Evaluating Simulated Fraction of Attributable Risk Using Climate Observations

FRASER C. LOTT AND PETER A. STOTT

Met Office Hadley Centre, Exeter, United Kingdom

(Manuscript received 12 August 2015, in final form 24 March 2016)

ABSTRACT

Although it is critical to assess the accuracy of attribution studies, the fraction of attributable risk (FAR) cannot be directly assessed from observations since it involves the probability of an event in a world that did not happen, the "natural" world where there was no human influence on climate. Instead, reliability diagrams (usually used to compare probabilistic forecasts to the observed frequencies of events) have been used to assess climate simulations employed for attribution and by inference to evaluate the attribution study itself. The Brier score summarizes this assessment of a model by the reliability diagram. By constructing a modeling framework where the true FAR is already known, this paper shows that Brier scores are correlated to the accuracy of a climate model ensemble's calculation of FAR, although only weakly. This weakness exists because the diagram does not account for accuracy of simulations of the natural world. This is better represented by two reliability diagrams from early and late in the period of study, which would have, respectively, less and greater anthropogenic climate forcing. Two new methods are therefore proposed for assessing the accuracy of FAR, based on using the earlier observational period as a proxy for observations of the natural world. It is found that errors from model-based estimates of these observable quantities are strongly correlated with errors in the FAR estimated in the model framework. These methods thereby provide new observational estimates of the accuracy in FAR.

1. Introduction

Attribution is the process of evaluating the relative contributions of multiple causal factors to a change or event with an assignment of statistical confidence (Hegerl et al. 2010). By its nature, an attribution study requires some sort of model, statistically or physically based, in order to quantify how different factors could have contributed to an observed change. It is clear from this that the model must be evaluated to establish whether it is representative of the observed reality.

To do this, a range of model validation techniques have been included in attribution studies in the past, to assess the suitability of the simulation by comparing its properties to observations. These have included spectra of variability, distributions comparing the range of observed climatology to that of the models used, and reliability diagrams [examples of all these techniques can be found in Christidis et al. (2013)]. However, the use of the reliability diagram has at times been contentious, since its meaning to attribution is not fully understood. This paper aims to increase this understanding and advise on the best manner for its use.

a. Reliability diagrams and the Brier score

Reliability diagrams (Wilks 2011) were devised for seasonal forecasts to express the ability of probabilistic forecasts to reproduce the statistically observed frequency of observed events. To do this, they show observed climatological frequency of an event over a given threshold (e.g., above-average temperature or lower-decile rainfall) against its modeled or forecast probability, usually in a given season. Typically this is binned by forecast probability obtained across a model ensemble (Fig. 1, left), which then enables the observed fraction to be obtained over the whole climatology. However, using area-pooling techniques considering individual grid points in a homogeneous region (Lott et al. 2014), an unbinned diagram may also be produced (Fig. 1, right), with each point representing a different region and season. This technique has been employed in a number of previous studies (e.g., Annan and Hargreaves 2010; Van Oldenborgh et al. 2013),

Denotes Open Access content.

Corresponding author address: Fraser C. Lott, Met Office Hadley Centre, Fitzroy Road, Exeter EX1 3PB, United Kingdom.
E-mail: fraser.lott@metoffice.gov.uk

DOI: 10.1175/JCLI-D-15-0566.1

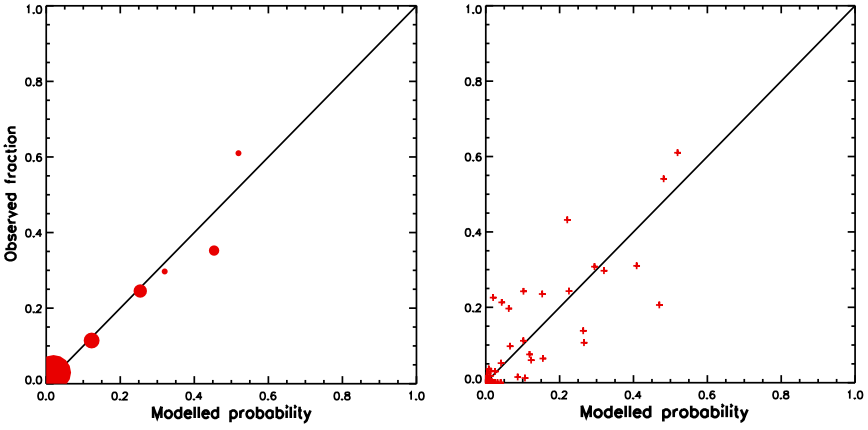

FIG. 1. (left) Binned and (right) unbinned reliability diagrams. This example is for upper-decile summer (June–August) temperatures in the Mediterranean SREX region (IPCC 2012). Point area for the binned diagram indicates the number of samples within that bin.

but here the method of Lott et al. (2014) is followed, which is summarized as follows:

1) Take the set of grid boxes within each study region and season as a data pool, as if each originated from a different ensemble member.
2) For the observations and for each model ensemble member, calculate the proportion of the region for which the event threshold is exceeded.
3) Calculate the mean proportion for all model ensemble members for that event. This is taken to represent the forecast probability of an event for locations in that region and season.
4) Plot the observed fraction of the region exceeding the threshold (representing its regional average frequency) against the modeled probability. This produces a figure with each point representing the studied region and season in each year.

