Weather and Climate Extremes 9 (2015) 25–35

Contents lists available at ScienceDirect

Weather and Climate Extremes

journal homepage: www.elsevier.com/locate/wace

Combining large model ensembles with extreme value statistics to improve attribution statements of rare events

Sebastian Sippel[a,*], Dann Mitchell[b], Mitchell T. Black[c], Andrea J. Dittus[c], Luke Harrington[d], Nathalie Schaller[b], Friederike E.L. Otto[b]

[a] Max Planck Institute for Biogeochemistry, Hans-Knöll-Str. 10, 07745 Jena, Germany
[b] Environmental Change Institute, University of Oxford, South Parks Road, Oxford OX1 3QY, UK
[c] School of Earth Sciences and ARC Centre of Excellence for Climate System Science, The University of Melbourne, Parkville, Victoria, Australia
[d] New Zealand Climate Change Research Institute, School of Geography, Environment and Earth Sciences, Victoria University of Wellington, Wellington 6012, New Zealand

ARTICLE INFO

Article history:
Received 23 December 2014
Received in revised form
15 June 2015
Accepted 16 June 2015
Available online 18 June 2015

Keywords:
Extreme value statistics
Ensemble modelling
Extreme weather events

ABSTRACT

Gaining a better understanding of rare weather events is a major research challenge and of crucial relevance for societal preparedness in the face of a changing climate. The main focus of previous studies has been to apply a range of relatively distinct methodologies to constrain changes in the odds of those events, including both parametric statistics (extreme value theory, EVT) and empirical approaches based on large numbers of dynamical model simulations.

In this study, the applicability of EVT in the context of probabilistic event attribution is explored and potential combinations of both methodological frameworks are investigated. In particular, this study compares empirical return time estimates derived from a large model ensemble with parametric inferences from the same data set in order to assess whether statements made about events in the tails are similar. Our analysis is illustrated using a case study of cold extremes and heavy rainfall in winter 2013/14 in Europe (focussing on two regions: North-West Russia and the Iberian Peninsula) for a present-day (including 'anthropogenic' influences) and an alternative 'non-industrial' climate scenario.

We show that parametric inferences made about rare 'extremes' can differ considerably from estimates based on large ensembles. This highlights the importance of an appropriate choice of block and sample sizes for parametric inferences of the tails of climatological variables. For example, inferences based on annual extremes of daily variables are often insufficient to characterize rare events due to small sample sizes (i.e. with return periods >100 years). Hence, we illustrate how a combination of large numerical simulations with EVT might enable a more objective assessment of EVT parameters, such as block and sample size, for any given variable, region and return period of interest.

By combining both methodologies, our case study reveals that a distinct warming of cold extremes in winter has occurred throughout Europe in the 'anthropogenic' relative to the non-industrial climates for given sea surface temperatures in winter 2013/14. Moreover, heavy rainfall events have become significantly more frequent and more pronounced in North and North-East Europe, while other regions demonstrate no discernible changes.

In conclusion, our study shows that EVT and empirical estimates based on numerical simulations can indeed be used to productively inform each other, for instance to derive appropriate EVT parameters for short observational time series. Further, the combination of ensemble simulations with EVT allows us to significantly reduce the number of simulations needed for statements about the tails.

1. Introduction

It is a major scientific challenge to better understand extreme meteorological events and potential changes in the odds of their occurrence in a warming climate (Seneviratne et al., 2012; Zhang et al., 2014). This is due to a number of reasons, including limitations of the observational record to capture rare extreme events, and issues of data availability and quality. Moreover, structural and parametric model uncertainties, as well as the proverbial chaotic

* Corresponding author.
E-mail address: ssippel@bgc-jena.mpg.de (S. Sippel).

http://dx.doi.org/10.1016/j.wace.2015.06.004

nature of weather (Lorenz, 1963) hinder any straightforward attribution of causality between climatic drivers and any particular extreme weather event.

To overcome these difficulties, many scientific studies use either one of the following approaches:

First, *extreme value theory* (EVT) has been developed to provide a means to model the tails of statistical distributions based on mathematical theory (Coles et al., 2001). Such an analysis allows statistical statements to be made based on parametric extreme value distributions (see Wigley, 2009, for illustrative examples). For example, scientific assessments have been made to investigate trends in temperature and precipitation extremes in the 21st century in atmosphere–ocean coupled models (Kharin and Zwiers, 2000; Kharin et al., 2007, 2013), allowing the estimation of return levels and their associated statistical uncertainties. Further illustrative applications of the (univariate) EVT framework elucidate causes for geophysical extremes, such as the connection between atmospheric modes of variability and cold extremes (Sillmann et al., 2011). However although EVT is increasingly used in climatological studies to constrain the odds of rare events (Katz, 2010), including extensions to account for non-stationarity, multivariate and spatial extremes (see Ghil et al., 2011, for a review), Katz et al. (2013) argues that its full potential has not yet been tapped for many geophysical applications.

Second, an alternative approach to improve the understanding of extremes and their changing odds in a non-stationary climate has been to deploy very large ensembles of dynamical models, namely *probabilistic event attribution* (PEA, Stone and Allen, 2005; Allen, 2003). This methodology is used extensively to sample rare events and subsequently estimate their probabilities under different climate forcing scenarios (Stott et al., 2004; Otto et al., 2012; Massey et al., 2014). The latter often serves to estimate the anthropogenic contribution (fraction of attributable risk) to changes in the meteorological risk of present-day weather and climate extremes (Allen, 2003; Stott et al., 2013; Bindoff et al., 2013; Christidis et al., 2013). Importantly, an assessment of this type addresses the odds of specific extreme weather events – often those that had happened in a particular year such as droughts, heat waves or cold spells (Herring et al., 2014). Notable extensions to the PEA methodology include the attempt to account for more impact-related variables, for instance through a coupling with hydrological models to assess floods (Pall et al., 2011). Nonetheless, PEA assessments are typically based on rather data-intensive empirical estimates of return times, and rely to a large extent on dynamical model simulations.

Our study addresses the following research questions:

1. Is the statistical framework of EVT applicable in the context of a probabilistic assessment of extreme events? Accordingly, can both methodological frameworks be productively combined to inform each other?
2. Using a *combined* methodology, how have meteorological extremes at daily time scales in the European winter of 2013/14 changed relative to a pre-industrial climate?

Based on our first research question, we envision an application in which both methodological frameworks could inform each other in order to (a) derive insights about appropriate parameter choices (i.e. required sample and block sizes) for the application of statistical models based on EVT for the meteorological characteristics of any variable or region of interest; and (b) given informed parameter choices, how many numerical simulations are actually needed to estimate a given 'target return period' to a satisfactory degree of accuracy?

Hence, our study details a joint assessment of both methodologies and evaluates whether statements made about the tails of meteorological variables such as temperature and precipitation are comparable. This methodological comparison might serve as a starting point to reconcile the two statistical frameworks for climatological applications, i.e. to inform each other about relevant parameter choices (EVT) or the number of samples needed to estimate a specific return level. To illustrate this comparison and to address the second research question, a large ensemble of atmosphere-only regional climate simulations for the European 2013/14 winter season is investigated as a case study along with a 'non-industrial' climate scenario of winter 2013/14 (i.e. with anthropogenic forcings removed Schaller et al., 2014, see Section 2).

The particular season of interest, winter 2013/14 in Europe, provides an interesting case study, because it came along with exceptionally mild temperatures, severe storm depressions, both winter dryness and heavy precipitation on regional to sub-continental scales. Significant but diverse societal impacts were associated with those events, for instance exceptionally early vegetation greening and a reduction of fossil fuel consumption for heating due to the absence of severe frosts in some regions.[1] Seasonal temperatures ranked among the highest ever recorded in a range of countries according to national weather services (e.g. Austria, Denmark, France, Germany, the Netherlands, Norway, Poland, Slovakia, Switzerland, and the UK, e.g. Fig. 1, Deutscher Wetterdienst, 2014). When it comes to seasonal rainfall anomalies, a remarkable east–west divide persisted over most of the winter, where central and south-eastern parts of the continent received exceptionally low rainfall, whereas its most western stretches, such as Ireland and the UK, experienced a record wet season (Huntingford et al., 2014, Fig. 1). These remarkable patterns resulted from a synoptic situation with many storm depressions that moved along the English Channel, over the British Isles and into the North Sea, hence advecting warm air into Central and East European regions, and causing rainfall and severe winds in Britain and along the Atlantic coast. This synoptic situation is also reflected by seasonal geopotential height anomalies (Fig. 1), which were strongly negative over the North Atlantic and the British Isles, whereas positive anomalies prevailed over Eastern Europe (see Huntingford et al., 2014 for a more detailed discussion).

