AGU PUBLICATIONS

Geophysical Research Letters

RESEARCH LETTER
10.1002/2015GL067189

Key Points:
- Model inadequacy can be accounted in event attribution studies using the notion of reliability
- Unreliable model ensembles overestimate attributable risk of extreme events
- Ensemble calibration ensures correct simulated probabilities of extreme events

Supporting Information:
- Supporting Information S1

Correspondence to:
O. Bellprat,
omar.bellprat@bsc.es

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

Received 26 NOV 2015
Accepted 16 FEB 2016
Accepted article online 19 FEB 2016

Attribution of extreme weather and climate events overestimated by unreliable climate simulations

Omar Bellprat[1,2] and Francisco Doblas-Reyes[1,2,3]

[1]Institut Català de Ciències del Clima (IC3), Barcelona, Spain, [2]Barcelona Supercomputing Center, Barcelona, Spain, [3]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain

Abstract Event attribution aims to estimate the role of an external driver after the occurrence of an extreme weather and climate event by comparing the probability that the event occurs in two counterfactual worlds. These probabilities are typically computed using ensembles of climate simulations whose simulated probabilities are known to be imperfect. The implications of using imperfect models in this context are largely unknown, limited by the number of observed extreme events in the past to conduct a robust evaluation. Using an idealized framework, this model limitation is studied by generating large number of simulations with variable reliability in simulated probability. The framework illustrates that unreliable climate simulations are prone to overestimate the attributable risk to climate change. Climate model ensembles tend to be overconfident in their representation of the climate variability which leads to systematic increase in the attributable risk to an extreme event. Our results suggest that event attribution approaches comprising of a single climate model would benefit from ensemble calibration in order to account for model inadequacies similarly as operational forecasting systems.

1. Introduction

Extreme weather and climate events are of general public concern due to their vast socioeconomic impacts. Understanding the causes that have led to event is crucial, particularly as studies are increasingly showing that certain extreme events are becoming more frequent under climate change [*Allen*, 2003]. While it is not possible to entirely attribute a single extreme weather and climate event to either anthropogenic or natural causes, it is possible to evaluate how the odds to experience an extreme event have changed due to the influence of an external driver [*Allen*, 2003; *Stott and Allen*, 2004]. Addressing this question has been an active area of recent research using different approaches [*Pall et al*., 2011; *Van Oldenborgh et al*., 2014; *Yiou and Cattiaux*, 2013; *King et al*., 2013; *Christidis et al*., 2013; *Schaller et al*., 2014], commonly carried out on recent extreme weather events (e.g., extreme precipitation and flooding), which are also equally applicable to extreme climate events (e.g., hot summers season or a warming hiatus).

An event attribution statement is formed by computing the probability that an extreme event occurs in the world as we observe it and in a counterfactual natural world that excludes an external driver, e.g., the one leading to climate change. In the remainder of the study these two probabilities describe one world with all radiative forcings (P_{ALL}) and one with only natural forcings (P_{NAT}), keeping in mind that an event attribution can be carried out on other phenomena too. The probabilities are usually simulated with a climate model where atmospheric radiative forcing can be controlled. A few studies discount the external driver by detrending the observed evolution of the climate [*Van Oldenborgh et al*., 2014] or by considering early observed periods of the last century where climate change was arguably small [*Yiou and Cattiaux*, 2013]. By comparing the derived probabilities (P_{ALL}, P_{NAT}) one can estimate how human influence has altered the risk to observe an extreme event. This is often expressed using the fraction of attributable risk (FAR $= 1 - P_{NAT}/P_{ALL}$) which measures how much of the event can be attributed to human influence from a probabilistic point of view.