The reliability of a forecast is indicated in these diagrams by how close the points lie to the 1:1 line. This may be summarized by a variety of statistics for different situations. For this study, the Brier score (Wilks 2011), calculated from the unbinned diagram, is chosen as a single figure to represent its reliability. As can be seen in Eq. (1.1), where k is the season of the year in question, n is the number of seasons per year, y is modeled probability, and o is observed frequency, it is essentially a mean-squared error, and therefore the score is better the closer it is to zero. (Note that the area-pool method changes the definition of the observed frequency, from what is usually a binary measure of whether the modeled event is observed to the fraction of the area in which the event is observed, assuming the homogeneity described above.) Unlike other statistics, Brier score does not need to be related to a model climatology. As will become clear, this is useful when applied to event attribution, as the climatology will often be changing with time:

$$B = \frac{1}{n}\sum_{k=1}^{n}(y_k - o_k)^2. \tag{1.1}$$

b. Event attribution

Event attribution (Allen 2003; Stott et al. 2013) seeks to determine whether weather and climate events have been made more or less likely as a result of a given climate forcing. As with the reliability diagram, the event is defined as exceeding a given threshold. The forcing is most commonly that of human-induced climate change and will be examined as such from here on in. It considers not just the probability of an event happening in the actual world P_{ALL} (where all known external climate forcings, both natural and anthropogenic, are present) but also the probability of the same event happening in "the world that might have been" (i.e., that unaffected by anthropogenic climate change, where only natural forcings are present and whose probability is given by P_{NAT}). This is quantified using the fraction of attributable risk [FAR, or simply F in Eq. (1.2) and subsequent equations for brevity]:

$$F = 1 - \frac{P_{\mathrm{NAT}}}{P_{\mathrm{ALL}}}. \tag{1.2}$$

The common techniques to calculate these probabilities (Pall et al. 2011; Christidis et al. 2013; King et al. 2013) are to compare two different ensembles of modeled climate, simulating all forcings and the natural world, respectively, and count the fraction of those ensemble members in which that event manifests. Uncertainties on these values are either derived from bootstrapping

from the model (Pall et al. 2011) or by comparison with other models (Lott et al. 2013; Christidis and Stott 2014). The problem with these techniques is that the observations only contribute through definition of the event (e.g., as a temperature or rainfall threshold) and through any bias correction undertaken. This means that estimates of uncertainties are always model based.

As described in the introduction, a range of model validation techniques have been included in event attribution studies to address this weakness by comparing various aspects of models to observations, including reliability diagrams. However, whether forecast reliability equates to attribution reliability has not previously been tested, and so the implications of such a diagram to the rest of the study are not fully understood. For example, while reliability for African climate prediction has been analyzed (Lott et al. 2014), its relationship to the positive attribution result of the East African drought of 2011 (Lott et al. 2013) has not yet been made clear (Lott et al. 2014). This is important because it has been argued that attribution does not require predictability because it is performed on past events (Christidis et al. 2013). Rather, an accurate, unbiased climatology in the worlds with and without human influence is thought to be required.

In seasonal forecasting, it has been found that realistic greenhouse gas forcings increase the reliability of probabilistic forecasts (Doblas-Reyes et al. 2006; Liniger et al. 2007) and that this is due to the presence of a trend (Scaife et al. 2009), which in some regions is the main contributor to predictability. It might be supposed that the presence or absence of this trend would make reliability a good indicator for the quality of attribution. However, seasonal forecasting does not have the same goal as attribution (i.e., to determine the probabilities with and without anthropogenic forcings, not just how much they or their associated variables have changed). Consequently, while a model's ability to reliably estimate the probability of events in the world as observed P_{ALL} can be effectively compared with observations using techniques derived from seasonal forecasting (provided observations of such events exist), the probability of such an event in the counterfactual world of natural forcings only P_{NAT} has no direct observational comparison. How, then, does one evaluate an error in FAR, and how does one relate such errors to reliability diagrams? In the absence of a clear theory giving such a relationship, this paper attempts to address these gaps in knowledge in a quantitative and empirical manner.

2. Applicability of Brier score for assessing errors in FAR

To investigate whether the Brier score is related to errors in FAR, this study considers a perfect model experiment. In this experiment, "pseudo-observations" both of the actual world and the world that might have been are taken from a single pair of members of the model simulation ensemble. The same area-pool technique as that used in the reliability diagrams (Lott et al. 2014) may also be used to determine the probabilities that make up the FAR. By counting the proportion of grid boxes exceeding the event threshold for each world, analogs to P_{NAT} and P_{ALL} are obtained, and in turn a FAR can be calculated. If this pseudo-observational FAR (labeled 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 for the model runs in question.

To perform this experiment, the general circulation model from phase 5 of the Coupled Model Intercomparison Project (CMIP5) database (Taylor et al. 2011) with the largest number of members was selected. This was CSIRO Mk3.6.0 (Rotstayn et al. 2012), with 10 all-forcings members and 10 natural-forcings members. The first member was taken to be pseudo-observations, while the other 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 differs only in its natural variability. It also does not include problems that might result from differing methods of measuring the observed variable. Thus it represents idealized conditions. The regions from the Special Report on Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation (SREX; IPCC 2012) were chosen as a global set, with the assumption that each of these approximate the regional homogeneity required by the area-pool technique (although in practice this homogeneity is highly variable with SREX region, and consequently this should be considered a relatively large approximation). 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–February, March–May, June–August, or September–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 and lower-decile precipitation were considered as simple indices. These represented heat waves and droughts, respectively, although these are rough generalizations, and in practice both should be defined more rigorously in a full event attribution study, taking into account other variables such as soil moisture. In addition, deciles are not particularly statistically extreme and as such cannot fully represent the accuracy of

