Very Rare Heat Extremes: Quantifying and Understanding Using Ensemble Reinitialization

CLAUDIA GESSNER,^a ERICH M. FISCHER,^a URS BEYERLE,^a AND RETO KNUTTI^a

^a *Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland*

(Manuscript received 27 November 2020, in final form 7 May 2021)

ABSTRACT: Heat waves such as the one in Europe 2003 have severe consequences for the economy, society, and ecosystems. It is unclear whether temperatures could have exceeded these anomalies even without further climate change. Developing storylines and quantifying the highest possible temperature levels is challenging given the lack of a long homogeneous time series and methodological framework to assess them. Here, we address this challenge by analyzing summer temperatures in a nearly 5000-yr preindustrial climate model simulation, performed with the Community Earth System Model CESM1. To assess how anomalous temperatures could get, we compare storylines generated by three different methods: 1) a return-level estimate, deduced from a generalized extreme value distribution; 2) a regression model, based on dynamic and thermodynamic heat wave drivers; and 3) a novel ensemble boosting method, generating large samples of reinitialized extreme heat waves in the long climate simulation. All methods provide consistent temperature estimates, suggesting that historical exceptional heat waves such as those in Chicago in 1995, Europe in 2003, and Russia in 2010 could have been substantially exceeded even in the absence of further global warming. These estimated unseen heat waves are caused by the same drivers as moderate observed events, but with more anomalous patterns. Moreover, altered contributions of circulation and soil moisture to temperature anomalies include amplified feedbacks in the surface energy budget. The methodological framework of combining different storyline approaches of heat waves with magnitudes beyond the observational record may ultimately contribute to adaptation and to the stress testing of ecosystems or socioeconomic systems to increase resilience to extreme climate stressors.

KEYWORDS: Atmosphere-land interaction; Atmospheric circulation; Extreme events; Summer/warm season; Surface temperature; Regression analysis; Statistical techniques; Ensembles

1. Introduction

Over recent decades, several outstanding heat waves have been recorded across densely populated regions of the world, including Chicago in 1995 (Changnon et al. 1996), western Europe in 2003 (García-Herrera et al. 2010), Russia in 2010 (Dole et al. 2011), and Australia in 2013 (Lewis and Karoly 2013). These climate extremes were characterized by exceptionally high temperatures that seriously impaired human and natural systems, such as by excess air pollution (Shaposhnikov et al. 2014; De Sario et al. 2013), mortality (Luber and McGeehin 2008; Robine 2008; Merte 2017), crop loss (Bastos et al. 2014; Ciais et al. 2005; De Bono et al. 2004; Barriopedro et al. 2011; van der Velde et al. 2012), or an increased chance of wildfires (De Bono et al. 2004; Gill and Malamud 2014; Parente et al. 2018). With increasing atmospheric greenhouse gas concentrations and continuing warming, heat wave frequency, duration, and intensity are projected to strongly increase (Della-Marta et al. 2007; Clark et al. 2006; Perkins et al. 2012;

Perkins-Kirkpatrick and Gibson 2017; Pfleiderer et al. 2019) and potential economic losses will exacerbate across the globe until the end of the twenty-first century (Easterling et al. 2000; Stein and Stein 2014; Orlov et al. 2019; Diffenbaugh and Burke 2019).

In light of such hazards, stakeholders and decision-makers are concerned about how anomalous temperatures could get in today's climate or the near future. Could more extreme anomalies unseen in the observational record have occurred without additional warming? Recent events can provide insight, but the highest possible anomalies of temperatures are unclear, mainly due to the following reasons: 1) Observational periods of climate and weather data are short and hence the sampling is incomplete particularly for very extreme values. During the relatively short observational record some regions may not have experienced the most pronounced heat wave magnitudes possible. 2) The climate system is currently in a transient state; drivers, feedbacks, and mean and extreme temperature levels continuously change over time (Stott et al. 2004; Shepherd 2014; Fischer and Knutti 2015; Christidis and Stott 2014). 3) The underlying physical mechanisms of extreme heat waves are known to a large extent qualitatively, but their magnitude and representation exhibit large uncertainties in global climate model simulations (Ehret et al. 2012; Diffenbaugh et al. 2017), affecting the assessment of

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JCLI-D-20-0916.s1>.

Corresponding author: Claudia Gessner, claudia.gessner@env.ethz.ch

DOI: 10.1175/JCLI-D-20-0916.1

© 2021 American Meteorological Society

This article is licensed under a Creative Commons Attribution 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

heat extremes (Sippel and Otto 2014; Angéilil et al. 2016) and their impacts (Hempel et al. 2013). 4) Overall, there is a lack of methodological approaches to overcome the listed obstacles and to estimate highest possible temperatures, based upon our physical understanding. Filling the gap of statistical, process-based, and model-based approaches to evaluate very rare heat extremes motivates this study. More precisely, we develop and compare different methodological frameworks to address the question of how much more intense recent heat waves could have gotten, and how their dynamic and thermodynamic drivers differ from more moderate events.

Daily observational data typically cover only a few decades so that moderate heat waves, occurring every few years, are well captured. Their driving mechanisms have been profoundly analyzed in climate science. High summer temperatures in the midlatitudes are induced by persistent anticyclonic circulation anomalies that yield collocated warming through descending air masses, high insolation, and advection of warm air from the surrounding area, mostly from lower latitudes (Pfahl and Wernli 2012; Horton et al. 2016; Schumacher et al. 2019; Petoukhov et al. 2013). These anticyclones are triggered by Rossby wave trains or subtropical ridges (Woollings et al. 2008; Davini et al. 2012; Coumou et al. 2015). Land–atmosphere processes can amplify heat waves, such as by the soil moisture–atmosphere feedback, enhancing sensible heat fluxes, and reducing latent cooling over dry soils (Seneviratne et al. 2010; Fischer et al. 2007; Miralles et al. 2014).

