the model study of Hoerling et al. (2012). The trend, however, is small and the presence of high interdecadal variability means that longer records would be required to robustly identify any long-term change.

In the example presented in this section, model evaluation has not yet assessed the representation of UK winter rainfall by models and extreme rainfall in particular, but only the background conditions in which it occurs. As the conditions vary between events, this kind of evaluation exercise has so far been employed in detailed attribution analyses after the events occur. It would, however, be possible to also identify several characteristic regimes associated with different types of extremes in different regions and evaluate the model over pre-selected circulation patterns in a proactive manner. This would be useful in fast-track attribution assessments that become available on timescales of days after the event.

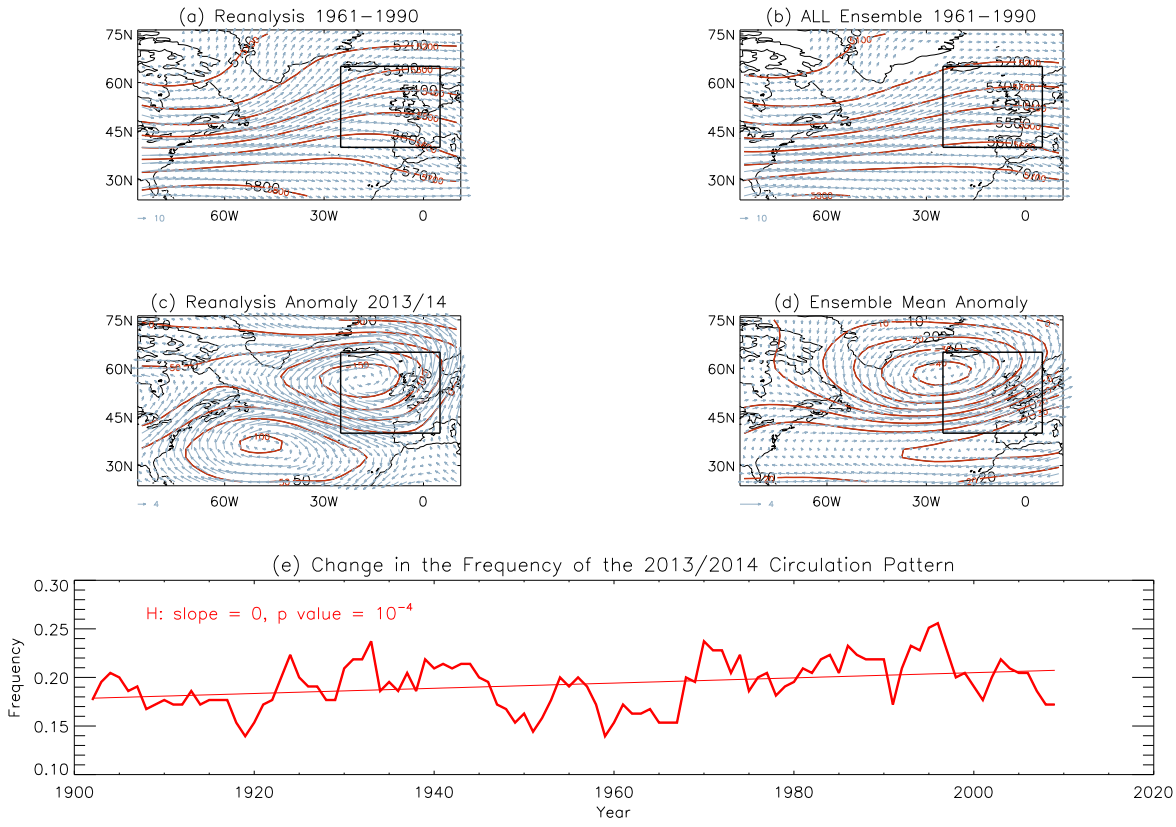

Figure 1. (a) Map of the winter mean 500 hPa height (red contours, m) and wind (blue vectors, m s^{-1}) in the North Atlantic during 1961-1990 produced with NCEP/NCAR reanalysis and (b) the mean of 43 simulations of the historical climate with seven CMIP5 models. The black box marks the wider UK region employed in the evaluation assessment. (c) Geopotential height and wind anomalies for winter 2013/2014 relative to the climatology constructed with NCEP/NCAR reanalysis data. (d) Geopotential height and wind anomalies relative to the climatology corresponding to the mean of simulated winters in the two most recent decades that correlate well (correlation coefficient greater than 0.6) with the reanalysis Z500 pattern over the UK region. (e) Timeseries of the frequency of winter circulation patterns similar to 2013/2014 from experiments with all historical forcings. The p-value refers to testing of the hypothesis that the least square fit has a zero trend.

3.3 Statistical characteristics

Although climate models may in some cases have little skill in forecasting extreme events, a minimum requirement for a model to be suitable for attribution would be that it is able to accurately represent the likelihood of occurrence of extremes in the present-day climate. Once this condition is met, it is then assumed that the model can also capture well the likelihood of extremes in a world without anthropogenic forcings. In attribution studies, a basic and informative assessment involves testing the simulated distribution of the climatic variable from which a probability estimate is derived (e.g. temperature for heatwave events) against the distribution from an independent observational or reanalysis dataset. In order for this assessment to be possible, a long multi-decadal simulation of the recent climate is required, or, better, an ensemble of such simulations. Event attribution experiments,

however, often span shorter periods centred on the event under consideration, as short simulations are computationally less expensive and so allow the generation of larger ensembles. In addition to the simulations conducted for various event attribution studies with the original attribution system based on HadGEM3-A, a small ensemble of simulations including all external forcings, covering the period 1960-2010 was also produced to enable necessary model evaluation assessments. A new ensemble of simulations is also to be produced with the high-resolution version of the model developed within EUCLEIA's WP8. More specifically, task 8.1 aims to deliver 15 simulations of the actual climate during 1960-2013, which will become the basis of detailed model evaluation by WP6 using, for example, methodologies described in this report.