This study's analysis focuses on cold temperature and heavy rainfall extremes, which allows us to state how the odds of occurrence of extremes in these two variables have changed between a 'non-industrial' climate and the contemporary winter climate in 2013/14. These two variables provide a good case study, because we expect temperature to be relatively spatially coherent, and precipitation to be somewhat noisier both in space and time. We illustrate our methodological approach as well as the attribution analysis for two spatially averaged regions, North-West Russia and the Iberian Peninsula, as well as for the entire European model domain.

In Section 2, we describe the experimental setup, evaluate and bias-adjust the regional climate model and outline the statistical methodology to estimate return times. In Section 3, we first outline the results of the methodological comparison (EVT vs. empirical return time estimates), and discuss related issues such as parameter choices for a potential combination of both methodologies. Second, the illustrative attribution case study of winter minimum temperatures and precipitation is presented. Lastly, we draw some conclusions about the applicability of EVT based return time estimates in the context of probabilistic event attribution (Section 4).

[1] http://www.pecad.fas.usda.gov/highlights/2014/03/EU_12march2014/

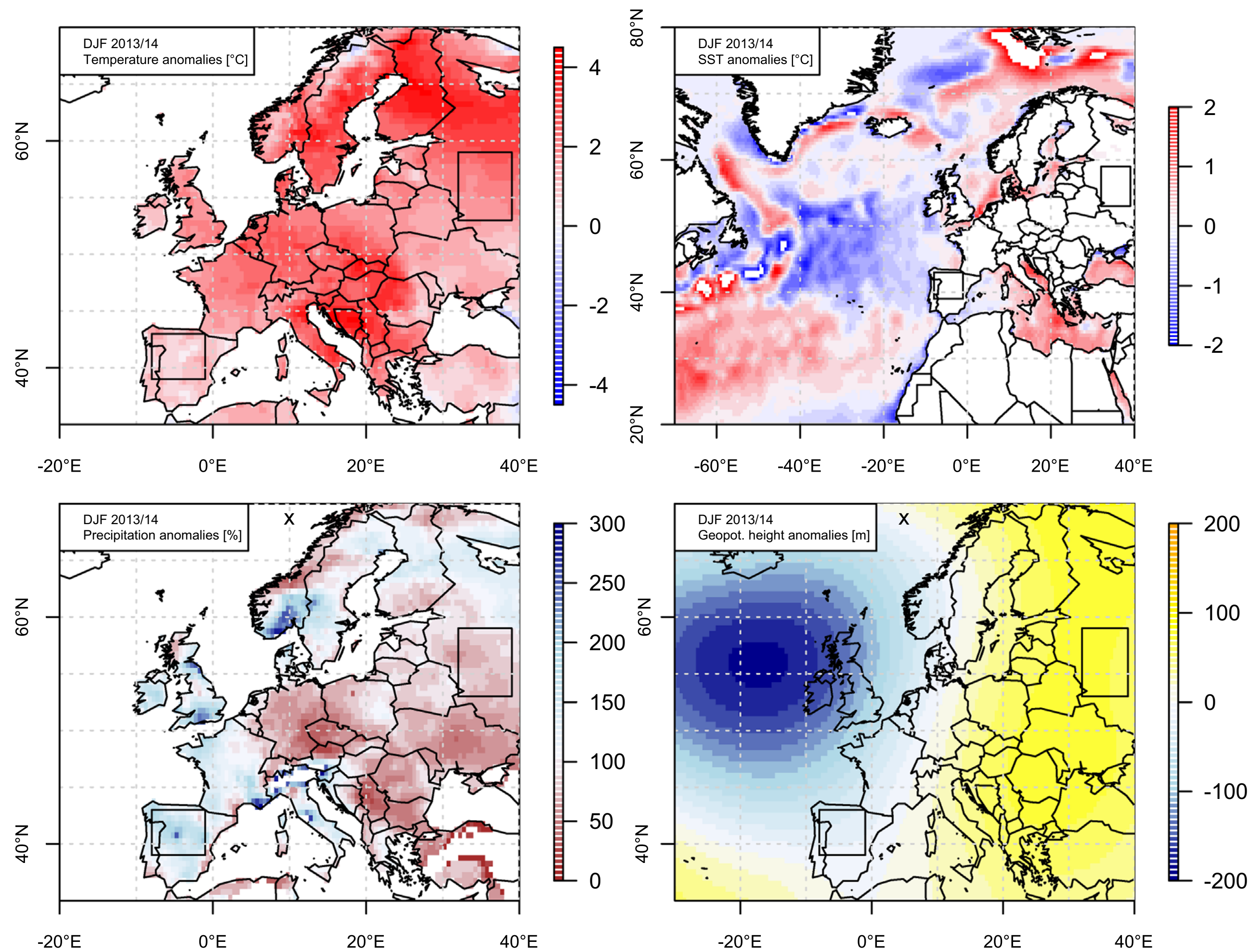

Fig. 1. Synoptic analysis of DJF 2013/14 in Europe: seasonal temperature anomalies (top left), SST anomalies (top right), anomalies in cumulative rainfall (bottom left), and geopotential height anomalies (bottom right). Temperature and precipitation data were taken from E-OBS, SSTs and geopotential height anomalies were calculated from ERA-Interim (reference period: 1981–2010). The study regions over Spain and Russia are drawn as rectangular boxes.

2. Materials and methods

Model structure and experiment setup. In this study, we analyze large ensemble simulations of the HadAM3P atmosphere-only, global circulation model with an embedded, identically formulated regional model for Europe (HadRM3P), which has been used extensively elsewhere (Jones, 2004; Massey et al., 2014). The global (nested regional) models are run with a spatial resolution of $1.875° \times 1.25°$ ($0.44° \times 0.44°$) on a rotated grid identical to the EURO-CORDEX region,[2] with 19 vertical levels and a temporal resolution of 15 (5) min (Massey et al., 2014). The model is based on the atmospheric component of the HadCM3 general circulation model (see Pope et al., 2000 for a full description) with improvements with respect to the calculation of clouds and convection, and a more realistic coupling of vegetated surfaces with the soil (Massey et al., 2014). Since atmosphere-only simulations were conducted, observed sea surface temperatures (SSTs) and sea ice fractions for the observed period (DJF 2013/2014) are provided to the model from the Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) dataset (Stark et al., 2007; Donlon et al., 2012). Further model drivers include the observed atmospheric composition (CO_2, CH_4, N_2O, halocarbons and ozone), natural and anthropogenic emissions of different sulfur species, and solar anomalies (see Massey et al., 2014 for a more detailed description and evaluation of the modelling framework). Initial conditions are perturbed in the global circulation model on 1st December for each ensemble member (ibid.).

As observed SST patterns from 'the world that might have been' (i.e. the 'non-industrial' scenario) in the absence of anthropogenic emissions are not known precisely, estimates are made using some of the state-of-the-art coupled ocean–atmosphere models taken from the Coupled Model Intercomparison Project, phase 5 (CMIP5, Taylor et al., 2012). Eleven of these models have run 'natural forcings only' simulations of the historical climate, and these are subtracted from the 'all forcings' simulations to obtain an estimate for the change in the SSTs (hereafter, delta SSTs). The differencing is performed on climatological monthly means over the last decade available, i.e. 1996–2005. The delta SSTs are then used to change the observed SSTs accordingly. To sample uncertainty, we use these different CMIP5 models which cover the main modelling groups from around the world (see Schaller et al., 2014, for details). All non-anthropogenic forcings such as aerosols, volcanoes and the solar cycle are kept constant in both scenarios.

The large model ensemble investigated in this paper is derived through the weather@home framework, in which citizen scientists donate idle computer time in order to perform computationally intensive calculations in a distributed manner. This approach provided model ensemble simulations for 13,260 DJF periods in an industrial world and 22,129 for the non-industrial case, using the variety of different SST reconstructions. Data preprocessing

[2] http://www.euro-cordex.net/About-Euro-Cordex.1864.0.html

Table 1
Regions used in this study and their geographical boundaries.