Values of FAR are arguably uncertain given the complexity of extreme events and the intrinsic uncertainty that arises from estimating probabilities in the extreme tails of a distribution. Previous studies have so far considered uncertainty in FAR by accounting for the fact that only a limited sample of model simulations are available (using resample techniques when estimating probabilities) [e.g., *Christidis et al*., 2013; *Bellprat et al*., 2015]. Model inadequacy, the aspect that the models can be systematically erroneous in simulating probabilities of an extreme event, has so far been considered by using multimodel ensembles [*Fischer and*

Knutti, 2015; *Bellprat et al*., 2015] or multimethod approaches [*Schaller et al*., 2014; *Otto et al*., 2015]. However, climate models share common deficiencies, e.g., in the description of land-surface coupling important for the variability of extreme summer temperatures [*Fischer et al*., 2012; *Bellprat et al*., 2013] or resolving small-scale convection [*Ban et al*., 2015] important for flooding events. Bias correction methods that correct the model variability have been proposed [*Sippel et al*., 2016], yet model evaluation remains a scarce practice in current event attribution studies [*Herring et al*., 2015]. As a consequence, the implication of model limitations on event attribution results is poorly understood. This argument does not apply to statistical attribution approaches [*Van Oldenborgh et al*., 2014; *Yiou and Cattiaux*, 2013] which are by construction bias-free.

The accuracy with which a model simulates probabilities to exceed a threshold can be measured with the notion of reliability, a common practice in weather and climate forecasting. Model reliability quantifies the agreement of simulated probabilities to exceed an event (e.g., rain $>$ 100 mm) with the observed frequencies in the past using a model hindcast (retrospective simulation). Reliability therefore measures how accurate FAR using a specific climate model is, since FAR builds on simulated probabilities to exceed a threshold. Although the concept of model reliability arises from weather and climate forecasting it does not measure whether a model has actual skill in predicting a certain event, which is not the aim of an event attribution. Reliability merely measures how accurate the probability to exceed a threshold is simulated. It is in this sense an integrative measure that evaluates the model variability over different time scales and statistical moments.

Estimating model reliability is difficult due to the limited length of hindcasts and available observations of past extreme events [*Weisheimer and Palmer*, 2014]. Since the effect of reliability on FAR is arguably statistical in nature, we propose here to study the relationship systematically in an idealized framework. The framework relies on a signal-plus-noise toymodel using Gaussian statistics described in *Weigel et al*. [2008] and *Siegert et al*. [2015]. It allows performing large numbers of simulations on generated observations for very long hindcasts. The model is here further extended to perform an event attribution on artificial observations. The first part of the study presents the statistical model, how it simulates worlds with and without climate change and how model reliability varies with model error. In a second step, an attribution exercise is carried out on sets of extreme events where the system is forced to have different levels of model reliability.

2. Synthetic Hindcasts With and Without Climate Change

2.1. Generating Synthetic Hindcasts

The statistical toy model is presented in *Weigel et al*. [2008] and is here extended to include a long-term component. The model mimics the main aspects of a climate hindcast: a predictable component of an observable, a model error, a perturbation that generates an ensemble, and a long-term trend.

The hindcast is constructed to simulate an observed time series (x_t) with length (T) by sampling Gaussian variability (x_t') with unit standard deviation and zero mean, superimposed on a linear trend (st) with zero intercept,

$$\begin{aligned} x_t &= x_t' + st, \\ x_t' &\sim N(0, \sigma_x = 1), t \in [0, T], \end{aligned} \tag{1}$$

where t defines the time step for the hindcast period T and s is the slope of the linear trend, similarly to *Rahmstorf and Coumou* [2011]. Based on the generated observations, a synthetic hindcast (y_t) can be specified,

$$\begin{aligned} y_t &= \alpha x_t' + \in_\beta + st + \left(\in_{1,\ldots,} \in_M\right), \\ \in_\beta &\sim N(0, \beta), \\ \in_{1,\ldots,M} &\sim N\left(0, \sigma_M = \sqrt{\left(\sigma_x - \alpha^2 - \beta^2\right)}\right), \end{aligned} \tag{2}$$

where α represents the predictable fraction of the observed anomaly (x_t') and β the standard deviation of the model error (ε_β). The predictability is kept low ($\alpha = 0.1$) throughout the study to mimic skill in hindcasts that for instance are forced by observed sea surface temperatures [*Pall et al*., 2011; *King et al*., 2013; *Christidis et al*., 2013; *Schaller et al*., 2014], yet the presented results will be independent of the level of predictability. The model has a zero-mean bias but has an error component conditional to a certain point in time. The individual hindcast members are generated with an ensemble spread (σ_M) such that the total variability of the hindcast