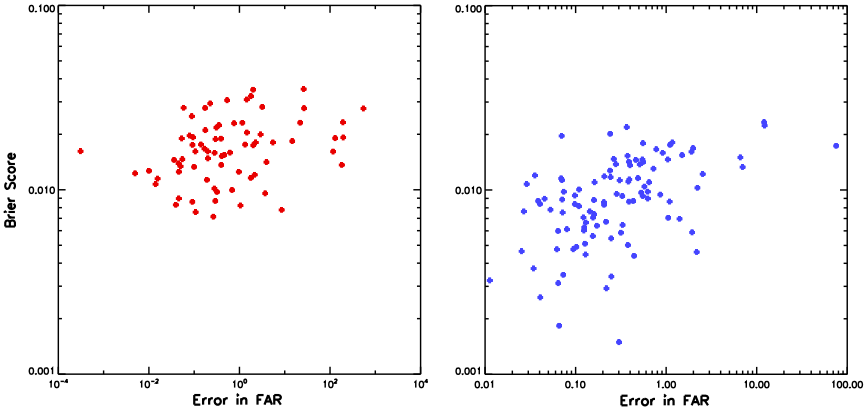

FIG. 2. Log–log plot of Brier score against error in FAR over the 1922–2011 period, using (left) upper-decile temperature and (right) lower-decile precipitation from perfect model experiments using CSIRO Mk3.6.0 model runs. Each point derives from a different season and SREX region. (Note the change in scale for the *x* axis.)

the model for the most severe events. However, this is all that can be considered in the framework of this experiment, as more extreme thresholds result in the sample of data being insufficiently large to make a meaningful study.

With F_{pseudo} defined for one season and member, an individual FAR F_m can be defined in turn for each model member m of a total of N members remaining (once the "observations" member is discounted). A mean-squared error over all the members [E_{FAR} in Eq. (2.1)] can therefore be produced to represent the error in FAR, with a form close to that of the Brier score (i.e., a variance), which represents the error in forecast probability:

$$E_{\text{FAR}} = \frac{1}{N}\sum_{m=1}^{N}(F_m - F_{\text{pseudo}})^2. \qquad (2.1)$$

Is it reasonable to suppose that Brier score represents a measure of the error in FAR? Bellprat and Doblas-Reyes (2016) indicate that statistically this should be the case, but does this extend to physical models? Figure 2 shows examples for temperature and precipitation, with each point representing a different season and region combination. Note that the point must be discarded if the event does not happen in one or more all-forcings members since this causes FAR, and in turn E_{FAR}, to become infinite. Similarly, log–log plots are used in this and all subsequent figures to make it easier to examine the pattern by eye, since the possible range of error in FAR is from 0 to infinity. (Pearson correlation statistics in this study will also assess the logarithmic correlation to prevent large error values greatly outweighing smaller ones in the statistics. The resultant p values will only be given if they are significant at the 5% level.)

From this figure, it is notable that there is some correlation between Brier score and error in FAR (0.49 for precipitation), although it is particularly weak in the case of temperature (0.34). This difference between the two variables is likely because precipitation is, in general, more localized in nature than temperature (Palmer et al. 2008). In most cases, this leads temperature to be more predictable, and as will be seen, this aids some aspects of this study. However, in the case of Brier score here, it appears that instead this is reducing the effectiveness of the area-pool technique as a resampling method. This is due to the long temperature decorrelation length scale, which means that the grid boxes are no longer samples that are as independent as they would be were they each obtained instead from different ensemble members. This is likely made worse by inhomogeneities present in the SREX regions. In contrast, the area-pool assumptions still hold for precipitation here because of its shorter decorrelation length scale. It is reasonable to hypothesize that this gives the clear increase seen in correlation between Brier score and error in FAR for rainfall over that seen for temperature in Fig. 2.

Broader applicability—imperfect models

While the perfect model study provides an excellent proof of principle, it does not completely represent a real-world experiment, precisely because models are not perfect and cannot be expected to have dynamics identical to those of the observed physical world. To test how well such an observationally based measure of model skill represents a model's ability to calculate FAR, an "imperfect model" experiment is conducted, in which one model is used to represent pseudo-observations and

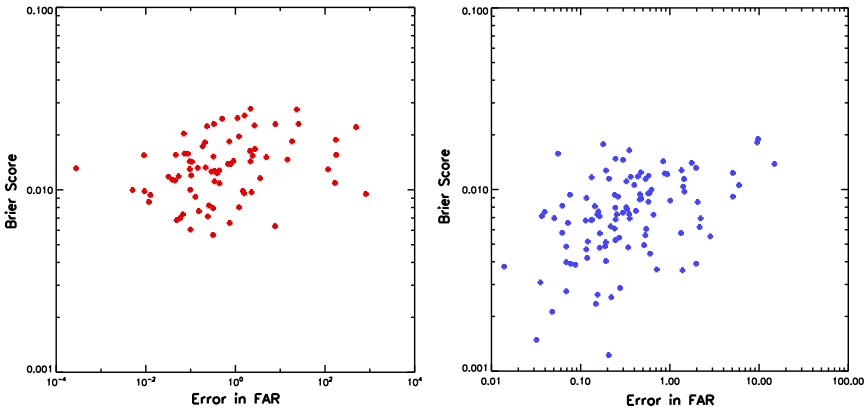

FIG. 3. As in Fig. 2, but for an imperfect model setup using CNRM-CM5 as pseudo-observations.

another model is used to calculate FAR. This takes account of imperfections in models' abilities to represent climate variability and change, imperfections represented by the difference between the climate models. At the same time, the correct value of FAR is still known, since unlike in the real-world case, it is known what the world would have done absent anthropogenic forcings (although there will be sampling uncertainty within this). Figure 3 repeats the experiments of comparing Brier score and error in FAR, now using CNRM-CM5 (Voldoire et al. 2011) 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 the HadGEM2-ES and CanESM2 models 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.