Region	Eastern boundary (°E)	Western boundary (°E)	Southern boundary (°N)	Northern boundary (°N)
Spain	−8	−1	39	43
NW Russia	32	39	53	59

consists of regridding the regional model ensembles to a regular 0.5° grid over Europe, using a second order conservative remapping scheme (Jones, 1999). Subsequently, a meteorological sanity check is conducted, in which all ensemble members with meteorologically implausible values are removed, before we validate our model and analyze the ensemble's statistics of extremes (see below). We derive both 1-day spatial averages over selected regions and 5-day grid-based averages for minimum temperatures and heavy precipitation, where both aggregation steps are computed on the original gridded time series. Two regions were chosen to represent different European climates with predominantly maritime/Mediterranean (Spain) and continental (North-West Russia) influences (Table 1). The grid-based European-scale analysis is noisier due to a lower level of aggregation, but nevertheless provides valuable spatially explicit details. Due to model spinup time, the first two weeks of December are disregarded, after which it was checked that no remainder spinup effects are detectable.

Model validation and bias adjustment: A high-quality grid-based European land-only observational data set in 0.5° resolution (EOBS, version 10.0, Haylock et al., 2008) is used in order to quantify biases in simulated meteorological variables and to conduct a simple synoptic assessment for winter 2013/14 (Fig. 1). Since our ensemble is based on the SST patterns of one winter season, an assessment of model performance would not be representative based on the ensemble alone. Hence, we use 50 randomly chosen ensemble members per year (i.e. 1300 model years in daily resolution) for the winter seasons from 1986 to 2010 from an identical model setup (Massey et al., 2014) for the purpose of validation. Differences in statistical distributions are assessed graphically by quantile–quantile plots. The spread of the ensemble is illustrated similar to Massey et al. (2014) by appending randomly chosen ensemble members without replacement in order to derive 50 winter time series for each of the years 1986–2010 (see Fig. 1 in Supplementary Information, S1).

The model's winter simulations of daily temperature show relatively good agreement with the distribution of daily minimum temperatures in E-OBS in both regions, although with a slight cold bias over NW Russia. For the whole European domain, larger biases are observed in Scandinavia, and towards the southern margins of the regional model domain (S1). Nonetheless, it can be noted that the regional model performs better in simulating temperatures in winter as compared to summer (Massey et al., 2014). Hence, we conclude that our model simulates temperatures to a reasonable degree, and this also holds for percentiles relatively far away from the mean (Fig. 1, S1).

Precipitation simulations do not always agree favourably with observations. Considerable wet biases towards the upper tails of the distributions of daily rainfall over the two regions remain, as well as for most grid cells throughout Europe (S1). Here, we use a very simple bias adjustment methodology to account for this bias. Due to the obvious positivity constraint, an additive correction of biases, which is often applied to climatic variables such as temperature (Hempel et al., 2013; Sippel and Otto, 2014), is not feasible for precipitation. Hence, we determine a multiplicative correction factor similar to Hempel et al. (2013), which quantifies biases in the 97.5th percentile:

$$c = \frac{OBS_{97.5th}}{MOD_{97.5th}} \tag{1}$$

Subsequently, daily rainfall values are scaled by c, which removes some of the biases in the high percentiles. Although using a single percentile is a somewhat subjective choice, we argue that it is relatively robust with respect to the observations, since in the period used for model validation (DJF 1986–2010), the 97.5th percentile corresponds approximately to the 50th largest value, hence a relatively robust sample. This simple multiplicative adjustment yields a better match of simulated and observed rainfall amounts also in higher quantiles, without any invasive changes to the distribution. Importantly however, scaling the absolute values with an adjustment factor does not affect any relative changes between fitted extreme value distributions.

Further, it is important to note that an acceptable simulation of daily precipitation statistics does not warrant satisfying simulation at monthly or seasonal time scales. For an evaluation and discussion for model performance at monthly time scales, we refer the interested reader to Massey et al. (2014). Moreover, while we acknowledge that the resolution of the regional model is too coarse to resolve local convection or thunderstorm-related activity, Fig. 1, S1 demonstrates that the distribution of daily rainfall events in the model agrees broadly with the observations for both regions, including its tails (S1).

Statistical estimates of return periods: The primary objective of this study is to compare statistical inferences for the tails of meteorological variables based on EVT with empirical return time estimates. The ensemble simulations are conducted for one season only (DJF 2013/14 in a 'natural' and 'anthropogenic' scenario), hence stationarity for the EVT based estimates of the tails is assumed. Further, we fit generalized extreme value (GEV) distributions of the form (Coles et al., 2001)

$$G(z) = \exp\left(-\left[1 + \xi\frac{z-\mu}{\sigma}\right]^{-(1/\xi)}\right), \tag{2}$$

to a sample of 1-day (5-day) minimum temperatures and maximum cumulative rainfall events for each simulated winter season for each area-averaged region (grid cell). Here, μ, σ and ξ denote the location, scale and shape parameter of the GEV distribution, respectively. Unless otherwise stated, confidence intervals representing 5–95% parametric uncertainty are given based on the normality of the GEV parameter estimates (Coles et al., 2001). To address the influence of GEV parameter choice (block and sample size) on the return time estimates (Section 3), we resample the large ensemble to derive different block and sample sizes for various return time estimates. This procedure is iteratively repeated for each parameter combination in order to derive resampling based 5–95% confidence intervals for return time estimates that are comparable to the empirical estimates.

For the analysis of rare winter extremes, a resampling strategy is used in order to avoid biases associated with an extrapolation from 1-yr extremes to several hundred year return level extremes (see Section 3), which might also entail a very different dynamical structure of the atmospheric circulation in the real world. Therefore, 10-yr block extremes are drawn from the large sample by a random selection of ten ensemble members, from which only the most extreme value is retained. This procedure is repeated 200 times (for both regions and for each CMIP5 model's SST reconstruction) to derive a statistical distribution of 10-year block extremes, which is subsequently used to fit a GEV distribution as specified above.

Throughout our analysis, a Generalized Maximum Likelihood Estimation (GMLE) approach is used for fitting the parametric model to the data (Martins and Stedinger, 2000), which is conducted using the extRemes software package (Gilleland and Katz, 2011). We also tested the GEV parameter estimation using the L-moment and MLE methods: these were found to yield estimates very similar to the GMLE method that we employ here. All statistical analysis is performed in the R statistical environment (R Development Core Team, 2008) using the add-on packages 'boot' and 'ADGofTest'. A guide to the here-deployed resampling strategy, including an illustration of the effects of parameter choices on return time estimates (Section 3.1) is provided in the Supplementary Information.

Empirical return time estimates are constructed by plotting the sorted values of the ensemble against its rank. To assess uncertainty of this empirical estimate, we derive bootstrapped uncertainty intervals (5–95%) by resampling ($n=5000$ ensemble members, $R=1000$ times).

Evaluation of fitted extreme value distributions: The parametric fits are evaluated in a three-fold approach:

First, we use adjusted mean residual life plots (Coles et al., 2001) in order to test whether the exceedance of any threshold u yields an approximately linear scaling of the 'residual means' (i.e. the average of the values exceeding the threshold u). This concept is frequently used to determine an appropriate threshold for peak over threshold models with prior declustering of extremes. It can be shown that the residual means follow a linear function of the threshold, if the peak over threshold model is appropriate (Coles et al., 2001, p. 79). Here, this idea is slightly modified, and we plot the 'mean residual life' of the seasonal block maxima, thus it could be seen similar to a seasonal declustering approach (i.e. assuming that any two extreme events in one season are not independent). Present non-linearities in these plots might indicate that extreme events are subject to different physical/dynamical climatic regimes, and will be further discussed/evaluated below.

Second, each fitted GEV (both regional and grid cell based) is tested for its goodness of fit using a parametric Anderson–Darling (AD) test based on a significance level of $\alpha=5\%$. We chose the AD test over a Kolmogorov–Smirnov (KS) test used in earlier studies (Kharin et al., 2007), because it is more sensitive to the tails of the distribution by implementing a weight function instead of a maximum distance approach such as the KS test.