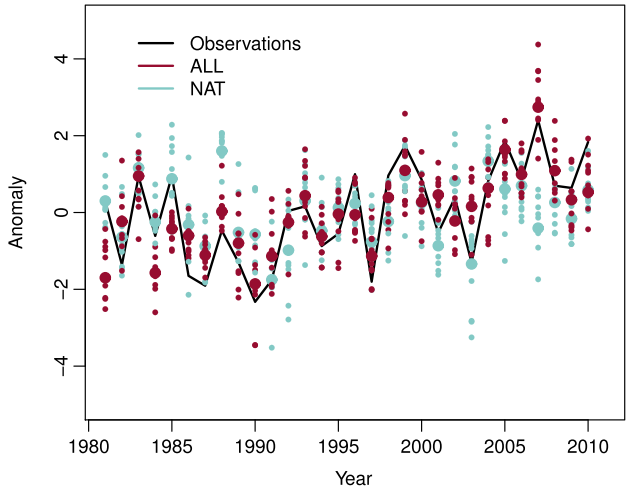

Figure 1. Example of a synthetic hindcast and the corresponding observations for a period of 30 years showing the anomalies from the entire period. The hindcast show the individual members (small dots, 15 members), the ensemble mean (large dot) for a hindcast with a long-term trend emulating an external forcing due to climate change (ALL, red) and a hindcast with the same model parameters without a trend describing a world without climate change (NAT, green). The ratio of the trend (S) and the residual variability (V) is 1.5 using the model parameters $\beta = 0.7$ and 0.1.

is equal to the total variability of the observations. The linear trend of the model is assumed to be perfectly simulated, an assumption that is not fulfilled in current models at a regional scale [*van Oldenborgh et al.*, 2013]. However, for simplicity we continue by defining only one source of model error while keeping in mind that climate models are unreliable not only because of their inadequacy to simulate short-term variability but also due to an erroneous long-term response to an external forcing [*van Oldenborgh et al.*, 2013].

The consideration of the linear trend allows to emulate a hindcast with climate change (hereafter, y_{ant}) including a nonstationarity term analogous to, e.g., increasing temperatures or monotonic changes in precipitation extremes. Consequently, a hindcast with the same model error but with no trend (y_{nat}) describes a counterfactual (counterfactual in the sense that it shares the same interannual variability) world without climate change. An example of a pair of these hindcasts is given in Figure 1 for a typical hindcast length of $T = 30$ years. The hindcast considers 15 ensemble members, which is typical ensemble size in climate hindcasts used to evaluate the model system for different conditions in the past [*Christidis et al.*, 2013]. An event attribution is consequently performed on a single artificially generated event with an ensemble size of 10,000 members. A large ensemble is required to accurately estimate probabilities of a rare event. The idealized model is not designed to detect and attribute a long-term trend [*Hegerl and Zwiers*, 2011] but to assess how the probability of an single extreme event differs in model with external forcing (the one with the linear trend) from the one without.

2.2. Variation of Model Reliability

In a first step, we explore how model error reflects on the model reliability in the simple model. A common way to illustrate the model reliability is using a reliability diagram as shown in Figure 2a. The diagram shows the binned simulated probabilities and corresponding observed frequencies of having an observation above the upper tercile with respect to the climatology. A system is perfectly reliable when the simulated probabilities match the observed frequencies, i.e., when the points lie on the diagonal (black line).