Christidis et al. (2013) used the long simulations with HadGEM3-A to demonstrate how a systematic evaluation of the attribution model can be conducted in terms of the representation of temperature and rainfall across several sub-continental regions. Evaluation of the model statistics may include:

a) A comparison between modelled and observed timeseries of the relevant climatic variable over a climatological period. For a model to be suitable for attribution, it would be expected that the observational timeseries generally fall within the range of the simulated ones.
b) A comparison between modelled and observed distributions. The probability density function (PDFs) of the climatic variable over the simulated period can be constructed with observational and model data and a comparison can then be made. Apart from a qualitative visual comparison of the PDFs, a Kolmogorov-Smirnov test is also often employed to assess whether the two distributions are significantly different (Lewis et al., 2014; Christidis et al., 2015).
c) An examination of the tails of modelled and observed distributions. Apart from checking the level of agreement of the overall distribution estimated from models and observations, a closer inspection of the tails is also essential in studies of extremes. The lack of large samples means that statistical distributions from the extreme value theory need to be applied to the data to estimate probabilities of extremes.
d) Power spectra analyses that compare the variability in model simulations across different timescales against the observed variability (Gillett et al., 2000). As with the timeseries analysis, observed spectra should lie within the range of the simulated ones for models with an adequate representation of climate variability.

We again consider the example of the extreme winter rainfall in the UK to illustrate how model statistics can be evaluated. We look at the ability of HadGEM3-A in its original low resolution configuration to provide a realistic representation of rainfall in the region and its variability. Fig. 2a shows the winter rainfall timeseries over the UK region from the NCEP/NCAR reanalysis and 5 simulations with HadGEM3-A. The model range encompasses the reanalysis timeseries and the model variability also appears to be generally consistent with the reanalysis. The winter rainfall PDFs over period 1960-2010 are shown in Fig. 2b. The similarity of the two distributions is commonly assessed in a statistical way using, for example, the Kolmogorov-Smirnov test. The application of the test shows that the distributions in Fig. 2b are not significantly different (P-value=0.45). A more stringent assessment of the modelled variability than the simple inspection of the plotted timeseries (Fig. 2a) is can be made on the basis of power spectra. Fig. 2c shows the spectra computed with reanalysis data and with data from the model simulations. The modelled variability is

consistent with the reanalysis, albeit possibly higher at multi-decadal timescales. Finally, focussing on the heavy rainfall tail of the distributions, the return time of extreme events has been estimated with a generalised-Pareto distribution (Fig. 2d) and again the model is found to agree well with the reanalysis.

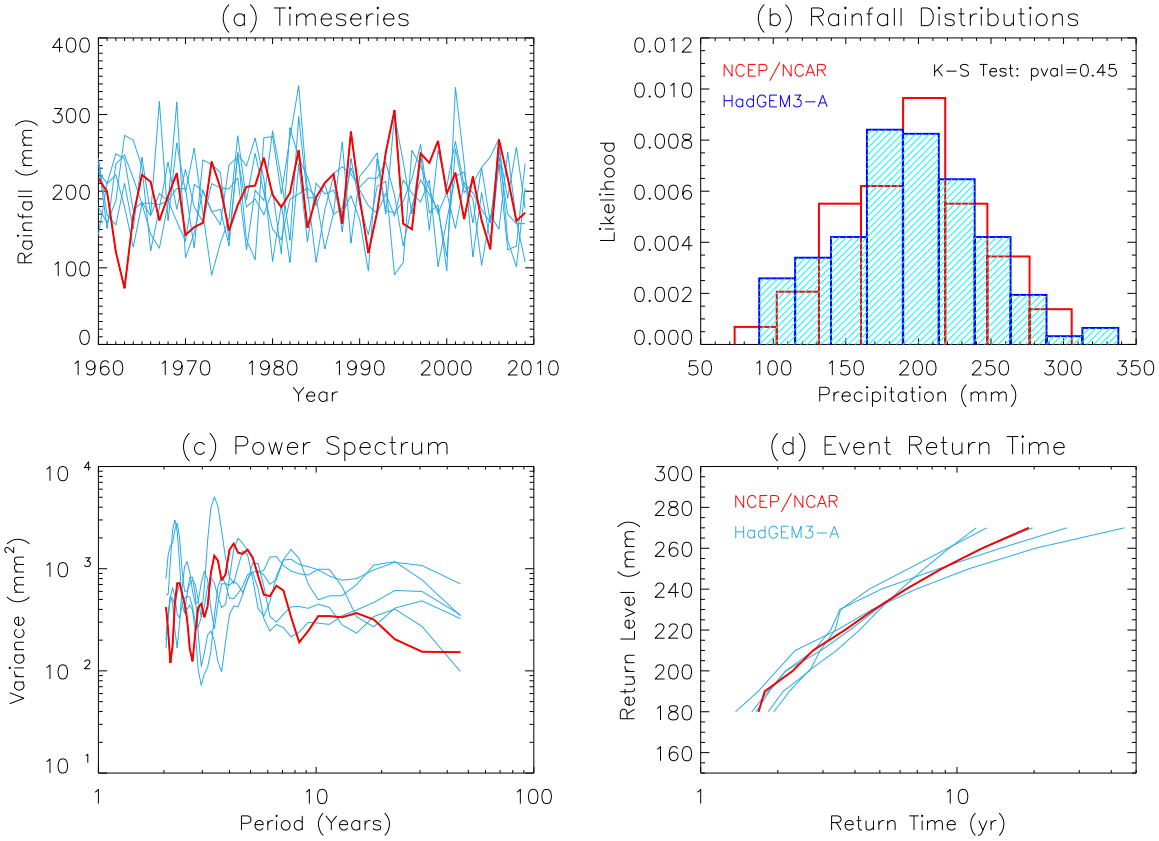

Figure 2. Evaluation assessments of HadGEM3-A to determine whether the model is suitable for attribution studies of extreme winter rainfall events in the UK. (a) Timeseries of winter precipitation in the UK region (b) Normalised distributions of the rainfall during 1960-2010. The p-value shown on the panel refers to a Kolmogorov-Smirnov test that assesses whether the distributions are distinguishable. (c) Power spectra corresponding to the timeseries shown in panel a. (d) Return time of winter rainfall events in the UK estimated with order statistics, or, for extreme thresholds, with the generalised Pareto distribution. Results shown in red correspond to the NCEP/NCAR reanalysis and results shown in blue to the HadGEM3-A simulations.