Lastly, in order to evaluate deviations of the fitted GEV distributions from the empirical large ensemble for rare events in the tails (Section 3.1), we adopt a somewhat ad hoc but practically useful definition of 'biases' (see Fig. 2 and associated discussion): since our focus is on 'rare events', we determine the maximum absolute difference in return levels in the interval of 100–1000 years (i.e. 99th to 99.9th percentile) between the fitted GEV's and the empirical return levels of the large ensemble, using monotonic Hermite spline interpolation to derive a continuous curve for the latter. To compare the biases in GEV fits from the empirical ensemble with the 'expected biases' inherent in any GEV model for a given block size (Section 3.1), we simulate a large number of random values from the fitted extreme value distributions for each region. Subsequently, we determine the distance ('bias', as defined above) between GEV fits from this data using each block size of interest from a 'large empirical GEV sample' ($n=15\,000$). Hence, these artificial simulations mimic the comparison between the empirical ensemble and GEV fits with different block sizes. The uncertainty of an empirical estimate of the tail is tested by resampling from a known GEV model (S2), the variance of which becomes large for very high return periods.

3. Results and discussion

In this section, we first test a combination of stationary EVT analysis with a large ensemble of numerical simulations and present a systematic evaluation of the parameter choices in EVT-based assessments regarding its effects on return time estimates for meteorological variables (Section 3.1, an illustrative guide is available as Supplementary Material). Subsequently, we analyze changes in cold temperature and heavy precipitation extremes in winter 2013/14 relative to a pre-industrial scenario in the large ensemble simulation using extreme value theory (Section 3.2).

3.1. Combining extreme value analysis with large ensemble simulations

A comparison of the fitted GEV distributions based on resampled sub-ensembles with the empirical estimate of the tail for the NW Russia region is presented in Fig. 2 (top), including a stationary GEV fit to the E-OBS observations for illustrative purposes only (1951–2014, black dots and line).[3]

The methodological comparison reveals that GEV-based inferences with large block sizes (e.g. 10-yr return periods, Fig. 2, top, dark-blue/dark-red line and shading) agree well with the empirical estimate (circles). However, inferences made for shorter return periods (e.g. 1-yr events: orange/light blue) overestimate (minimum temperatures) or underestimate (maximum rainfall) return levels of rare events (e.g. 100+ year return levels). This analysis is presented in for the NW Russia region, and occurs similarly over Spain (S2), although less pronounced. These differences are important to consider, because a relatively large proportion of the GEV's fitted to resampled sub-ensembles ($n=1000$, annual block extremes) is not rejected by a statistical Goodness-of-Fit test,[4] and could thus be misinterpreted if only a small ensemble were available. However, these differences in the inferences about the tails can be readily detected in the mean residual life plots, for example in the NW Russia region (Fig. 2, middle) with a non-linear breakpoint approximately around the median (marked as 50th percentile, corresponding to 2-yr return events). Hence, extreme value statistics of seasonal minimum daily temperatures or precipitation might not be 'rare enough' in order to satisfactorily constrain events that are located far in the tails. This could potentially lead to notorious biases in statistical models, which are most pronounced for large return periods (i.e. 99th to 99.9th percentile in this case) if the chosen block size is too small. Comparing these biases with biases in the tails for independent and identically GEV-distributed artificial data (see Section 2 for a detailed description of the resampling strategy to obtain these 'expected biases in the tails') for any given block size shows that for large enough block sizes these biases are reduced (Fig. 2). Hence, although the ultimate reason for non-adequate statistical model fits for rare events are limited sample and block sizes (Fisher and Tippett, 1928; Coles et al., 2001), characteristics of climatological variables such as serial correlation, climatic variability and noise, or potential dynamic regime changes under extreme conditions might considerably amplify these deviations.

The illustrative example highlights that the choice of block size is critical in climatological applications of EVT. To further and more systematically investigate this issue, we conducted a range of resampling experiments to assess the influence of parameter choice on an EVT-based estimation of climatological events in the extreme tails. Those parameter choices are inherently a trade-off between bias (short block size) and variance (due to small sample sizes for large blocks Coles et al., 2001), which is illustrated in

[3] However, it should be kept in mind that the observations are based on a 63-year period, including potential non-stationarities and cover a variety of synoptic conditions, whereas the model ensemble is run conditional on 2013/14 SST's.

[4] Based on the AD-test, the proportion of the null hypothesis *not rejected* is for the anthropogenic (natural) ensemble: 41% (30%) NW Russia, 100% (100%) Spain (minimum temperatures), and 99% (98%) NW Russia, 99% (100%) Spain (heavy rainfall).

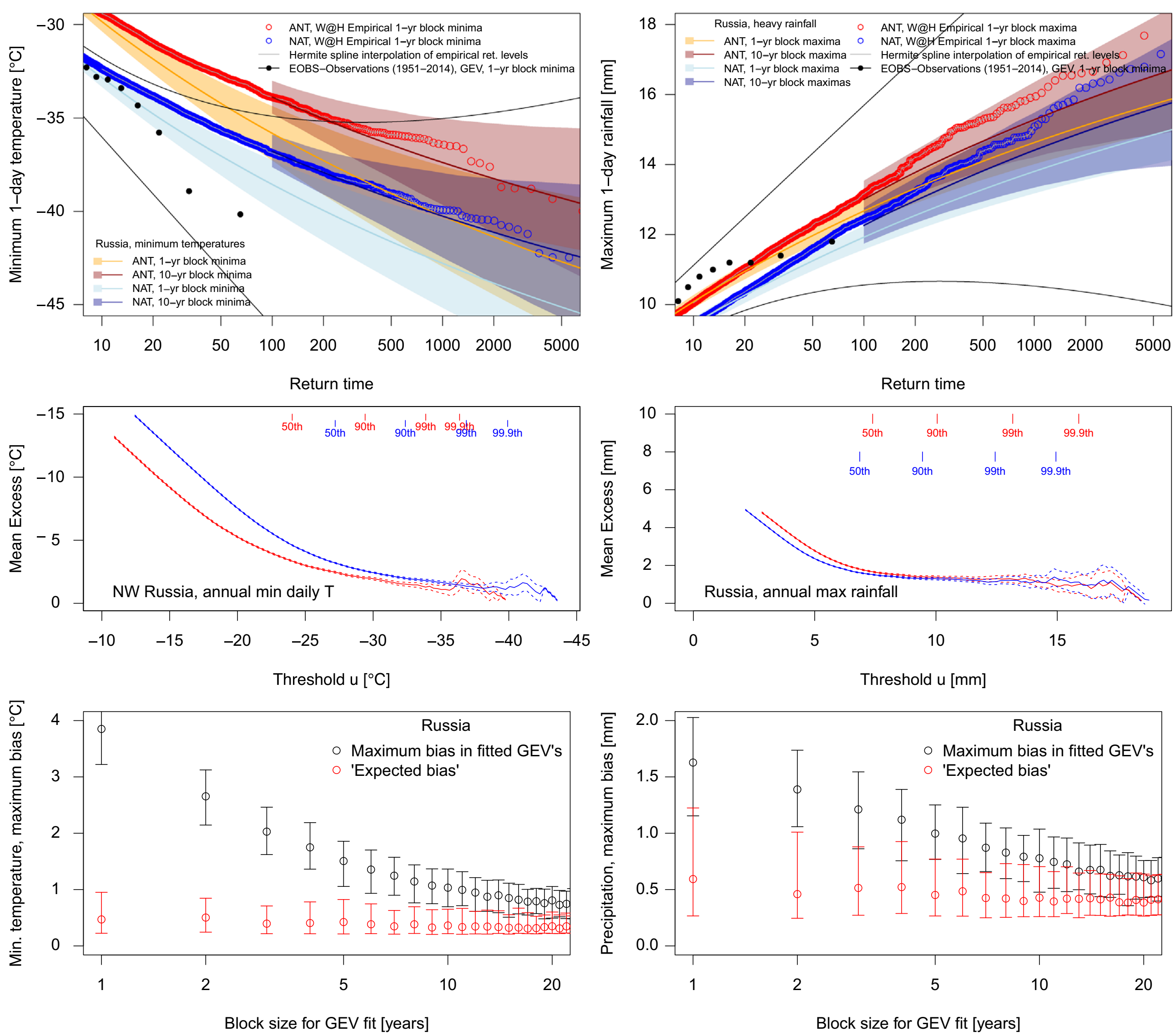

Fig. 2. Return level plots of GEV distributions fitted to 1-yr ($n=1000$) and 10-yr ($n=100$) block extremes of daily minimum temperatures (top left) and heavy precipitation (top right). Coloured dots reflect the 'empirical' large ensemble, observations are denoted in black (EOBS, 1951–2014). Shading represents parametric uncertainty as taken from the fitted generalized extreme value distributions. (Middle panels) Mean residual life plot of annual block minima of daily temperatures (left) and block maxima of daily rainfall (right) for NW Russia. (Bottom) Biases in rare events (100–1000 year return periods, as defined in Section 2) estimated from GEV distributions as a function of block size for the model ensemble and as would be expected by sampling from 'ideal' GEV distributions. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