The example shows that the system chosen is too confident, simulating the occurrence of upper tercile events too often. This apparent deviation from the diagonal can be measured in different ways. A traditional measure is the reliability component of the Brier score [*Brier*, 1959], which measures the weighted squared deviations between the points and the diagonal. An alternative and more recent measure presented in *Weisheimer and Palmer* [2014] defines reliability by the weighted linear regression through the points. The example shows this linear regression with a dashed line and an uncertainty estimate by resampling the hindcast. One advantage of this measure is that the slope of the regression scales the reliability (R) between $R = 1$ for a perfect reliable hindcast, $R = 0$ for unreliable (overconfident) hindcast similar to other verification scores, regardless of the number of bins considered. It is important to note that ensemble overconfidence can arise from a model bias, i.e., an erroneous shift of the ensemble away from the observation or from a weak perturbation of the ensemble resulting in low ensemble spread. Both effects will lead to an ensemble that does not entail the observations.

Varying the standard deviation of the model error leads to different levels of R as shown in Figure 2b. Increasing the error reduces R until having no reliability for $\beta = 0.99$. The uncertainty of R is large as shown

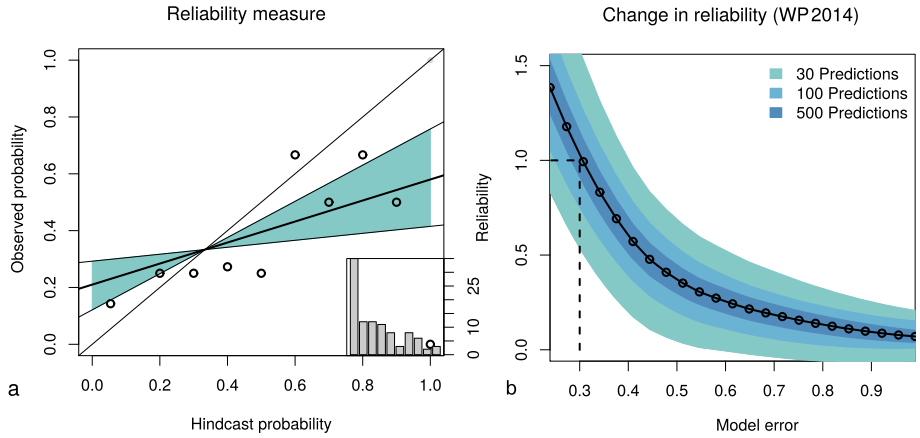

Figure 2. (a) Reliability diagram for a synthetic hindcast (100 years and 15 members) for simulating events above the upper tercile. The observed probabilities (relative frequencies, circles) are binned (10 bins) for a range of probabilities. The number of hindcasts years for each bin is shown in the lower right histogram, also known as sharpness diagram. The black dashed line shows a linear weighted regression with its 25–75% confidence level in agreement with *Weisheimer and Palmer* [2014]. The model parameters are the same as in Figure 1. (b) Reliability measured as the slope of the regression line in Figure 2a as a function of the standard deviation of the model error (β). The black line shows the median estimate by repeating the hindcasts 1000 times with different number of hindcast length (30, 100, and 500 years). Increasing length reduces the uncertainty in reliability as shown by the width of the colored areas. The ensemble size is chosen to be 1000 to avoid a systematic underestimation of the ensemble spread [*Ferro*, 2013].

by the colored areas for low numbers of hindcast years but can be reduced substantially in the idealized model by considering 500 model years (blue dark area). The variation of R as a function of β is robust also when considering the reliability component of the Brier score (see Figure S1 in the supporting information) and when considering different parameter values (see Figure S2). The level of β is hereafter interpreted as inversely proportional to the reliability. However, we decide against fitting R as an inverse function of the model error to remain general with respect to the choice of the reliability measure.

The reader will note that perfect reliability ($R = 1$) is reached at a certain threshold with β larger than zero, which arises from the condition that the ensemble spread needs to sample the total model error for a model to be reliable [*Slingo and Palmer*, 2011]. Given the definition of the ensemble spread (equation (2)) an optimal level of β (in the sense of perfect reliable) arises that depends on the level of the predictability (see Text S1). Values of β below this level (dashed line) are therefore underconfident ($R > 1$), indicating that the ensemble spread is larger than the variance of the model error.