Evaluation of model statistics is a valuable guide that ought to precede any attribution assessment to establish the credentials of the model employed. This becomes particularly useful in regions where the model has low skill in reproducing the event in question, as indicated, for example, by the reliability diagram. For example, it has been shown that HadGEM3-A has no predictive skill with regard to heavy winter precipitation in Northern Europe (Christidis et al., 2013), However, the evaluation methodology presented in this section provides sufficient evidence that the simulated rainfall distribution, variability and climatological probability of extremes are realistic and the model would therefore be suitable for an attribution study of the 2013/2014 extreme rainfall event. The attribution results in that case would not be greatly constrained by the prescribed SSTs in the model simulations, as

these are found not to provide a strong predictive factor, but would express the change of probabilities of events similar to the one observed irrespective of the oceanic conditions. The following sections will explore links between forecasting skill and attribution in more detail.

3.4 Reliability diagrams

The reliability diagram, a model evaluation tool popular in seasonal forecasting, has also been applied to event attribution systems to assess the predictive skill of the model. A typical reliability diagram (Wilks, 2011) shows observed frequency against modelled/forecast probability. Points along the 1 to 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. 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, to increase the data sample size for these studies, a new area-pool method was devised to measure regional reliability.

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 standardised WMO (2002) technique, but this requires a large sample size, as well as there being an implicit assumption that the probability in each box is independent. Instead, the area-pool technique was 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 ensembles used at the different timescales of seasonal prediction or attribution. The set of grid boxes within the study region and season 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:

a) For the observations and for each model ensemble member, calculate the proportion of the region for which the event criterion is met.
b) 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.
c) 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 Fig. 3.

This method was applied to seasonal and decadal forecasts of wet seasons in African regions in Lott et al. (2014) as well as to attribution validation runs. This proved to be a useful analysis for forecasting, but at that point the link between reliability and attribution statistics, such as the fraction of attributable risk, was still not established. The subsequent investigations are detailed in the sections to follow.

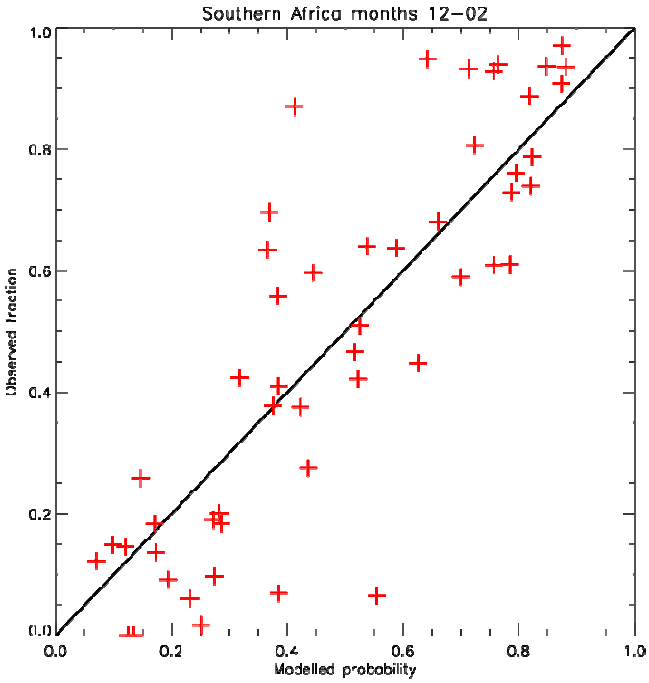

Figure 3. An unbinned reliability diagram created using the area-fraction technique, in this case for above-median Southern African wet-season (December-February) temperatures between 1960 and 2010, from multidecadal climate simulations in Lott et al. (2014).

3.5 Links between forecast reliability and the Fraction of Attributable Risk

Reliability describes the trust we put into forecasted probabilities that an event occurs. An event attribution statement is built on these probabilities too, in that it compares the probability of occurrence of an extreme event, both considering and excluding the effects of climate change. The forecast reliability is thus assumed to affect the ratio of probabilities expressed as a fraction of attributable risk (FAR), yet it is currently unknown whether this assumption holds.

Since the effect of forecast reliability on the FAR is arguably statistical in nature, Bellprat and Doblas-Reyes (2015) systematically study the relationship in an idealised framework of a synthetic forecast model, created statistically rather than physically. A synthetic forecast generator is considered for this purpose (Weigel et al., 2008) that allows to perform large numbers of predictions on generated observations. The model mimics the main aspects of a forecast system: a predictable component of an observable variable, an error of the forecast that is conditional to each specific forecast and a perturbation that generates an ensemble. The model is further extended to include a long-term component to account for a linear trend due to an external forcing such as the one leading to climate change. Including and excluding this linear trend allows to perform an attribution exercise on artificially generated observations. Details of the model are discussed in Bellprat and Doblas-Reyes (2015).

An example of the statistical model is given in Fig. 4 for a typical hindcast period of 30 years. The predictions in red show the forecast including the effect of climate change and results plotted in blue depict a prediction in the “world that might have been” without human perturbation of the climate. The black line shows the evolution of the observations generated

artificially, deliberately mimicking the evolution of global temperature anomalies as in Rahmsdorf et al., 2011.

An important aspect of the model is that the forecast is perfectly calibrated to the observations with respect to its climatology. The model has hence zero bias, the true trend and the exact variability as used in the observations. The evaluation described in section 3.3 would hence deem a perfect model, while its forecast reliability could still be imperfect. This is because looking at each individual predictions the forecast might not entail the observations, thought he overall variability and mean state for the entire period might still be consistent with the observations.

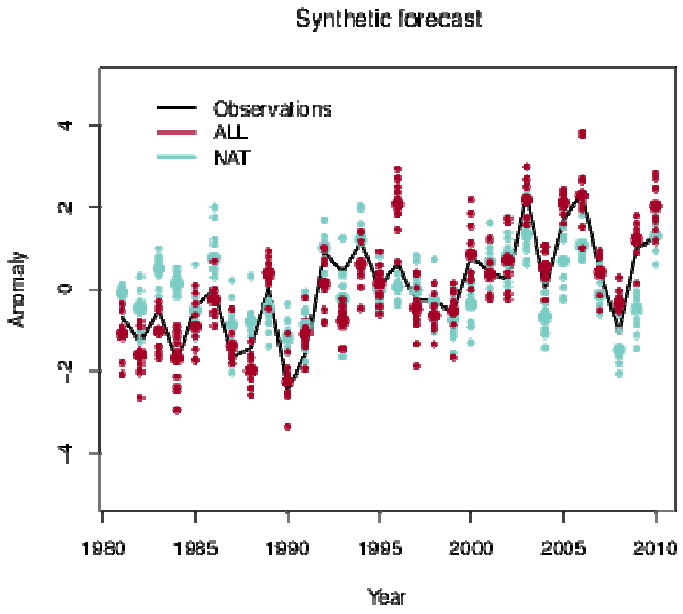

Figure 4. Example of a synthetic forecast system and the corresponding observations for a period of 30 years showing the anomalies relative to the entire period. This is typical of a retrospective seasonal hindcast. The forecasts show the individual members (small dots, ten members), the ensemble mean (large dot) for a forecast with a long-term trend emulating an external forcing due to climate change (ALL, red) and a forecast with the same model parameters without a trend describing a world without climate change (NAT, green).