Fig. 3 for two different return times (20 and 1000-years) in NW Russia. We note that from a practical perspective, for example for the analysis of relatively short observational time series, the choice of block size depends not only on the available sample size and climatological variable of interest, but also on the 'target return time' upon which a statement should be made (Fig. 3, see also tabulated values in S2, and the 'code tutorial' provided as Supplementary Material). To this end, it is interesting to note that these biases require careful consideration if, for example, statistical models are derived based on annual extremes of daily variables, which is widely being done (see for example: Coles et al., 2001). On the other hand however, GEV-based inferences with larger block sizes allow us to derive very consistent statements for high return intervals, for which a reduced number of ensemble simulations are already sufficient (e.g. compare GEV-based inference with an ensemble of size $n=1000$ in Fig. 3 with the empirical estimate, $n=13\,260$).

At this point, a couple of cautionary remarks might be appropriate. First, it should be noted that in this paper we investigate the simplest case of an application of EVT: daily extremes determined from seasonal blocks under stationary conditions (i.e. Winter 2013/14 under anthropogenic *or* natural forcing conditions). Hence, it should be stressed that EVT can also be applied under non-stationary conditions (Kharin and Zwiers, 2000; Kharin et al., 2007, 2013) and with covariates accounting for additional information (see for example Sillmann et al., 2011). Furthermore, peak-over-threshold models constitute an important alternative to modelling block maxima with GEV's (Coles et al., 2001); a detailed investigation of this in terms of informed parameter choices based on ensemble simulations could be a topic for future study.

Second, as we are concerned here about the statistics of rare events, and thus the dynamical structure of such events is not investigated. In Europe, such rare events might be related to relevant modes of atmospheric variability, such as for example the North Atlantic Oscillation (NAO) (Sillmann et al., 2011). Therefore, our analysis and estimation of return periods of extremes is conditional on sea surface temperature patterns that were present in winter 2013/14 in the Euro-Atlantic region with the NAO being in

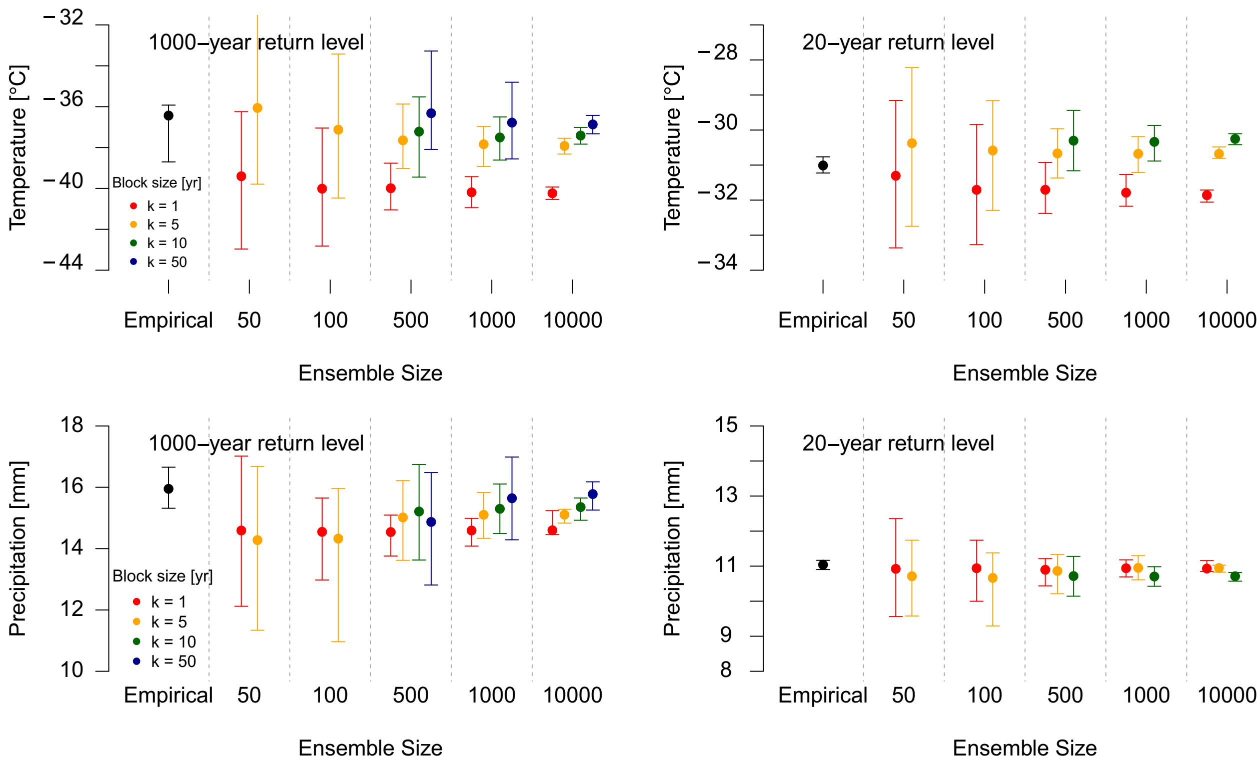

Fig. 3. Illustration of bias–variance trade-offs for estimates of 1000-year (left) and 20-year (right) return time events of daily minimum temperature (top) and maximum precipitation (bottom) in the Russia region for the bootstrapped empirical ensemble and GEV's fitted to resampled ensembles using various sample and block sizes.

its positive phase (Huntingford et al., 2014).

Third, it should be pointed out that an analysis of rare events is inherently uncertain. In this paper, we are addressing statistical (Section 3.1) and scenario reconstruction uncertainties (Section 3.2). Hence, possibly large uncertainties that might stem from the models' (imperfect) structure or parametrization are not examined here, although an attempt was made to implicitly account for such issues using the empirical bias correction for precipitation (Section 2).

To summarize, our analysis reveals that seasonal block extremes in an ensemble of regional model simulations of daily meteorological variables might not be robust enough to infer statistical statements on the odds of particularly rare events of both temperature and precipitation. A resampling scheme is shown to improve the fits to the tails based on larger than annual block sizes. Therefore, the combination of a large number of dynamical model simulations with statistical extreme value models might enable a more informed selection of parameter choices for EVT-based inferences. In return, EVT-based estimates might point at the number of numerical simulations needed to adequately constrain a given return period of interest (Fig. 3). Hence, we conclude that for climatological applications both methodologies might benefit from a statistical setup in which EVT and large numerical simulations inform each other, for example to choose EVT parameters for the analysis of relatively short observational time series.

3.2. The anthropogenic influence on European minimum temperatures and precipitation in winter 2013/14

In this subsection, we present and discuss how climatic changes between the counterfactual scenario and the present might have altered the odds of cold extremes and heavy precipitation as an application of the extreme value analysis outlined above. We also illustrate for two regions how uncertainties in the reconstruction of a counterfactual past might induce uncertainties in attribution statements. Finally, we discuss our results in the context of changes in extremes throughout Europe.