3. 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. These are the climatological FAR and the combined pre- and postchange score, which will be detailed first.

a. *Adapting Brier score to attribution*

For any $f = f(x, y, \ldots)$, standard propagation of errors (Hughes and Hase 2010) gives

$$\varepsilon_f^2 = \left(\frac{\partial f}{\partial x}\right)^2 \varepsilon_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \varepsilon_y^2 + \ldots , \tag{3.1}$$

where ε_n is the standard error on n. Applying this to FAR, taking partial derivatives of Eq. (1.2) gives

$$\frac{\partial F}{\partial P_{\mathrm{ALL}}} = \frac{P_{\mathrm{NAT}}}{P_{\mathrm{ALL}}^2} \quad \text{and} \quad \frac{\partial F}{\partial P_{\mathrm{NAT}}} = \frac{-1}{P_{\mathrm{ALL}}}.$$

To complete the application to the error on FAR, it is then necessary to obtain the errors on P_{NAT} and P_{ALL}. Since the previous section has found that there is some correlation of Brier score with the error on FAR as obtained with the area-pool method, it might be reasonable to assume that the Brier score, as a mean-squared error, can be taken to represent the error variance ε^2. It may be further supposed that there are separate Brier scores related to each of the all-forcings and natural-forcings worlds, B_{ALL} and B_{NAT}, respectively. Substituting this into Eq. (3.1) gives

$$B_{\mathrm{FAR}} = \frac{P_{\mathrm{NAT}}^2}{P_{\mathrm{ALL}}^4} B_{\mathrm{ALL}} + \frac{1}{P_{\mathrm{ALL}}^2} B_{\mathrm{NAT}}.$$

This can be simplified by substitution of FAR from Eq. (1.2):

$$B_{\mathrm{FAR}} = \frac{B_{\mathrm{NAT}} + (1 - F)^2 B_{\mathrm{ALL}}}{P_{\mathrm{ALL}}^2}. \tag{3.2}$$

The B_{ALL} should correspond to B in Eq. (1.1) since y and o values correspond to individual probabilities; however, B_{NAT} cannot be observed. However, it can be supposed

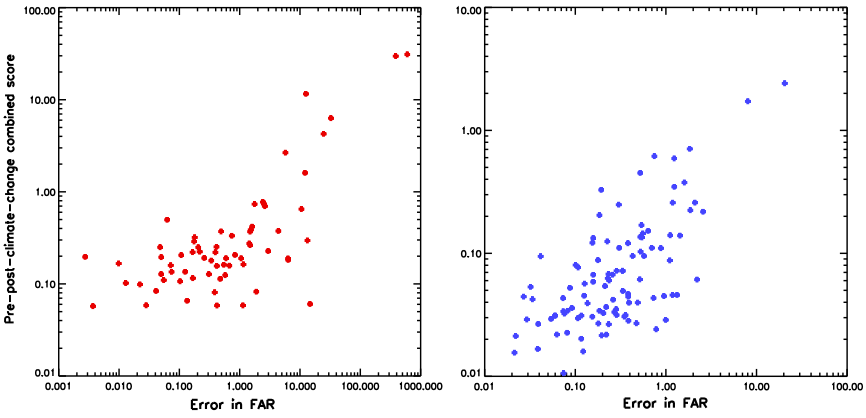

FIG. 4. As in Fig. 3, but showing combined pre- and post-climate-change Brier score, again with (left) temperature and (right) precipitation. (Note the change in scales for the x and y axes.)

that the world with natural forcings only is closer to early real-world observations when the climate had changed less. Therefore this value can be approximated by dividing up the record into prechange and postchange periods and calculating B_{ALL} for each section, calling them B_{pre} and B_{post}. Substituting these in for B_{NAT} and B_{ALL} in Eq. (3.2) gives Eq. (3.3), an alternative score that should account for both sets of forcings (where overbars denote ensemble averages, which are necessary because of the area-pool method used in this study):

$$B_{\mathrm{FAR}} \approx \frac{B_{\mathrm{pre}} + (1-\overline{F})^2 B_{\mathrm{post}}}{\overline{P_{\mathrm{ALL}}}^2}. \tag{3.3}$$

Ideally, observations from the preindustrial period would be used to represent the natural world in this evaluation technique. Of course, almost no measurements are available from this period. Depending on the variable and region in question, the observation record may start as late as the end of the twentieth century. A practical decision must therefore be made as to what constitutes prechange. For the initial purposes of this study, this period was chosen to be 1922–51 because both temperature and precipitation are well observed by that point. A corresponding 30-yr period, 1982–2011, was chosen as the postchange climatology.

b. *A simpler alternative—climatological FAR*

What if the assumption that Brier score may be used as an analog for error in FAR does not hold true? It may sometimes be found that the Brier score is sufficiently abstracted from the process of calculating FAR that it cannot be used in place of an error variance as assumed in Eq. (3.2) onward. It is therefore worth considering a simpler quantity. To do this, F_{pseudo} in Eq. (2.1) is replaced with $F_{\mathrm{pre/post}}$, a more measurable quantity that estimates FAR using the pre- and postchange climatological probabilities as defined in Eq. (3.4), to give the uncertainty $E_{\mathrm{pre/post}}$. In effect, this is now evaluating the error between using the model to obtain FAR and using a method based on the observational climatology. The $F_{\mathrm{pre/post}}$ is analogous to methods such as that of Van Oldenborgh (2007), who uses observational information without modeling to estimate changes in probabilities. Therefore, by comparing $E_{\mathrm{pre/post}}$ in Eq. (3.4) to E_{FAR} in Eq. (2.1), it will be possible to evaluate whether this is a stronger or weaker method for evaluating error in FAR than those that use the Brier score:

$$E_{\mathrm{pre/post}} = \frac{1}{N}\sum_{m=1}^{N}(F_m - F_{\mathrm{pre/post}})^2, \quad \text{where}$$
$$F_{\mathrm{pre/post}} = 1 - \frac{\overline{P_{\mathrm{pre}}}}{\overline{P_{\mathrm{post}}}}. \tag{3.4}$$

4. Results

Following this analysis, the imperfect model experiments were reexamined, now comparing error in FAR to the prechange–postchange combined score and to the error from climatological FAR, to assess whether either outperforms Brier score. Note that once again, CSIRO Mk3.6.0 is used for model members and CNRM-CM5 for pseudo-observations in the results that follow (HadGEM2-ES and CanESM2 pseudo-observations give very similar results and are not shown).