In the example considered here the forecast reliability is chosen to be reasonably good, as the observations fall mostly into the spread (small dots) of the predictions. The level of the forecast reliability can be varied by changing the level of the forecast error by adding a random error with standard deviation of β and zero mean. The forecast will then have on average a zero error, but will still be imperfect considering each individual year (event). Increasing the forecast error (β) decreases the forecast reliability for which a robust inverse relationship is found described in detail in Bellprat and Doblas-Reyes (2015). Having a model where the forecast reliability can be adjusted enables a robust study of its relationship with the FAR in event attribution.

Fig. 5 shows the outcome of this analysis for an extreme event with a return period of (a) 10 years and (b) 50 years. The level of FAR is shown for different levels of reliability and different signal (S, trend in the observations) to variability (V, residual variability from the trend) ratios. The blue area shows the confidence band of FAR using re-sampling. The figure shows that unreliable forecasts overestimate the attributable risk. The overestimation is particularly strong when the S/V is low and FAR can take any value between zero and one given its large uncertainty in the 1-in-10 years event.

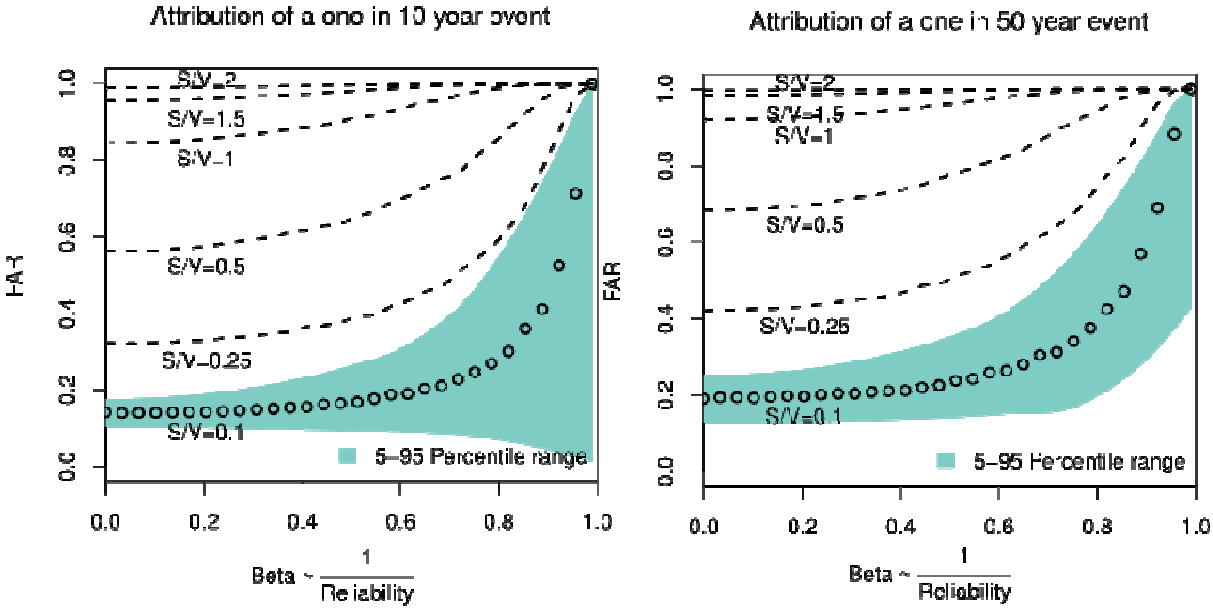

Figure 5. Variation of the FAR for a 1-in-10 years event (a) and 1-in-50 years (b) event as a function of the standard deviation of the forecast error (β) which is inversely proportional to the forecast reliability. The variation of FAR is shown for different signal (S, trend) to variability (V, residual variability from the trend) ratios where the dots show the number of different levels of β computed. The blue area shows the 5-95 % confidence interval sampled by repeating the predictions 1000 times allowing the forecast error to vary within the chosen standard deviation (β).

The systematic increase in FAR arises from an overconfidence of the prediction. As the reliability decreases the forecast becomes too confident about the possible outcomes and the event under consideration becomes very unlikely in the prediction. Unreliable systems will hence tend to underestimate the probability that an extreme event occurs. This underestimation is particularly strong if the prediction excludes climate change, as it is deliberately chosen to explain part the extreme outcome. The probability that the extreme event occurs without climate change in an unreliable forecast converges thus quickly to zero and consequently the value of FAR to one given the definition in 3.1.

The level of reliability of state-of-the-art forecasting systems for extreme events tends to be small, especially for precipitation extremes (Weissheimer and Palmer, 2014) and the identified relationship should hence become a relevant aspect in event attribution studies. It is important to note that forecast reliability is not a property that is bound to seasonal forecast systems, but the concept equally applies to free running historical simulations and also predictions made from mere observations. As a rough criterion, we conclude that

forecast reliability becomes relevant for attribution statements when forecasts become "marginally useful" and the values of attributable risk should be particularly questioned if forecasts fall into classes with lower reliability.

3.6 Evaluating error in FAR from physical models

We next consider a perfect model experiment to investigate whether the reliability is related to errors in FAR for physical models as well as statistical ones. In this experiment, "pseudo-observations" both of the actual world and the "world that might have been" (without the effect of human influence) are taken from a single pair of members of the model simulation ensemble. Using the same area-pool technique discussed in Section 3.5 the probabilities that make up the FAR can be determined exactly. If this pseudo-observational FAR (labelled F_{pseudo} in subsequent equations) is then compared to the estimate of the FAR in the rest of the ensemble, it becomes possible to calculate the error in the estimate of FAR using the model. This may then be compared to the Brier scores (Wilks, 2011) which summarise the reliability diagrams for the model runs in question.