Temperatures: In a winter season such as DJF2013/14 in Europe, minimum temperatures have warmed significantly and unambiguously in both study regions (Fig. 4, S3) and throughout Europe (S4). For example, the location parameter of GEV distributions fitted to 10-yr resampled minimum temperatures in NW Russia has shifted significantly under all scenarios (Fig. 4). However, the reconstruction of a 'non-industrial world' scenario induces considerable uncertainties, with a warming of roughly two and four degrees at the lower and upper end, respectively, of the CMIP5 models used for reconstruction (Fig. 4). Hence, scenario uncertainties are larger in magnitude than statistical uncertainties resulting from fitting statistical models in this type of study. Furthermore, the decreasing odds of extremely cold temperatures in the two regions studied in this paper seem to a very large proportion caused by a shift in the location parameter of the GEV, rather than by changes in the scale or shape of the distribution (S4). In fact, none of the different SST reconstructions shows a significant change in scale or shape in any region under study (not shown), and computing GEV parameters over each grid cell of the European model domain yields only minor and largely non-significant changes in the shape and scale parameters of seasonal cold extremes (S4). This finding indicates that the year-to-year variability of seasonal cold extremes (around the shifting mean) has not changed markedly in our model, though the interpretation of individual GEV parameters is to be made with caution (Gilleland, pers. comm.).

Nonetheless, testing different assumptions about potential changes in the scale and shape of the tails is a highly topical issue in climatology – not least because recent findings point at a decreasing temperature variability at the sub-seasonal scale in northern latitudes (Screen, 2014). Consequently, we further investigate this issue in our model ensemble with a focus on the tails. To do so, we compare the

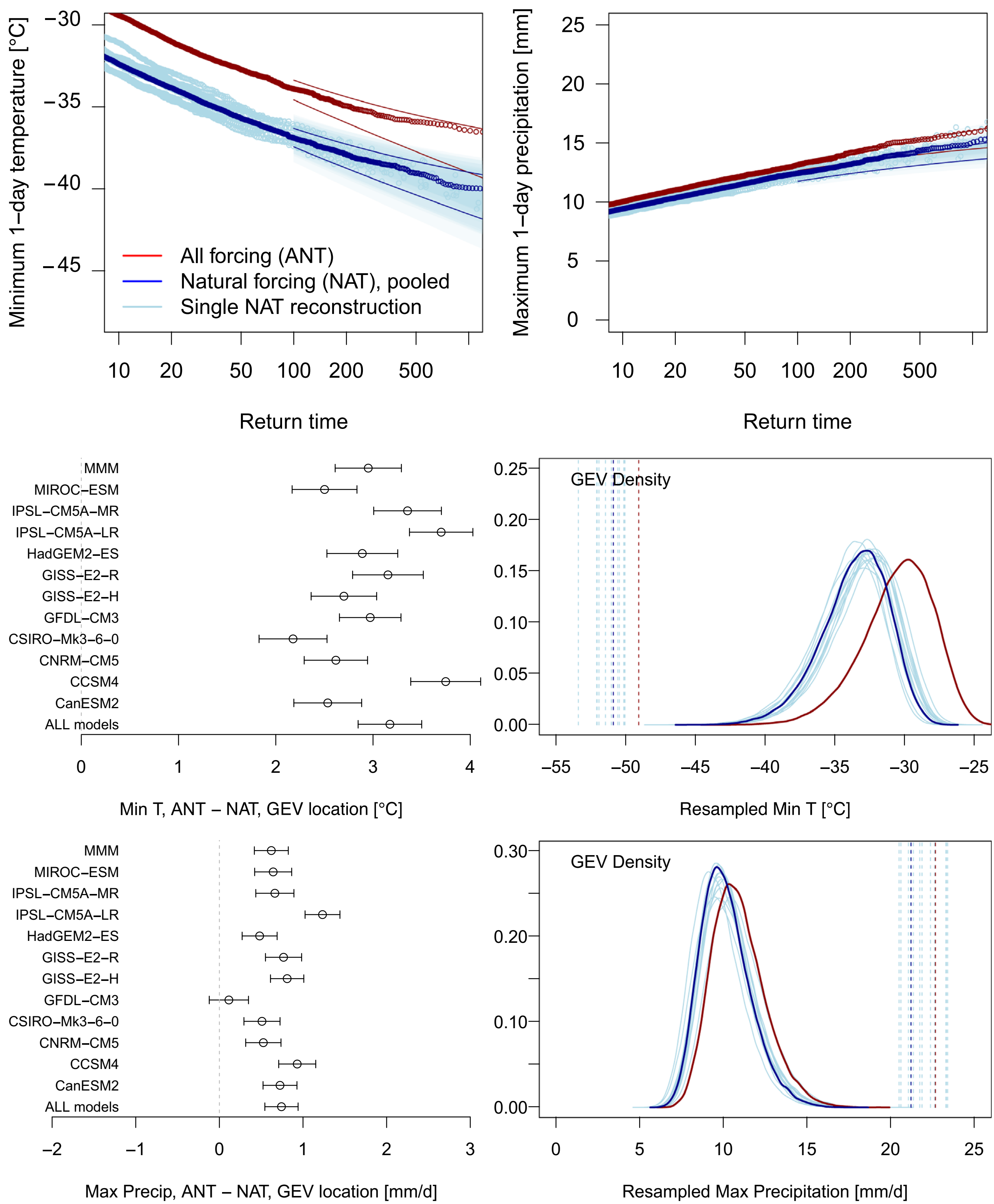

Fig. 4. (Top panel) Return periods of seasonal block minimum temperatures (left) and heavy rainfall (right) in NW Russia. (Middle panel) Warming in the GEV's location parameter (left) and densities of the fitted GEV distributions for 10-yr resampled block minimum 1-d temperatures. (Bottom) Changes in the location parameter (left) and GEV density (right) of 10-yr resampled heavy precipitation events. Changes in the GEV's location parameter (middle and bottom left panel) and single SST reconstructions (turquoise lines) are given for the 11 different natural SST estimates, and the CMIP5 models from which the estimates were obtained are listed in Schaller et al. (2014).

present-day warming relative to the pre-industrial scenario in winter maximum temperatures with the warming in the coldest winter temperatures (i.e. a 'differential warming of winter temperature extremes' is defined as the difference between the warming in the warmest and coldest winter temperatures expressed through 100-year return levels). To this end, we find a clear, spatially coherent and widely significant pattern (Fig. 5): in large areas of North and Central Europe, cold extremes have warmed considerably stronger than warm extremes. Only in the Mediterranean region and towards the eastern edges of our model domain this pattern is not as clearly

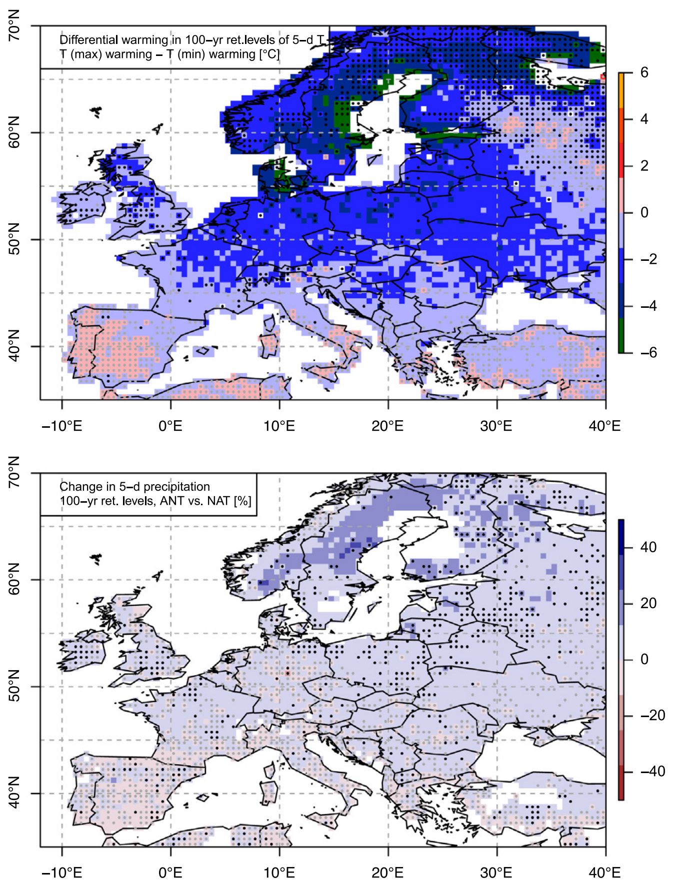

Fig. 5. (Top) Difference in warming of the warm and the cold tail of 100-year return periods of daily winter extremes. (Bottom) Percent changes in 100 year return levels of 5-day rainfall sums in Europe between DJF2013/14 and a counterfactual 'non-industrial' winter season with a similar sea surface temperature pattern. Black dots indicate poor goodness-of-fit as indicated by the AD-test, while grey stippling indicates non-significant changes in return levels.

pronounced. This finding is qualitatively consistent with previous studies that have shown that daily minimum temperatures (nights) are warming faster than maximum temperatures (days) in observations (Alexander et al., 2006; Donat et al., 2013) and that sub-seasonal temperature variability in northern latitudes is decreasing (Screen, 2014), both of which might contribute to the differential warming seen here. Mechanisms behind the day–night asymmetry might indeed include stronger night-time effects of increased greenhouse gas forcing, whereas changes in sub-seasonal variability in northern latitudes might be driven by the Arctic amplification, i.e. temperatures of northerly winds might have warmed faster than southerly winds over the last decades (Screen, 2014). Disentangling these effects is not the focus of this paper, but would provide an interesting topic for further study.