Beginning with the combined score [Eq. (3.3)], shown in Fig. 4 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

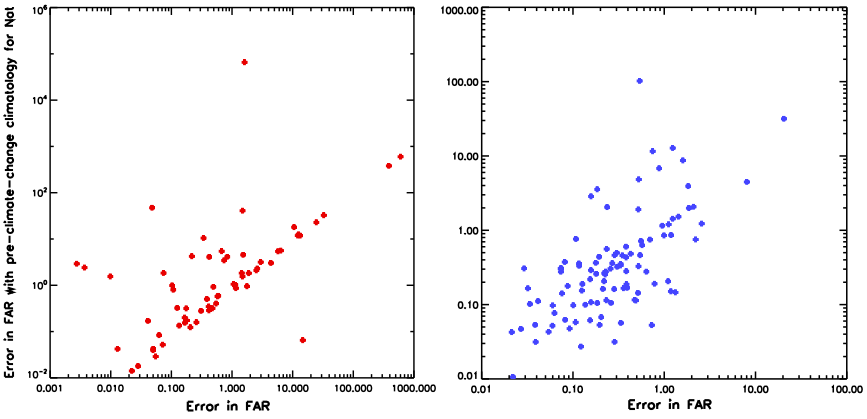

FIG. 5. As in Fig. 3, but showing error in FAR calculated using pre- and post-climate-change climatology against that using natural run pseudo-observations in the calculation (as in previous diagrams). (Note the change in scales for the *x* and *y* axes.)

approximations breaking down when correlation length scales become extended.

Turning to the simpler quantity of the climatological FAR [$E_{\text{pre/post}}$ in Eq. (3.4)], it is plotted in Fig. 5 against the imperfect model error in FAR [E_{FAR} in Eq. (2.1)]. 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 an imperfect model experiment, with only a few points straying from the 1:1 line (although these are enough to reduce the correlation to 0.63). Given the contrast to the detrimental effect that it appears to have in Figs. 2–4, it would seem that the area-pool method is not affected by decorrelation length for climatological FAR in the same detrimental way as the measures originating in reliability diagrams. Instead, the stronger effect is that of increased predictability for temperature over precipitation owing to its length scales, giving in turn a much reduced scatter for the temperatures in Fig. 5 compared to precipitation. Consequently, the observationally based measure of climatological FAR constitutes a very good check on the accuracy of modeled FAR values alongside the combined score or equivalent reliability diagrams.

5. Application to the satellite period

As discussed earlier, the observations in the region or variable of interest often do not extend back to a period that can really be considered to be before the climate has changed from its natural state. The ability to use a shorter period to evaluate the attribution must consequently be considered. The following figures consider a common scenario, where the data are satellite derived and consequently only start in the 1980s [such as the TAMSAT precipitation dataset (Tarnavsky et al. 2014; Maidment et al. 2014), which only starts in 1983]. This late start, combined with the fact that the time series is of insufficient length to support two 30-yr climatology periods for comparison, necessitates shortening the postchange period to 1997–2010 and moving the prechange period to 1983–96. This then makes it possible to reexamine the three measures (Brier score, prechange–postchange combined score, and climatological FAR) in this reduced dataset.

The reduction of the dataset from 90 years to a mere 28 years has a definite impact on the ability of the simple Brier score to represent error in FAR (Fig. 6), with the correlation reduced to 0.30 for precipitation and becoming insignificant for temperature.

For the prechange–postchange combined score, the reduction in data, and indeed the reduction in climate change taking place between the start and the end of the time series, might be expected to have a substantial negative effect on the correlation to error in FAR. As can be seen in Fig. 7, this proves not to be the case, with the correlation maintained at 0.70 for temperature and only sustaining a small drop to 0.59 for precipitation. This indicates that the change over the satellite period is still sufficient to represent the overall direction of change in the climate, if not its magnitude, and that this is what is necessary to reveal differences between the model and the observations.

As can be seen in Fig. 8, the correlation between climatological and imperfect model error in FAR actually appears enhanced by the shift to the satellite period, both for temperature (0.74) and for precipitation (0.67).

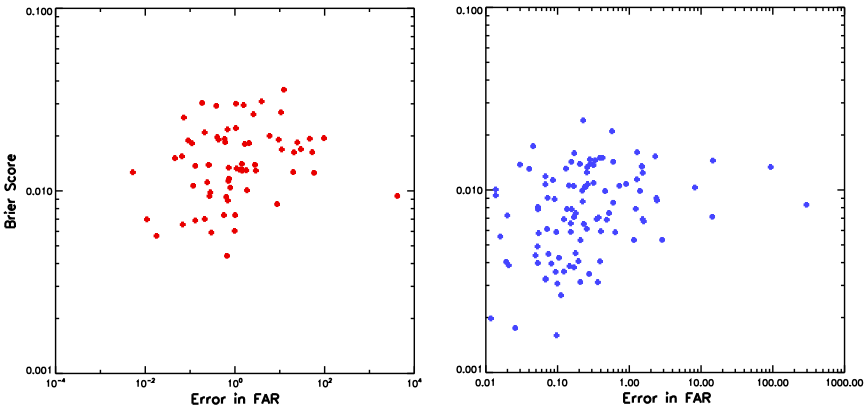

FIG. 6. As in Fig. 3, but for the period 1983–2010, for (left) temperature and (right) precipitation.