To perform this experiment, the general circulation model from CMIP5 database (Taylor et al., 2012) with the largest number of members was selected. This was CSIRO Mk3.6.0, with 10 all-forcings members and 10 natural-forcings members. The first member was taken to be "pseudo-observations", while the other model members are used to represent normal "model" members. This constitutes a "perfect model" experiment, as the physics in the model perfectly match those found in the pseudo-observations, and only differs in its natural variability. It thus represents idealised conditions. The SREX regions (Field et al., 2012) were chosen as a global set which approximate the regional homogeneity required by the area-pool technique, where the probabilities of an event in a given season are, for evaluation purposes, considered to be the fraction of that region in which the event occurs. This was considered for a fixed season (either December to February, March to May, June to August or September to November) over all years between 1922 and 2011. The event itself is defined in the same way for both FAR and Brier Score. In this study, upper decile temperatures (representing heatwaves) and lower decile precipitation (representing droughts) were considered.

With F_{pseudo} defined for one season and member, this means an individual FAR, F_m, can be defined for each model member m of a total on N members. A mean-squared error over all the members, E_{FAR}, can therefore be produced to represent the error in FAR (Equation 1), with a form close to that of the Brier Score, which represents the error in forecast probability. This thereby maximises the chance of the two measures correlating.

$$E_{F_{AR}} = \frac{1}{N}\sum_{m=1}^{N}(F_m - F_{\text{pseudo}})^2 \qquad (1)$$

In results below, log-log plots are used to make it easier to examine the pattern by eye, since the possible range of error in FAR is from 0 to infinity. (Correlation statistics in this study will also assess the logarithmic correlation to prevent large error values greatly outweighing smaller ones in the statistics. They will only be given if they are significant at the 5% level.)

It was found that that there is some correlation between Brier score and error in FAR (0.49 for precipitation), though it is particularly weak in the case of temperature (0.34).

Precipitation is, in general, more localised in nature than temperature (Palmer et al., 2008). This leads temperature to be more predictable, aiding some aspects of this study. However, it seems likely here that instead the effectiveness of the area-pool technique as a resampling method is breaking down due to the long temperature decorrelation length scale. The area-pool assumptions still hold for precipitation here due to its shorter scale.

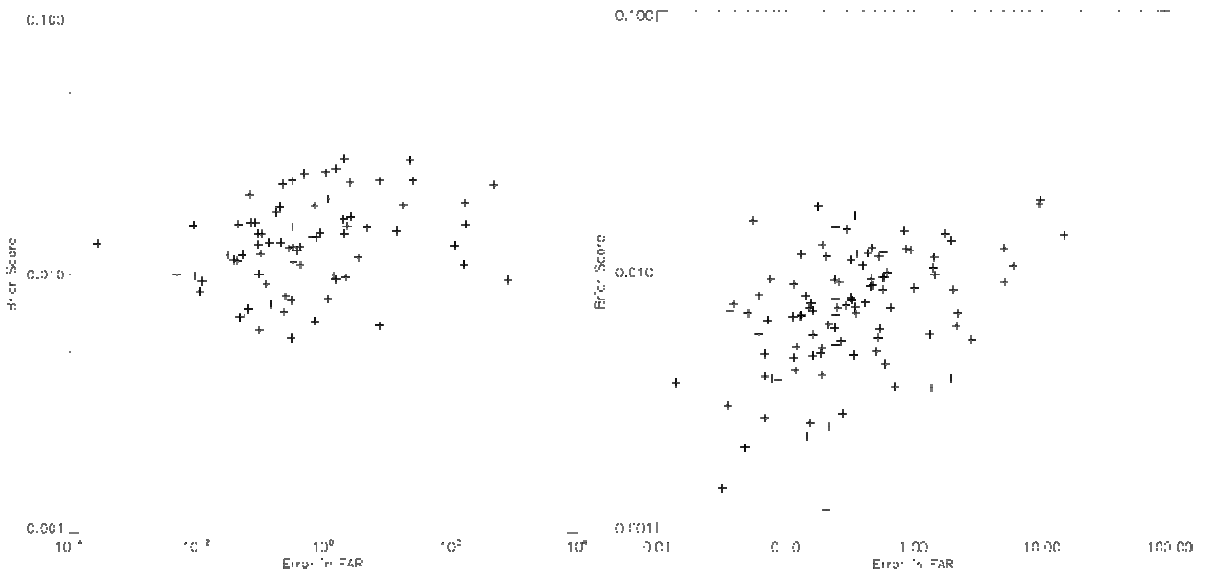

Figure 6. Log-log plot of Brier Score against error in FAR over the 1922-2011 period, from temperature (left) and precipitation (right) imperfect model experiments using CSIRO Mk3.6.0 model runs and CNRM-CM5 as pseudo-observations. Each point derives from a different season and SREX region.

Fig. 6 repeats the experiments of comparing Brier score and error in FAR, now using CNRM-CM5 pseudo-observations (again from the CMIP5 archive) with the CSIRO Mk3.6.0 model data. The result is very similar to that seen for perfect model data, with correlations of 0.29 and 0.45 for temperature and precipitation respectively. (Similar figures were also produced using HadGEM2-ES and CanESM2 for pseudo-observations, with negligibly different results, and are not shown.) This provides validation for this technique and its results. Consequently subsequent experiments in this study will continue to use the same imperfect model setup.

New methods for assessing the error in FAR

The relative weakness of the correlation between Brier Score and error in FAR might be expected, given that Brier Score is computed purely by comparing the all-forcings simulations with the observations. To calculate a measure of the errors in FAR it might be wiser to assume that the uncertainty on an attribution would depend on errors in both the all-forcings and natural probabilities of the event, given both are required to compute FAR. This study suggests two alternatives to the Brier score for this assessment.