3. Reliability and Attribution of an Extreme Event

The model described in the previous section has the advantage that its model error can be varied. This allows to explore systematically the link between the model reliability and an event attribution. We consider for this purpose an extreme event in the observations (x_{EX}) that would occur given the climatology every 10 and 50 years. This artificial "extreme" event is consequently simulated with the model including a linear trend (world with climate change, y_{ant}) and one without a trend (world without climate change, y_{nat}). We assume here that the predictability of the extreme event (x'_{EX}) is the same in both worlds regardless of whether climate change has occurred or not (which is equivalent to assuming that climate change does not affect the predictability of extremes beyond the conditioning provided by the trend) and assume also that the model error conditional to this event is the same in both hindcasts.

Using the two hindcasts, we can compare the probability of the event in the two worlds by computing the cumulative probability that $P_{NAT} = \mathrm{Prob}(F_{NAT} > x_{EX})$ and $P_{NAT} = Prob(F_{NAT} > x_{EX})$. The reader will note that the threshold is the same in both cases. The change in the probability is expressed as

$$\mathrm{FAR} = 1 - P_{NAT}/P_{ALL} \tag{3}$$

Values of FAR > 0 denote that the event has become more likely due to climate change. We consider here only examples where FAR is positive, the same arguments are though valid for negative values of FAR. The

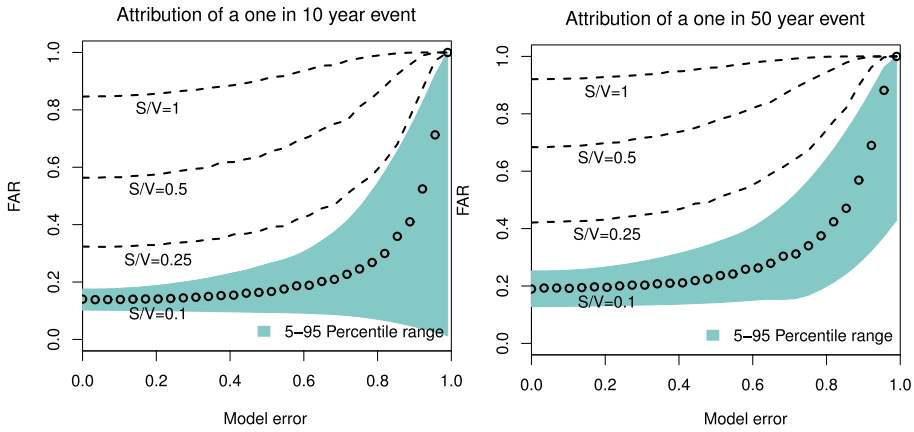

Figure 3. Variation of FAR for a (a) 1-in-10 year and (b) 50 year event as a function of the standard deviation of the model error (β) which is inversely proportional to the model reliability (see Figure 2). The variation of FAR is shown for different signal (S, trend) to variability (V, residual variability from the trend) ratios where the dots show the number of different levels of β computed. The green area shows the 5–95 % confidence interval sampled by repeating the predictions 1000 times allowing the model error to vary within the chosen standard deviation (β).

values of FAR for the two events and for different levels of model error (and hence model reliabilities) are shown in Figure 3 as a median value and an uncertainty estimate using 10,000 ensemble members and repeating the attribution 1000 times. The values of FAR are shown for different ratios of the external signal ($S = sT$, the total change due to the trend) and the level of natural variability (V, residual variability from the trend). Values of FAR increase with increasing model error, i.e., low reliability. The increase is particularly strong when the S/V ratio is small, relevant for events occurring at small scales [*Sippel and Otto*, 2014; *Schaller et al*., 2014]. Along with the systematic increase of FAR the uncertainty of FAR (green shade) increases, taking any value between 0 and 1 for low model reliability.