However, great care should be taken in concluding that there is an intrinsic benefit of a shorter climatology. For example, this might indicate that by chance some confounding noise from natural variability or forcings are not seen during the shorter period. Decadal variability will also play a role, so such short periods may not always be effective divining the difference between the pre- and postchange climatologies. Similarly, because this is a model study, variations in coverage do not come into play. However, with real observations, the increased observational coverage provided by satellites may also help considerably here, although the in situ gauge network is better for the earlier period so the relative usefulness of different climatologies may be found to vary between temperature and rainfall.

These results would seem to indicate that the two new diagnostics are complementary measures of the error in FAR and that they are still suitable for use in the satellite era. In turn, this would suggest 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 should replace the single reliability diagram currently in use. This would give a graphic representation of the possible error in FAR and is particularly preferable for studies using short time series where the weaknesses of a single diagram appear magnified. Furthermore, when producing a FAR value from the simulations, this should be compared against a climatological FAR such as that can be produced by KNMI Climate Explorer (Van Oldenborgh 1999) as a secondary check of potential FAR errors. Note that the Climate Explorer calculations of climatological FAR are superior to the simplified versions used in this study since the climatology is fitted to a generalized extreme value (GEV) distribution with continuously varying parameters. This both avoids assuming stationary pre- and postchange worlds and enables the assessment of more extreme events than those seen in the climatology. The use of GEV distributions could also be used in the future to resample the axes of

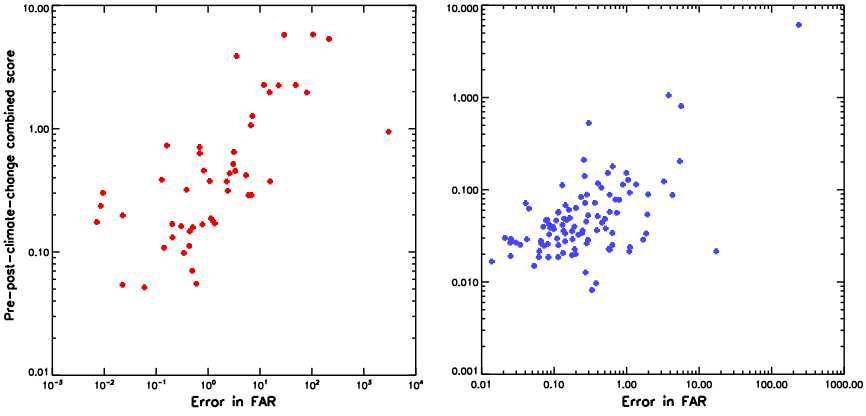

FIG. 7. As in Fig. 6, but for prechange–postchange combined score. (Note the change in scales for the *x* and *y* axes.)

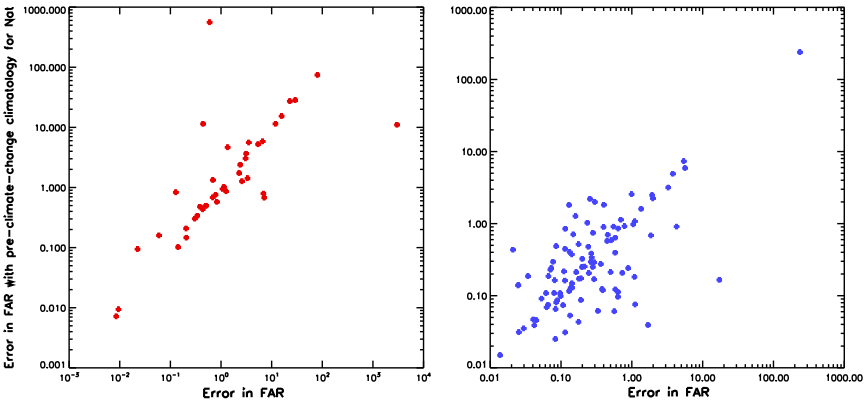

FIG. 8. As in Fig. 6, but for error from climatological FAR. (Note the change in scales for the x and y axes.)

reliability diagrams, which would make it possible to assess events with much lower probability.

6. Discussion and future work

The results obtained regarding the relationship between reliability diagrams and FAR, as well as their improvement through use of the pre- and postchange climate, do not seem likely to be an artifact of this technique since Bellprat and Doblas-Reyes (2016) also show the relationship using a statistical model. Consequently, the use of the area-pool method to demonstrate the importance of accounting for changes in forcings through time should not imply that it must be used in all further calculations of reliability and FAR. While the technique is useful for increasing sample size and for obtaining the estimation of FAR error in this paper, area pooling nonetheless produces substantial complications through the necessity to assume the homogeneity of regions and the effects of varying decorrelation length scale. Provided there is sufficient sample size, it is preferable to use the more common means of constructing reliability diagrams (WMO 2002) and to base probability on intraensemble variability, which should be more statistically useful for full attribution studies. If data are still insufficient to sample for a reliability diagram, it may be worth considering alternate means, such as fitting generalized extreme value distributions to the model and observations and then producing reliability diagrams based on sampling from the distribution. This would also be a solution to the problem of there being insufficient observations to sample extremely rare events (a problem that area pool does not solve in any case).