Pre-post-climate-change combined Brier score uses standard propagation of errors (Hughes and Hase, 2010) along with the assumption that the Brier score calculated using only data from early in the time series (pre-climate-change) will correspond to that for the natural-forcings-only world, which cannot be observed. The later part of the time series is then used

in place of the all-forcings Brier score, giving Equation 2 that should account for both sets of forcings.

$$B_{F_{AR}} \approx \frac{B_{\text{pre}} + (1-\overline{F_{AR}})^2 B_{\text{post}}}{\overline{P_{\text{ALL}}}^2} \qquad (2)$$

Alternatively it is worth considering a simpler quantity. To do this, F_{pseudo} in Equation 1 is replaced with $F_{\text{pre/post}}$, a more measurable quantity which is effectively the climatological FAR in observations, as defined in Equation 3 to subsequently give the uncertainty $E_{\text{pre/post}}$.

$$E_{\text{pre/post}} = \frac{1}{N}\sum_{m=1}^{N}(F_m - F_{\text{pre/post}})^2 \quad \text{where} \quad F_{\text{pre/post}} = 1 - \frac{\overline{P_{\text{pre}}}}{\overline{P_{\text{post}}}} \qquad (3)$$

In effect, this is now evaluating the relationship between errors using the model to obtain FAR and errors in using a method based on the observational climatology. This is analogous to methods such as that of van Oldenborgh (2007), who uses observational information without modelling to estimate changes in probabilities.

Results

Following this analysis, the imperfect model experiments were re-examined, now comparing error in FAR to the pre-post-change combined score, and to the error from climatological FAR, to assess whether either outperforms Brier score. Beginning with the combined score (Equation 2), shown in Fig. 7 against error in FAR, it is clear that there is improvement for both temperature (correlation 0.71) and precipitation (0.69). Notably, by eye this correlation looks weaker for temperature. Any weakness in this regard, again, seems likely to be due to the area-pool approximations breaking down when correlation length scales become extended.

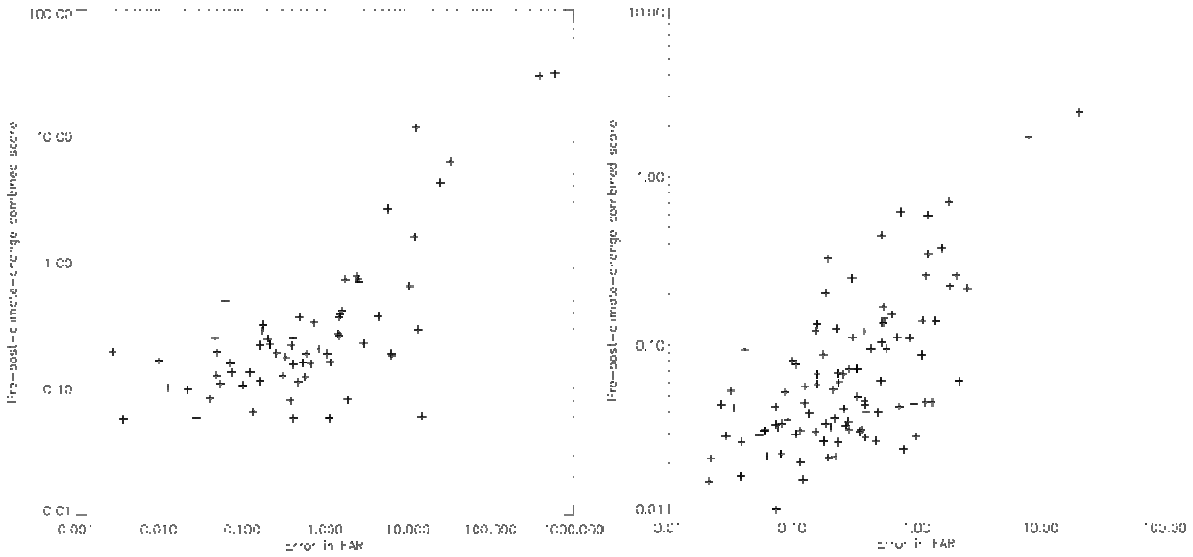

Figure 7. As Fig. 6, but showing combined pre- and post-climate-change Brier score, again with temperature (left) and precipitation (right).

Turning to the simpler quantity of the climatological FAR ($E_{\text{pre/post}}$ in Equation 3), it is plotted in Fig. 8 against the imperfect model error in FAR (E_{FAR} in Equation 1). While the correlation

for precipitation is negligibly different to combined score (0.65), for temperature there is a strong relationship between errors calculated from the observationally-based measure and the "true" error in FAR known from this being a imperfect model experiment, with only a few points straying from the 1:1 line (though these are enough to reduce the correlation to 0.63). It would seem that the area-pool weaknesses in other measures do not affect climatological FAR in the same way as the measures originating in reliability diagrams, and instead the stronger effect of the increased predictability for temperature over precipitation due to its length scales. Consequently, the observationally-based measure of climatological FAR constitutes a very good check on the accuracy of modelled FAR values alongside the combined score or equivalent reliability diagrams.

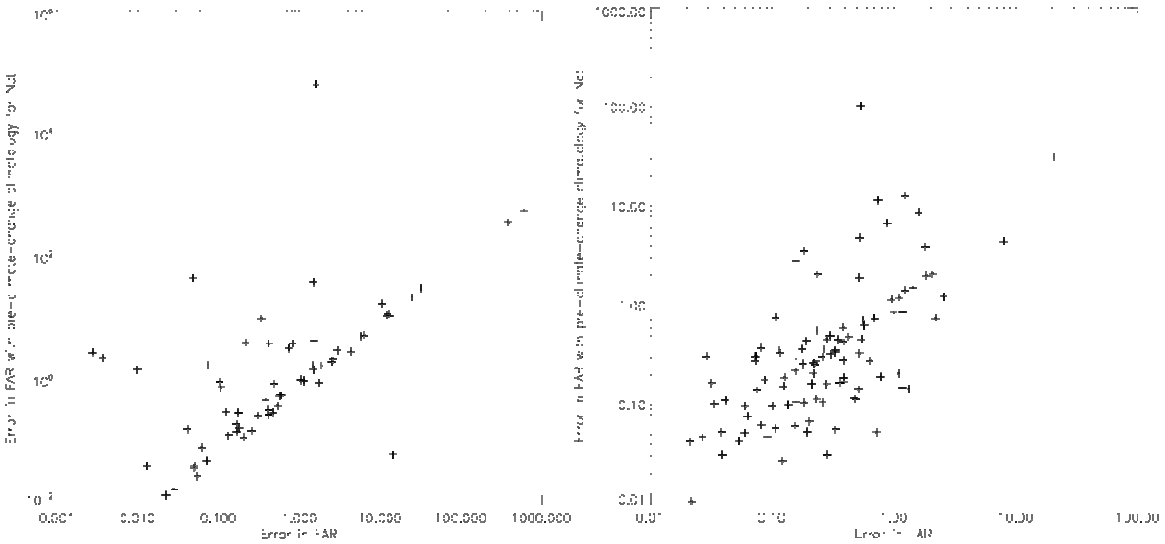

Figure 8. As Fig. 6 but showing error in FAR calculated using pre/post-climate-change climatology against that using natural run pseudo-observations in the calculation.