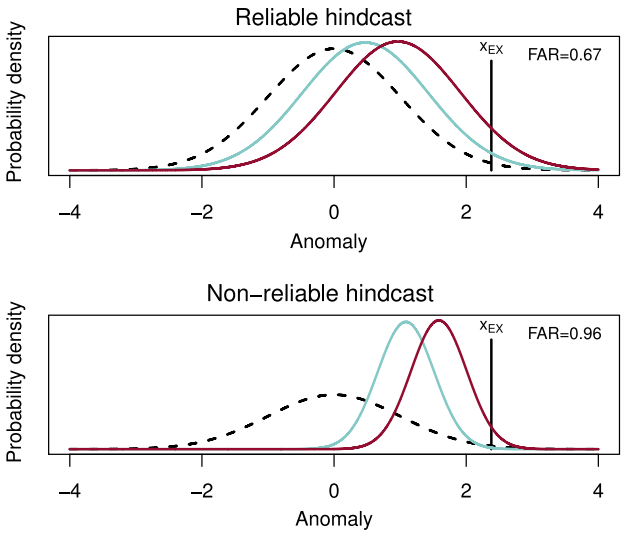

Figure 4. Illustration of an event attribution in an example of a reliable ($\beta = 0.3$) and an unreliable hindcast ($\beta = 0.95$) for the 1-in-50 year event. The probability density function with the dashed black line shows the climatology of the natural world, together with the hindcast of the world without (green) and with climate change (red). The extreme event threshold (x_{EX}) is shown as a thick vertical line. The center of the hindcast distribution is defined by the predictability, the model error, and the externally forced signal. The value of the FAR is shown for each hindcsat in the upper right corner, which is also graphically indicated by the ratio of the areas of the predictions above the threshold.

In order to explain the systematic increase in FAR we select one illustrative example (Figure 4) for the one in 50 years event using a configuration of high reliability and low reliability, respectively. The hindcasts (y_{ANT}, y_{NAT}) are shown as two distributions, separated mainly by the external signal (S). The unreliable hindcast has a too narrow (also referred as overconfident) ensemble spread. This is the typical condition of unreliability in current generation of climate models [*Weisheimer and Palmer*, 2014]. The overconfident ensemble underestimates the probabilities that the event will occur, both in the actual and counterfactual hindcast. Values of FAR increase because P_{NAT} decreases at a higher pace than P_{ANT} with lower ensemble spread. This result can be extended to other types of distributions with heavier tails such as a generalized extreme value distribution (GEV; see Figure S2a). We also find that the result is robust when

considering increased variability in the climate change world as could be expected for temperature [*Schär et al*., 2004, Figure S3b] and when attributing a model quantile instead of an observed (fixed) threshold (Figure S3c).

To demonstrate the relevance of the result in practice, we show how reliability affects FAR in a physical climate model ensemble developed under the EUropean Climate and weather Events: Interpretation and Attribution (EUCLEIA) project (http://eucleia.eu) to establish an event attribution prototype [*Christidis et al*., 2013]. Reliability cannot be varied in a such physical model ensemble for a given region and type of events, yet the reliability can be corrected. Ensemble calibration, a common practice in weather and climate forecasting [*Doblas-Reyes et al*., 2005], is one way to achieve this. Using a technique known as ensemble inflation [*von Storch*, 1999], the model ensemble variability is corrected to achieve high reliability (shown for a seasonal temperature example in Figure S4). The attributable risk of an extreme event that would occur once in 30 years at each grid point [*Fischer and Knutti*, 2015] is consequently computed using the raw and calibrated model output. Values of FAR systematically decrease over many regions in the globe after the calibration (Fig. S5), which confirms the obtained result in the toy model framework. Note that the change in FAR is substantial in the example (up to −0.8 FAR) which illustrations the relevance of reliability in event attribution studies.

4. Discussion and Conclusions

Current generation of climate models imperfectly simulate extreme events due to limitations of model resolution and erroneous representation relevant physical mechanisms. Its implication when attributing an extreme event is largely unknown, a gap which is here filled by relying on the notion of model reliability. Model reliability measures how accurate a model ensemble simulates the probability of an extreme event and thus quantifies the model uncertainty of the fraction attributable risk (FAR). Using an idealized model framework, we find that unreliable models are prone to overestimate FAR due to overconfident ensemble spread, a common and well-known deficit of current climate model systems [*Slingo and Palmer*, 2011]. This result is valid for other types of distributions and independent whether the model is conditioned (e.g., with fixed SSTs) on the event. Note that ensemble overconfidence is a consequence of weak perturbation of model physics to generate an ensemble and due to model bias; both arguments are interchangeable.