In brief, our analysis suggests that cold and warm temperatures extremes have warmed considerably since pre-industrial times, but the upper and the lower (extreme) tail might indeed warm at different rates. However, our present analysis does not show any evidence that the year-to-year variability of seasonal cold temperature extremes has changed.

Precipitation: When it comes to wintertime heavy rainfall events, changes between the 'non-industrial' (NAT) and anthropogenic (ANT) scenarios are less pronounced and vary among regions and the CMIP5 models used for reconstructing the SST patterns. We find a significant shift towards stronger heavy precipitation events in NW Russia (Fig. 4, bottom), whereas in Spain no significant overall changes are shown by our model (S3). Moreover, the counterfactual world reconstructions clearly show that scenario uncertainty is large when it comes to heavy rainfall: in NW Russia all except one SST reconstructions for the 'non-industrial' scenario lead to a significant increase in the location parameter, however results for Spain show the sign of the location parameter to differ between SST estimates leading to an overall small but non-significant increase in the location parameter. Like the results for temperature, we also observe that the scale and shape parameters are not changing significantly across the studied regions.

To understand further, we derived GEV fits for heavy precipitation for each grid cell of the European model domain (S5). Most attribution studies conducted to date have been looking at regional averages, mainly because spatial (or temporal) aggregation reduces the level of noise. Although we acknowledge that this type of spatially explicit analysis presented here might involve considerable uncertainties, particularly as local features such as processes on a sub-grid cell scale might not be well-represented in the model, we argue that the figure presented in S5 allows us to identify European regions that show a spatially coherent signal of human-induced changes in 100-year return levels of daily rainfall. Whilst the overall pixel-based signal is much noisier and does not point at strongly pronounced changes in extreme winter rainfall in Central or Southern Europe, we are able to identify regions in North and North-East Europe that exhibit a spatially coherent signal of increasing 100-yr return levels (Fig. 5). Here again, those changes can be attributed to a shift in the location parameter of the GEV, rather than changes in scale or shape (S5). In conclusion, we find clear indications that winter rainfall extremes are changing in parts of North Europe, whilst in southern regions, particularly in the Mediterranean no clear statement can be made at present. Although the mechanisms behind intensified extreme rainfall are still debated (O'Gorman and Schneider, 2009), they can be conceptualized as a subtle interplay between thermodynamical effects (i.e. the amount of moisture held within a fixed volume of air, described by the well-known Clausius–Clapeyron relationship, e.g. Held and Soden, 2006) and large-scale atmospheric dynamics in a warming climate (Emori and Brown, 2005). Nonetheless, our results in European regions agree qualitatively well with previous findings of intensified daily rainfall in model simulations for the mid-high latitudes, and relatively minor changes in the extreme percentiles over the Mediterranean (Pall et al., 2007). Likewise, Westra et al. (2013) show that maximum precipitation events at daily time scales are becoming more intense in the observational record for most stations globally, with least pronounced changes occurring in drier sub-tropical regions, such as the Mediterranean.

4. Conclusion

In this study we examined two commonly used techniques for assessing the odds of extreme weather events in a changing climate. The purpose of which was to test if using statistical inferences on relatively small sample sizes (as is common in observational studies) would give quantitatively similar results to using large sample sizes of the simulated climate (as is used in complementary experiments, e.g. Stott et al., 2004; Otto et al., 2012). When it comes to attribution statements it is important to account for such potential differences, because statements are often based on the evaluation of relatively subtle changes in the tails. The analysis is then used to understand how the lower and (upper) tails of temperature (rainfall) distributions change under anthropogenic climate change in a European winter season.

We show that for some regions of Europe, the definition of what counts as 'extreme' data can drastically change how the extreme value distribution (in this case the GEV) is fitted. In some cases, for instance winter temperatures over NW Russia, the GEV model based on an annual block size does not fit well to empirical estimates of the tails from thousands of ensemble members of a climate simulation. The reason for the observed disparity may well be because of different dynamic regimes under very extreme meteorological conditions that do not occur in every seasonal simulation. As such, a careful choice of parameters is crucial when using EVT for understanding extreme events, especially if small sample sizes akin to observational data are used. We argue that large ensemble simulations might offer a route to test the robustness of such parameter choices for any particular variable or region of interest. Further, a combination of GEV-based inference with ensemble simulations allows us to reduce the number of required simulations substantially for estimating high return periods. For example, we show that when analysing extreme temperatures over Russia, a statement regarding the 1000-year return period can be made by fitting an extreme value distribution to a sample size of 1000 years, whereas empirical estimates would require an order of magnitude larger sample size. Similar conclusions can be drawn for Spain.

Using the refined resampling technique for understanding rare extremes and with respect to the case study of the unusual winter 2013/14, we find a widespread warming pattern throughout Europe, which led to a reduction of return periods of very cold winter days (as derived from seasonal minima). This is accompanied by an increase in warm winter anomalies both in frequency and magnitude throughout the model domain. Crucially, the observed warming of daily winter temperature minima is larger than the maxima, showing an asymmetry in the changes in extremes. Finally, predominantly northern parts of Europe show significant increases in unusually extreme daily rainfall events, emphasizing the importance of considering extreme events on a regional basis.

Acknowledgements

We thank Philippe Naveau and three anonymous reviewers for very helpful comments on a previous version of the manuscript.

We acknowledge the World Climate Research Programme (WCRP) and the International Centre for Theoretical Physics (ICTP), which have supported this research through the WCRP-ICTP Summer School on Extremes (2014). Furthermore, we thank the participants of the climate*prediction*.net project, who generated the large ensemble simulations by donating idle computer time. Moreover, we acknowledge the E-OBS dataset from the EU-FP6 project ENSEMBLES (http://ensembles-eu.metoffice.com) and the data providers in the ECA&D project (http://www.ecad.eu), which was used for model evaluation. S. Sippel acknowledges funding provided by the German National Academic Foundation. A.J. Dittus and M.T. Black are supported by Australian Research Council Grant CE110001028.

Appendix A. Supplementary data

Supplementary data associated with this paper can be found in the online version at http://dx.doi.org/10.1016/j.wace.2015.06.004.

References

Alexander, L., Zhang, X., Peterson, T., Caesar, J., Gleason, B., Klein Tank, A., Haylock, M., Collins, D., Trewin, B., Rahimzadeh, F., et al., 2006. Global observed changes in daily climate extremes of temperature and precipitation. J. Geophys. Res.: Atmos. (1984–2012) 111.

Allen, M., 2003. Liability for climate change. Nature 421, 891–892.

Bindoff, N.L., Stott, P.A., AchutaRao, M., Allen, M.R., Gillett, N., Gutzler, D., Hansingo, K., Hegerl, G., Hu, Y., Jain, S., et al., 2013. Detection and Attribution of Climate Change: From Global to Regional. IPCC, Cambridge University Press, Cambridge, United Kingdom; New York, NY, USA.

Christidis, N., Stott, P.A., Scaife, A.A., Arribas, A., Jones, G.S., Copsey, D., Knight, J.R., Tennant, W.J., 2013. A new hadgem3-a-based system for attribution of weather- and climate-related extreme events. J. Clim. 26, 2756–2783.

Coles, S., Bawa, J., Trenner, L., Dorazio, P., 2001. An Introduction to Statistical Modeling of Extreme Values, vol. 208. Springer, London, UK.