These results were found to hold true for shorter periods than those shown, including for the satellite period (1983 onwards). For further details see Lott and Stott (2015, in prep).

This work suggests that in the suite of attribution evaluation diagnostics used in future studies, a pair of reliability diagrams from the first and second halves of the time series in question could provide more useful information than the single reliability diagram currently in use, as that would give a graphic representation of the possible error in FAR. Furthermore, a FAR value produced from the simulations, can be compared against a climatological FAR such as that can be produced by van Oldenborgh's KNMI Climate Explorer (http://climexp.knmi.nl) as a secondary check of potential FAR errors. Further evaluation of the scores and techniques discussed here are underway as part of the EUCLEIA project. Through these studies, it should be possible to better indicate the level of confidence that can be held in the results of event attribution studies in the future.

3.7 Publications

Evaluation methodologies and their application to attribution studies are part of several studies by EUCLEIA authors published in the most recent special reports on extreme events of the Bulletin of the American Meteorological Society (BAMS):

- Explaining Extreme Events of 2013 from a Climate Perspective, 2013, Bull Amer Meteor Soc, 95
- Explaining Extreme Events of 2014 from a Climate Perspective, 2015, Bull Amer Meteor Soc, 96 (to be published in September 2015)

A detailed discussion on model evaluation is also included in a new paper that provides an overview of extreme events attribution:

- Attribution of extreme climate events (2015), Stott PA, Christidis N, Karoly D, Vanderlinden J-P, von Storch H, van Oldenborgh GJ, Otto F, Sun Y, Vautard R, Walton P, Yiou P, Zwiers FW, WIRES (submitted)

New research on model reliability that explores the links between prediction, attribution and the accuracy of the FAR is presented in the following papers:

- Bellprat O, Doblas-Reyes F (2015), Unreliable climate simulations overestimate attributable risk of extreme events, in preparation, Geophys Res Lett (submitted)
- Lott FC, Gordon M, Graham RJ, Scaife AA, Vellinga M (2014), Reliability of African climate prediction and attribution across timescales, Environ Res Lett 9, doi:10.1088/1748-9326/9/10/104017
- Lott FC, Stott PA (2015), Evaluating simulated fraction of attributable risk using observations, in preparation

4. Lessons Learnt

Model evaluation in event attribution studies is a multifaceted exercise that on the one hand increases confidence in attribution statements, but on the other needs to overcome challenges arising from the quality of the datasets used as a benchmark. The work presented in this report keeps the focus on methodological approaches and relies on the advances in other WPs and projects with regard to the quality of observational or reanalysis datasets (more discussion in Section 5). Given the uncertainties in the data, we suggest that models should ideally be evaluated against different datasets, e.g. products from two or more reanalyses.

Evaluation of the model's ability to capture physical processes and synoptic conditions that favour the occurrence of extremes is essential. A model that provides a seemingly realistic climatological likelihood of extremes, but fails to adequately reproduce important mechanisms, circulation patterns and/or modes of variability should be deemed unsuitable for attribution analyses. With regard to circulation patterns, comparisons between the simulated flow and the one derived from reanalysis Z500 fields, as demonstrated in Section 3.2 provides a simple yet useful evaluation test. Regional assessments over a range of

conditions that commonly increase the chances of extremes may be carried out in advance to enable fast-track assessments.

Testing the modelled statistical characteristics of extremes against observations can be done using timeseries and power spectra analyses, as well as qualitative and quantitative comparisons of the climatological distributions of the relevant climatic variable. These tests are especially valuable in attribution studies of events with low or no predictive skill, as they help decide whether a model that fails to successfully forecast the event is suitable for an attribution assessment, provided it can still yield reliable estimates of the climatological likelihood of events similar to the one under consideration. While good quality observational datasets for this kind of work are available for temperature and in some cases for precipitation, there are still observational gaps in other important variables. EUCLEIA aims to identify and to some extent close these gaps in future studies with a European focus.

Reliability diagrams are typically employed to determine whether predictive factors are present that allow the model to reliably forecast weather or climate events. It is shown that low reliability tends to lead to an overestimate of the FAR and an increase in its uncertainty. New work done in WP6 introduced an area-pool technique for the construction of reliability diagrams that helps a) increase the size of the sample from which probabilities are derived and b) reduce the sensitivity to the grid box size and the number of the available simulations. Probability estimates required in the estimation of the FAR were derived by application of this new approach in the context of perfect model analyses to assess the error in FAR. It is found that the error correlates well with the error in an observational estimate for which an early period is used as an approximation of the non-anthropogenic climate. Taking advantage of this correlation, new indicators of the uncertainty in the FAR can next be developed and this work is already underway as part of the EUCLEIA project.

5. Links Built

Deliverable 6.5 is key to the success of EUCLEIA as it helps establish the quality and reliability of model-based attribution assessments delivered by the attribution system developed within the project. It therefore underpins all the scientific WPs of EUCLEIA in ways explained below:

- WP4: The deliverable ensures that stakeholders are provided with scientifically robust information.
- WP5: Scientific methodologies developed in WP5 may frame the attribution problem in such a way that the changing risk of events may be constrained by dominant patterns of the atmospheric flow (T5.3). The deliverable investigates whether such patterns are well represented in the models used for attribution analyses.
- WP6: The work discussed in D6.5 is a major component of WP6. The methodologies presented here together with the additional work on key physical processes (T6.1-T6.3) will facilitate a thorough evaluation of the new attribution system.
- WP7: The reliability of attribution statements associated with the targeted test cases carried out by WP7 can be examined with methodologies discussed in this report.

- WP8: The deliverable provides ways to assess the level of confidence in the attribution products from the near real time attribution service developed by WP8, some of which will feature as event studies in the popular BAMS report on extremes.