The study suggests that event attribution approaches using single climate model would benefit from ensemble calibration [*Doblas-Reyes et al*., 2005] and other bias correction approaches [*Sippel et al*., 2016] in order to avoid systematic overestimation of FAR. Calibration ensures that the model ensemble variability at different temporal scales follows the one observed, including the variability arising from a long-term trend (the response to an external forcing). These trends are known to be deficient in current models on regional scales [*Van Oldenborgh et al*., 2013], and although the implication of incorrect trends has not been explored in this study, we illustrate that ensemble calibration is also an elegant solution to that particular problem. Model calibration could hence become a standard in event attribution studies in order to consider model limitations.

However, statistical correction of model ensembles remains a challenge due to the small number of observed extreme events. Ensemble calibration and bias correction approaches are therefore uncertain by itself, and its uncertainty should be propagated onto the estimate of FAR as well (as proposed in the supporting information). The consideration of physical constraints of how biases evolve [e.g., *Bellprat et al*., 2013] in bias correction methods may further aid to overcome the sample limit when correcting climate model extremes as recently proposed in *Sippel et al*. [2016]. Ultimately, the conclusion that can be drawn from this study is that climate models tend to overestimate attribution results and that future studies should increasingly consider model correction approaches in order to account for model uncertainties.

Acknowledgments
We would like to acknowledge valuable discussions and feedback received from François Massonnet, Nathalie Schaller, Chloé Prodhomme, and Fraser Lott. This work was supported by the EUropean CLimate and weather Events: Interpretation and Attribution (EUCLEIA), funded by the European Union's Seventh Framework Programme [FP7/2007-2013] under grant agreement 607085 and the ESA Living Planet Fellowship Programme under the project VERITAS-CCI. We are further indebted to the s2dverification (http://cran.r-project.org/web/packages/SpecsVerification/index.html) and specs-verification (http://cran.r-project.org/web/packages/s2dverification/index.html) with which the calculations have been carried out. The synthetic hindcast generator has been implemented into s2dverfiction. No further data were used in producing this manuscript

References

Allen, M. (2003), Liability for climate change, *Nature*, *421*, 891–892.
Ban, N., J. Schmidli, and C. Schär (2015), Heavy precipitation in a changing climate: Does short-term summer precipitation increase faster? *Geophys. Res. Lett.*, *42*, 1165–1172, doi:10.1002/2014GL062588.
Bellprat, O., S. Kotlarski, D. Lüthi, and C. Schär (2013), Physical constraints for temperature biases in climate models, *Geophys. Res. Lett.*, *40*, 4042–4047, doi:10.1002/grl.50737.
Bellprat, O., F. C. Lott, C. Gulizia, H. R. Parker, L. A. Pampuch, I. Pinto, A. Ciavarella, and P. A. Stott (2015), Unusual past dry and wet rainy seasons over Southern Africa and South America from a climate perspective, *Weather Clim. Extremes*, doi:10.1016/j.wace.2015.07.001.
Brier, G. W. (1950), Verification of forecasts expressed in terms of probability, *Mon. Weather Rev.*, *78*(1), 1–3.

Christidis, N., P. A. Stott, 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. Clim.*, *26*, 2756–2783.

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 A*, *57*, 234–252, doi:10.1111/j.1600-0870.2005.00104.

Ferro, C. A. T. (2014), Fair scores for ensemble forecasts, *Q. J. R. Meteorol. Soc.*, *140*(683), 1917–1923.

Fischer, E. M., and R. Knutti (2015), Anthropogenic contribution to global occurrence of heavy-precipitation and high-temperature extremes, *Nat. Clim. Change*, *5*(6), 560–564.