Deutscher Wetterdienst, 2014. Seasonal Bulletin on the Climate in WMO Region VI—Europe and Middle East. Technical Report. Deutscher Wetterdienst. URL ⟨http://www.seevccc.rs/SEECOF/SEECOF-11/STEP%201/RCC_CM_DWD_Seasonal ClimReport_2013_2014_DJF.pdf⟩.

Donat, M., Alexander, L., Yang, H., Durre, I., Vose, R., Dunn, R., Willett, K., Aguilar, E., Brunet, M., Caesar, J., et al., 2013. Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century: the hadex2 dataset. J. Geophys. Res.: Atmos. 118, 2098–2118.

Donlon, C.J., Martin, M., Stark, J., Roberts-Jones, J., Fiedler, E., Wimmer, W., 2012. The operational sea surface temperature and sea ice analysis (OSTIA) system. Remote Sens. Environ. 116, 140–158.

Emori, S., Brown, S., 2005. Dynamic and thermodynamic changes in mean and extreme precipitation under changed climate. Geophys. Res. Lett. 32.

Fisher, R.A., Tippett, L.H.C., 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In: Mathematical Proceedings of the Cambridge Philosophical Society. Cambridge University Press, pp. 180–190.

Ghil, M., Yiou, P., Hallegatte, S., Malamud, B., Naveau, P., Soloviev, A., Friederichs, P., Keilis-Borok, V., Kondrashov, D., Kossobokov, V., et al., 2011. Extreme events: dynamics, statistics and prediction. Nonlinear Process. Geophys. 18, 295–350.

Gilleland, E., Katz, R.W., 2011. New software to analyze how extremes change over time. Eos, Trans. Am. Geophys. Union 92, 13–14.

Haylock, M., Hofstra, N., Klein Tank, A., Klok, E., Jones, P., New, M., 2008. A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. J. Geophys. Res.: Atmos. (1984–2012) 113.

Held, I.M., Soden, B.J., 2006. Robust responses of the hydrological cycle to global warming. J. Clim. 19, 5686–5699.

Hempel, S., Frieler, K., Warszawski, L., Schewe, J., Piontek, F., 2013. A trend-preserving bias correction—the isi-mip approach. Earth Syst. Dyn. 4, 219–236.

Herring, S.C., Hoerling, M.P., Peterson, T.C., Stott, P.A., 2014. Explaining Extreme Events of 2013 from a Climate Perspective, vol. 95. American Meteorological Society.

Huntingford, C., Marsh, T., Scaife, A.A., Kendon, E.J., Hannaford, J., Kay, A.L., Lockwood, M., Prudhomme, C., Reynard, N.S., Parry, S., et al., 2014. Potential influences on the united kingdom's floods of winter 2013/14. Nat. Clim. Change 4, 769–777.

Jones, P.W., 1999. First-and second-order conservative remapping schemes for grids in spherical coordinates. Mon. Weather Rev. 127, 2204–2210.

Jones, R., 2004. Generating High Resolution Climate Change Scenarios Using PRECIS. MetOffice.

Katz, R.W., 2010. Statistics of extremes in climate change. Clim. Change 100, 71–76.

Katz, R.W., Craigmile, P.F., Guttorp, P., Haran, M., Sansó, B., Stein, M.L., 2013. Uncertainty analysis in climate change assessments. Nat. Clim. Change 3, 769–771.

Kharin, V.V., Zwiers, F., Zhang, X., Wehner, M., 2013. Changes in temperature and precipitation extremes in the cmip5 ensemble. Clim. Change 119, 345–357.

Kharin, V.V., Zwiers, F.W., 2000. Changes in the extremes in an ensemble of transient climate simulations with a coupled atmosphere–ocean gcm. J. Clim. 13, 3760–3788.

Kharin, V.V., Zwiers, F.W., Zhang, X., Hegerl, G.C., 2007. Changes in temperature and precipitation extremes in the ipcc ensemble of global coupled model simulations. J. Clim. 20, 1419–1444.

Lorenz, E.N., 1963. Deterministic nonperiodic flow. J. Atmos. Sci. 20, 130–141.

Martins, E.S., Stedinger, J.R., 2000. Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. Water Resour. Res. 36, 737–744.

Massey, N., Jones, R., Otto, F., Aina, T., Wilson, S., Murphy, J., Hassell, D., Yamazaki, Y., Allen, M., 2014. weather@ home-development and validation of a very large ensemble modelling system for probabilistic event attribution. Q. J. R. Meteorol. Soc http://dx.doi.org/10.1002/qj.2455.

O'Gorman, P.A., Schneider, T., 2009. The physical basis for increases in precipitation extremes in simulations of 21st-century climate change. Proc. Natl. Acad. Sci. 106, 14773–14777.

Otto, F., Massey, N., Oldenborgh, G., Jones, R., Allen, M., 2012. Reconciling two approaches to attribution of the 2010 Russian heat wave. Geophys. Res. Lett. 39.

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

Pall, P., Allen, M., Stone, D., 2007. Testing the Clausius–Clapeyron constraint on changes in extreme precipitation under CO_2 warming. Clim. Dyn. 28, 351–363.

Pope, V., Gallani, M., Rowntree, P., Stratton, R., 2000. The impact of new physical parametrizations in the Hadley centre climate model: Hadam3. Clim. Dyn. 16, 123–146.

R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL ⟨http://www.R-project.org⟩. version 2.10.0.

Schaller, N., Otto, F., van Oldenborgh, G., Massey, N., Sparrow, S., Allen, M., 2014. The heavy precipitation event of May–June 2013 in the upper Danube and Elbe basins. In: Herring, S., Hoerling, M., Peterson, T., Stott, P. (Eds.), Explaining Extreme Events of 2013 from a Climate Perspective, Bulletin of the American Meteorological Society. pp. S69–S72.

Screen, J.A., 2014. Arctic amplification decreases temperature variance in northern mid-to high-latitudes. Nat. Clim. Change 4, 577–582.

Seneviratne, S.I., Nicholls, N., Easterling, D., Goodess, C.M., Kanae, S., Kossin, J., Luo, Y., Marengo, J., McInnes, K., Rahimi, M., et al., 2012. Changes in climate extremes and their impacts on the natural physical environment. In: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation, pp. 109–230.

Sillmann, J., Croci-Maspoli, M., Kallache, M., Katz, R.W., 2011. Extreme cold winter temperatures in Europe under the influence of north Atlantic atmospheric blocking. J. Clim. 24, 5899–5913.

Sippel, S., Otto, F.E., 2014. Beyond climatological extremes-assessing how the odds of hydrometeorological extreme events in south-east Europe change in a warming climate. Clim. Change 125, 381–398.

Stark, J.D., Donlon, C.J., Martin, M.J., McCulloch, M.E., 2007. Ostia: An operational, high resolution, real time, global sea surface temperature analysis system. In: OCEANS 2007-Europe, IEEE. pp. 1–4.

Stone, D.A., Allen, M.R., 2005. The end-to-end attribution problem: from emissions to impacts. Clim. Change 71, 303–318.

Stott, P.A., Allen, M., Christidis, N., Dole, R.M., Hoerling, M., Huntingford, C., Pall, P., Perlwitz, J., Stone, D., 2013. Attribution of Weather and Climate-related Events. In: Climate Science for Serving Society. Springer, Dordrecht, Heidelberg, New York, London pp. 307–337.

Stott, P.A., Stone, D.A., Allen, M.R., 2004. Human contribution to the European heatwave of 2003. Nature 432, 610–614.

Taylor, K.E., Stouffer, R.J., Meehl, G.A., 2012. An overview of cmip5 and the experiment design. Bull. Am. Meteorol. Soc. 93, 485–498.

Westra, S., Alexander, L.V., Zwiers, F.W., 2013. Global increasing trends in annual maximum daily precipitation. J. Clim. 26, 3904–3918.

Wigley, T.M., 2009. The effect of changing climate on the frequency of absolute extreme events. Clim. Change 97, 67–76.

Zhang, X., Hegerl, G., Seneviratne, S., Stewart, R., Zwiers, F., Alexander, L., 2014. WCRP Grand Challenge: Understanding and Predicting Weather and Climate Extremes. Technical Report. World Climate Research Program. URL 〈 http://www.wcrp-climate.org/images/documents/grand_challenges/GC_Extremes_v2.pdf〉 White Paper.