The work discussed in D6.5 can potentially strengthen the links with other FP7 projects by communicating relevant information through the overarching activity of WP9. Reanalysis products developed by ERA-CLIM2 may be employed in evaluation methodologies. Information from satellite data provided by QA4ECV can also be useful for the evaluation of modelled physical processes. While reanalysis data are often employed as a benchmark to assess model reliability, it is important to consider the effect of uncertainties in the data and the work of UERRA can be helpful to this end. Finally, easy access to datasets facilitated by CLIPC is also expected to benefit model reliability assessments.

Finally, the application of evaluation methodologies to attribution systems developed outside EUCLEIA will improve understanding of the sensitivity of attribution products to the models they are derived from. Such an intercomparison can be coordinated by the attribution component of the C20C project, which brings together the international research community working on the development of event attribution systems.

References

Allen MR (2003) Liability for climate change, Nature 421:892

Bellprat O, Doblas-Reyes F (2015) Unreliable climate simulations overestimate attributable risk of extreme events, Geophys Res Lett (submitted)

Bellprat O, Lott F, Gulizia C, Parker HR Pampuch LA, Pinto I, Ciavarella A, Stott PA (2015) Unusual past dry and wet rainy seasons over Southern Africa and Southern South America from a climate perspective. Weather and Climate Extremes (accepted)

Christidis N, Stott PA, Scaife A, Arribas A, Jones GS, Copsey D, Knight JR, Tennant WJ (2013) A new HadGEM3-A based system for attribution of weather and climate-related extreme events, J Climate 26:2756-2783

Christidis N, Stott PA, Ciavarella A (2014) The effect of anthropogenic climate change on the cold spring of 2013 in the UK. In Explaining Extreme Events of 2013 from a Climate Perspective, Bull Amer Meteor Soc 95:S79-S82

Christidis N, Stott PA (2015) Extreme rainfall in the United Kingdom during winter 2013/2014: The role of atmospheric circulation and climate change. In Explaining Extreme Events of 2014 from a Climate Perspective, Bull Amer Meteor Soc (accepted)

Diffenbaugh NS, Scherer M (2013) Likelihood of July 2012 US temperatures in preindustrial and current forcing regimes. In Explaining Extreme Events of 2012 from a Climate Perspective, Bull Amer Meteor Soc 94:S6-S9

Field CB (2012), ed. Managing the risks of extreme events and disasters to advance climate change adaptation: Special report of the intergovernmental panel on climate change (SREX), Cambridge University Press

Fischer EM, Seneviratne SI, Vidale PL, Lüthi D, Schär C (2007) Soil Moisture–Atmosphere Interactions during the 2003 European Summer Heat Wave, J. Climate 20:5081–5099

Gillett NP, Allen MR, Tett SFB (2000) Modelled and observed variability in atmospheric vertical temperature structure, Clim Dyn 16:49-61

Hewitt HT, Copsey D, Culverwell ID, Harris CM, Hill RSR, Keen AB, McLaren AJ, Hunke EC (2011) Design and implementation of the infrastructure of HadGEM3: the next-generation Met Office climate modelling system, Geosci Model Dev, 4:223-253

Hoerling M, Eischeid J, Perlwitz J, Quan X, Zhang T, Pegion P (2012) On the increased frequency of Mediterranean drought, J Climate 25:2146–2161

Hughes IG, Hase TPA (2010), Measurements and their Uncertainties, Oxford University Press

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Leetmaa A, Reynolds R, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Jenne R, Joseph D (1996) The NCEP/NCAR 40-year reanalysis project, Bull Amer Meteor Soc, 77, 437-470

King AD, Lewis SC, Perkins SE, Alexander LV, Donat MG, Karoly DJ, Black T (2013) Limited evidence of anthropogenic influence on 2011-12 extreme rainfall over southeast Australia. In Explaining Extreme Events of 2012 from a Climate Perspective, Bull Amer Meteo. Soc, 94:S55-S58

Lewis SC, Karoly DJ, Yu M (2014) Quantitative estimates of anthropogenic contributions to extreme national and State monthly, seasonal and annual average temperatures for Australia, Australian Meteorological and Oceanographic Journal 64:215-230

Lott FC, Christidis N, Stott PA (2013), Can the 2011 East African drought be attributed to human-induced climate change? Geophys Res Lett 40:1177-1181

Lott FC, Gordon M, Graham RJ, Scaife AA, Vellinga M, Reliability of African climate prediction and attribution across timescales (2014), Environ Res Lett 9, doi:10.1088/1748-9326/9/10/104017

Lott FC, Stott PA (2015), Evaluating simulated fraction of attributable risk using observations, in preparation

Pall P, Aina T, Stone DA, Stott PA, Nozawa T, Hilberts AGJ, Lohmann D, Allen MR (2011) Anthropogenic greenhouse gas contribution to flood risk in England and Wales in autumn 2000, Nature 470:382–385

Palmer TN, Doblas-Reyes FJ, Weisheimer A, Rodwell MJ (2008) Towards seamless prediction: Calibration of climate change projections using seasonal forecasts, Bull Amer Meteor Soc 89:459-470

Stott PA, Allen M, Christidis N, Dole RM, Hoerling M, Huntingford C, Pall P, Perlwitz J, Stone D (2013) Attribution of weather and climate-related extreme events. Climate Science for Serving Society: Research, Modeling and Prediction Priorities, G. R. Asrar and J. W. Hurrell (eds), doi:10.1007/978-94-007-6692-1_12

Taylor K, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experimental design, Bull Am Meteorol Soc 93: 485-498

van Oldenborgh GJ (2007), How unusual was autumn 2006 in Europe? Clim Past 3:659–668

Wilks DS (2011), Statistical Methods in the Atmospheric Sciences (3rd edition), Elsevier

WMO (2002), Standardised Verification System (SVS) for Long-Range Forecasts (LRF), in New Attachment II-9 to the Manual on the GDPS (WMO-No. 485), Volume I, WMO (Geneva)

Yiou P, Cattiaux J (2013) Contribution of atmospheric circulation to wet north European summer precipitation of 2012. In Explaining Extreme Events of 2012 from a Climate Perspective, Bull Amer Meteor Soc 94:S39-S41

Yiou P, Cattiaux J (2014) Contribution of atmospheric circulation to wet southern European winter of 2013. In Explaining Extreme Events of 2012 from a Climate Perspective, Bull Amer Meteor Soc 95:S66-S69