Similarly, it should be considered that FAR is not always the best way of quantifying the change in event probability, even if it is the most common. Its asymmetry (where events only possible in the all-forcings world have a FAR of 1, but events only possible in the natural world have a negatively infinite FAR) has already been seen in this study to necessitate the application of logarithms to deal with the large variation in the scale of the error. The solution to this problem of asymmetry is to use a new metric that incorporates this logarithm, which is introduced here as difference of binary logarithms of probability (DBLP), or in mathematical notation $\Delta \mathrm{lb} P$, where lb is the binary (i.e., base 2) logarithm. This is simply related to risk ratio $P_{\mathrm{ALL}}/P_{\mathrm{NAT}}$, used elsewhere (e.g., Angélil et al. 2014), as shown by

$$\Delta \mathrm{lb} P = \mathrm{lb} P_{\mathrm{ALL}} - \mathrm{lb} P_{\mathrm{NAT}} \equiv \mathrm{lb} \frac{P_{\mathrm{ALL}}}{P_{\mathrm{NAT}}}. \qquad (6.1)$$

The choice of binary logarithm over the natural logarithm or that of any other base is to ensure that DBLP remains a human-readable index. Its value increases by 1 each time the probability doubles, so an event that is twice as likely because of climate change has a DBLP of 1, while an event that is 4 times as likely has a DBLP of 2. Similarly, events that are a half or a quarter as likely have a DBLP of −1 and −2, respectively, and no change gives a DBLP of 0. Thus it is neither mathematically nor psychologically biased, as its use, unlike FAR, does not imply a greater interest in events whose probability increases with climate change [since FAR only represents a true conditional probability when it is positive (Hansen et al. 2014)]. Combining this with the symmetry, future studies should be able to relate these errors to their sources considerably more easily.

It is unfortunate, therefore, that DBLP is not suited to use in this study. While on face value it appears that its symmetry will clarify a number of calculations made in earlier chapters, in fact DBLP does not sit well with the area-pooled frequencies that are used to approximate

probabilities. In Eq. (2.1), E_{FAR} blows up for a given region and season if any member has no grid boxes that report the event in the all-forcings world, and that event must be discarded in the subsequent correlation plots. If a similar mean-squared error over all DBLP values E_{DBLP} is formulated in its place, this blowup now also takes place for events not seen in any given natural member, which in a warming world is particularly likely to take place in upper-decile temperature events. With these points discarded in addition to those already lost from the E_{FAR} plot, there remains insufficient data availability to correlate E_{DBLP} using the area-pool method. If this method were replaced with the fitting of the region's events to a function such as a generalized extreme value distribution, this would alleviate the problem, but this would also necessitate a change to every other calculation. Consequently, this is left as the possible subject of future studies.

7. Conclusions

By assuming that SREX regions are sufficiently homogeneous that the statistics of an area pool are equivalent to that of an ensemble averaged over that region, this paper shows that single reliability diagrams provide some indication of the accuracy of an estimate of fraction attributable risk (FAR). However, if instead, pre- and post-climate-change reliability diagrams are produced, they can be used to provide a much better indication of whether the modeled statistics of the all-forcings and natural worlds taken together provide attribution results consistent with the observed reality. An improved indication of uncertainty can also be achieved by calculating a FAR based upon climatological probabilities to compare with that derived from models. Using imperfect model experiments, this study shows that both techniques are good at representing the errors in modeled FAR that would be found were it possible to observe, in the present, a parallel world with only natural climate forcings. It is therefore recommended to always compare FAR (or a similar metric), which has been derived from a climate simulation, to its counterpart produced from observed climatology, as it is indicative of its grounding in reality. The use of pairs of reliability diagrams is also recommended to provide individual qualitative assessment of both the all-forcings and the naturally forced model against observations. Work is under way to provide an even closer quantitative relationship between these diagrams and the uncertainties within FAR, to enable the use of reliability in recalibration in a manner similar to that used for seasonal forecasting (e.g., Doblas-Reyes et al. 2005), and thus to apply it to provide tailored observation-based uncertainty estimates on specific event attribution studies.

Through the studies in this paper, 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.

Acknowledgments. Many thanks to Dáithí A. Stone for the initial inspiration for this study and to Omar Bellprat, Nikos Christidis, Andrew Ciavarella, Matthew Mizielinski, Colin P. Morice, Adam A. Scaife, Geert Jan van Oldenborgh, and others on the EUCLEIA project for discussions throughout its gestation. Further thanks go to the reviewers of this paper for the improvements suggested. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement 607085. This work was supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101).

REFERENCES

Allen, M. R., 2003: Liability for climate change. *Nature*, **421**, 891–892, doi:10.1038/421891a.

Angélil, O., D. A. Stone, and P. Pall, 2014: Attributing the probability of South African weather extremes to anthropogenic greenhouse gas emissions: Spatial characteristics. *Geophys. Res. Lett.*, **41**, 3238–3243, doi:10.1002/2014GL059760.

Annan, J. D., and J. C. Hargreaves, 2010: Reliability of the CMIP3 ensemble. *Geophys. Res. Lett.*, **37**, L02703, doi:10.1029/2009GL041994.

Bellprat, O., and F. Doblas-Reyes, 2016: Attribution of extreme weather and climate events overestimated by unreliable climate simulations. *Geophys. Res. Lett.*, **43**, 2158–2164, doi:10.1002/2015GL067189.

Christidis, N., and P. A. Stott, 2014: Change in the odds of warm years and seasons due to anthropogenic influence on the climate. *J. Climate*, **27**, 2607–2621, doi:10.1175/JCLI-D-13-00563.1.

——, ——, A. A. Scaife, A. Arribas, G. S. Jones, D. Copsey, J. R. Knight, and W. J. Tennant, 2013: A new HadGEM3-A-based system for attribution of weather- and climate-related extreme events. *J. Climate*, **26**, 2756–2783, doi:10.1175/JCLI-D-12-00169.1.

Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—II. Calibration and combination. *Tellus*, **57A**, 234–252, doi:10.1111/j.1600-0870.2005.00104.x.

——, ——, ——, and J.-J. Morcrette, 2006: Impact of increasing greenhouse gas concentrations in seasonal ensemble forecasts. *Geophys. Res. Lett.*, **33**, L07708, doi:10.1029/2005GL025061.

Hansen, G., M. Auffhammer, and A. R. Solow, 2014: On the attribution of a single event to climate change. *J. Climate*, **27**, 8297–8301, doi:10.1175/JCLI-D-14-00399.1.

Hegerl, G. C., and Coauthors, 2010: Good practice guidance paper on detection and attribution related to anthropogenic climate change. IPCC Rep., 9 pp. [Available online at http://www.ipcc-wg2.gov/meetings/EMs/IPCC_D%26A_GoodPracticeGuidancePaper.pdf.]

Hughes, I. G., and T. P. A. Hase, 2010: *Measurements and Their Uncertainties*. Oxford University Press, 160 pp.

IPCC, 2012: *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*. Cambridge University

Press, 582 pp. [Available online at http://www.ipcc-wg2.gov/SREX/images/uploads/SREX-All_FINAL.pdf.]

King, A. D., S. C. Lewis, S. E. Perkins, L. V. Alexander, M. G. Donat, D. J. Karoly, and M. T. Black, 2013: Limited evidence of anthropogenic influence on the 2011–12 extreme rainfall over southeast Australia [in "Explaining Extreme Events of 2012 from a Climate Perspective"]. *Bull. Amer. Meteor. Soc.*, **94** (9), S55–S58.

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

Lott, F. C., P. A. Stott, and N. Christidis, 2013: Can the 2011 East African drought be attributed to human-induced climate change? *Geophys. Res. Lett.*, **40**, 1177–1181, doi:10.1002/grl.50235.

——, M. Gordon, R. J. Graham, A. A. Scaife, and M. Vellinga, 2014: Reliability of African climate prediction and attribution across timescales. *Environ. Res. Lett.*, **9**, 104017, doi:10.1088/1748-9326/9/10/104017.

Maidment, R., D. Grimes, R. P. Allan, E. Tarnavsky, M. Stringer, T. Hewison, R. Roebeling, and E. Black, 2014: The 30 year TAMSAT African rainfall climatology and time series (TARCAT) data set. *J. Geophys. Res. Atmos.*, **119**, 10 619–10 644, doi:10.1002/2014JD021927.

Pall, P., T. Aina, D. A. Stone, P. A. Stott, T. Nozawa, A. G. J. Hilberts, D. Lohmann, and M. R. Allen, 2011: Anthropogenic greenhouse gas contribution to flood risk in England and Wales in autumn 2000. *Nature*, **470**, 382–385, doi:10.1038/nature09762.

Palmer, T. N., F. J. Doblas-Reyes, A. Weisheimer, and M. J. Rodwell, 2008: Toward seamless prediction calibration of climate change projections using seasonal forecasts. *Bull. Amer. Meteor. Soc.*, **89**, 459–470, doi:10.1175/BAMS-89-4-459.

Rotstayn, L. D., S. J. Jeffrey, M. A. Collier, S. M. Dravitzki, A. C. Hirst, J. I. Syktus, and K. K. Wong, 2012: Aerosol- and greenhouse gas-induced changes in summer rainfall and circulation in the Australasian region: A study using single-forcing climate simulations. *Atmos. Chem. Phys.*, **12**, 6377–6404, doi:10.5194/acp-12-6377-2012.

Scaife, A. A., C. Buontempo, M. Ringer, M. Sanderson, C. Gordon, and J. F. B. Mitchell, 2009: Toward seamless prediction: Calibration of climate change projections using seasonal forecasts. *Bull. Amer. Meteor. Soc.*, **90**, 1549–1551, doi:10.1175/2009BAMS2753.1.

Stott, P. A., and Coauthors, 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., Springer, 307–337, doi:10.1007/978-94-007-6692-1.

Tarnavsky, E., D. Grimes, R. Maidment, E. Black, R. P. Allan, M. Stringer, R. Chadwick, and F. Kayitakire, 2014: Extension of the TAMSAT satellite-based rainfall monitoring over Africa and from 1983 to present. *J. Appl. Meteor. Climatol.*, **53**, 2805–2822, doi:10.1175/JAMC-D-14-0016.1.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2011: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, doi:10.1175/BAMS-D-11-00094.1.

Van Oldenborgh, G. J., 1999: KNMI Climate Explorer. [Available online at http://climexp.knmi.nl.]

——, 2007: How unusual was autumn 2006 in Europe? *Climate Past*, **3**, 659–668, doi:10.5194/cp-3-659-2007.

——, F. J. Doblas-Reyes, S. S. Drijfhout, and E. Hawkins, 2013: Reliability of regional climate model trends. *Environ. Res. Lett.*, **8**, 014055, doi:10.1088/1748-9326/8/1/014055.

Voldoire, A., and Coauthors, 2011: The CNRM-CM5.1 global climate model: Description and basic evaluation. *Climate Dyn.*, **40**, 2091–2121, doi:10.1007/s00382-011-1259-y.

Wilks, D. S., 2011: Statistical forecasting. *Statistical Methods in the Atmospheric Sciences*, 3rd ed. Elsevier, 215–300.

World Meteorological Organization, 2002: Standardised verification system (SVS) for long-range forecasts (LRF): New attachment II-9 to the *Manual on the GDPS*. WMO Rep. 485, 23 pp. [Available online at http://clima1.cptec.inpe.br/gpc/pdf/svs.pdf.]