Fischer, E. M., J. Rajczak, and C. Schär (2012), Changes in European summer temperature variability revisited, *Geophys. Res. Lett.*, *39*, L19702, doi:10.1029/2012GL052730.

Hegerl, G., and F. Zwiers (2011), Use of models in detection and attribution of climate change, *WIREs Clim. Change*, *2*, 570–591, doi:10.1002/wcc.121.

Herring, S. C., M. P. Hoerling, J. P. Kossin, T. C. Peterson, and P. A. Stott (2015), Explaining extreme events of 2014 from a climate perspective, *Bull. Am. Meteorol. Soc.*, *96*(12), S1–S172.

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. Am. Meteorol. Soc.*, *94*(9), S55–S58.

Otto, F. E. L., et al. (2015), Factors other than climate change, main drivers of 2014/15 water shortage in Southeast Brazil, *Bull. Am. Meteorol. Soc.*, *96*(12), S35–S40.

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.

Rahmstorf, S., and D. Coumou (2011), Increase of extreme events in a warming world, *Proc. Natl. Acad. Sci. U.S.A.*, *108*, 17,905–17,909.

Schaller, N., F. E. L. Otto, G. J. van Oldenborgh, N. R. Massey, S. Sparrow, and M. R. Allan (2014), The heavy precipitation event of May–June 2013 in the upper Danube and Elbe basins, *Bull. Am. Meteorol. Soc.*, *95*(9), S69–S72.

Schär, C., P. L. Vidale, D. Lüthi, C. Frei, C. Häberli, M. A. Liniger, and C. Appenzeller (2004), The role of increasing temperature variability in European summer heatwaves, *Nature*, *427*(6972), 332–336.

Siegert, S., D. B. Stephenson, P. G. Sansom, A. A. Scaife, R. Eade, and A. Arribas (2015), A Bayesian framework for verification and recalibration of ensemble forecasts: How uncertain is NAO predictability? *J. Clim.*, *29*, 995–1012.

Sippel, S., and F. E. L. Otto (2014), Beyond climatological extremes—Assessing how the odds of hydrometeorological extreme events in South-East Europe change in a warming climate, *Clim. Change*, *125*(3–4), 381–398, doi:10.1007/s10584-014-1153-9.

Sippel, S., F. E. L. Otto, M. Forkel, M. R. Allen, B. P. Guillod, M. Heimann, M. Reichstein, S. I. Seneviratne, K. Thonicke, and M. D. Mahecha (2016), A novel bias correction methodology for climate impact simulations, *Earth Syst. Dyn.*, *7*, 71–88, doi:10.5194/esd-7-71-2016.

Slingo, J., and T. N. Palmer (2011), Uncertainty in weather and climate prediction, *Philos. Trans. R. Soc. A*, *369*(1956), 4751–4767.

Stott, P. A., and M. R. Allen (2004), Human contribution to the European heatwave of 2003, *Nature*, *432*, 610–614.

van Oldenborgh, G. J., F. Doblas-Reyes, S. S. Drijfhout, and E. Hawkins (2013), Reliability of regional climate model trends, *Environ. Res. Lett.*, *8*(1), 014055.

van Oldenborgh, G. J., R. Haarsma, H. de Vries, and M. R. Allen (2014), Cold extremes in North America vs. mild weather in Europe: The winter 2013/2014 in the context of a warming world, *Bull. Am. Meteorol. Soc.*, doi:10.1175/BAMS-D-14-00036.1.

von Storch, H. (1999), On the use of inflation in statistical downscaling, *J. Clim.*, *12*, 3505–3506.

Weigel, A. P., M. A. Liniger, and C. Appenzeller (2008), Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts?, *Q. J. R. Meteorol. Soc.*, *134*, 241–260.

Weisheimer, A., and T. N. Palmer (2014), On the reliability of seasonal climate forecasts, *J. R. Soc. Interface*, doi:10.1098/rsif.2013.1162.

Yiou, P., and J. Cattiaux (2013), Contribution of atmospheric circulation to wet North European summer precipitation of 2012, *Bull. Am. Meteorol. Soc.*, *93*, 1054–1057.