In contrast, little work has been done on very intense and, in a statistical sense, very rare temperature extremes that occur once in several centuries to millennia, although the probability of historically unprecedented hot extremes increase with anthropogenic global warming (Lewis and King 2017; Diffenbaugh et al. 2018; Fischer and Schär 2010). In the literature, such a low probability–high impact event is also referred to as a perfect storm or gray swan, defining an unseen but predictable event (Paté-Cornell 2012; Stein and Stein 2014). In the following, the term “very rare heat waves” is used to describe low probability–high impact events of very high near-surface air temperature anomalies.

Commonly, the rareness of extremes is quantified by their return period in years, which expresses the average waiting time between two events of the same kind. The inverse is the corresponding likelihood of occurrence within a year. However, due to comparatively short records for most climate data, it is challenging to quantify return periods of very rare events as they reach beyond the period length (Coles et al. 2001).

The generalized extreme value (GEV) distribution is a popular and well-established statistical approach to extrapolate into the far tails of meteorological datasets and estimate magnitudes of climate events more extreme than those found in observations (Coles et al. 2001). Most GEV distributions of local or regional mean temperature maxima are bounded, suggesting a theoretical upper temperature limit (Wang et al. 2016), not necessarily consistent with physical constraints. Referring to physical laws, an upper temperature bound is conceivable, since specific environmental conditions limit the energy input into the climate system, whereas outgoing longwave radiation scales with surface

air temperature T^4 , providing a very strong negative feedback on net surface energy. In a thought experiment, maximum surface air temperatures could be attained under, for instance, daytime clear-sky conditions and heat accumulation over several days in the boundary layer (Black et al. 2004; Miralles et al. 2014), providing the highest transmission of solar radiation through the atmosphere and/or completely dry soils, entirely converting absorbed energy into sensible heat. Based on this idea, a key point of this study is to assess and compare different approaches to develop storylines of very rare heat waves near the upper temperature bounds. Storylines here refer to self-consistent and plausible unfoldings of a heat wave event [following the definition introduced in Shepherd et al. (2018)] and developed using statistical, process-based, and climate model-based approaches. Note that storylines provide a qualitative understanding of certain climate events rather than quantitative precision; that is, they cannot be interpreted in a probabilistic way, but they can serve as a plausible worst case to stress test a system (Shepherd 2019; Shepherd et al. 2018).

Previous studies apply different methods to increase the sample size of independent climate extremes in long climate data, using extreme value theory or rare event sampling approaches based on numerical models or empirical importance sampling. The idea of rare event sampling is to repeatedly generate more realizations of climate events in a model (Lin and Emanuel 2016; Ragone et al. 2018; Webber et al. 2019; Plotkin et al. 2019). Depending on the context, this method can be computationally efficient, but it may suffer from model biases and the plausibility of simulation scenarios. Recently empirical importance sampling in combination has been proposed as a powerful way of efficiently generating very rare heat waves (Yiou and Jézéquel 2020). Nevertheless, resampling of very rare events is still in its infancy and subject to providing unphysical climate events.

Other studies analyze very long climate simulations, which have a higher chance to catch very rare events than shorter simulations, due to the chaotic nature of the climate systems (Huang et al. 2016; Sriver et al. 2015). Furthermore, long climate periods allow for a robust representation of natural and forced variability. Long climate simulations are computationally expensive, but they provide a physically consistent time series that samples largely independent climate extremes. This reduces the risk of biases in the return period during “disaster gaps” (i.e., if the distribution is fit to a too short nonrepresentative period).

Here we use different methods to develop storylines for very rare temperature anomalies unseen in observations by combining both statistical methods and large ensembles with physical understanding. The exceptionally long climate simulation offers a test bed for the robustness of methodological frameworks to explore the far tails of climate distributions. To avoid challenges in a transient climate state, our analysis is based on a millennial GCM simulation of equilibrated preindustrial climate that surpasses the period length of most available climate data. For this purpose, we put observed events in the context of a preindustrial quasi-equilibrium climate state. The newly introduced ensemble boosting method further increases the sample of temperature

anomalies by generating storylines of the most extreme simulated heat waves.

This paper is organized into two major sections. The first part evaluates summer heat waves in the long preindustrial simulation and identifies the main physical drivers of the most extreme temperature anomalies over three historical hotspots in central North America (CNA), western Europe (WEU), and western Russia (WRU). The second part compares different methods to develop storylines of very rare heat waves with the focus on WEU. Introducing the novel ensemble boosting method additionally provides deeper insight into the contribution of heat wave drivers.

2. Methods

This study analyses very rare heat extremes in an exceptionally long preindustrial control simulation that is carried out with the fully coupled Community Earth System Model version 1.2 (CESM-PI), including the Community Atmosphere Model version 5 (CAM5) and the Community Land Model (CLM4) at about 2° horizontal resolution and the ocean and sea ice model at about 1° horizontal resolution (Hurrell et al. 2013). The preindustrial climate forcing remains constant throughout the full simulation period and the simulation is ensured to be equilibrated by excluding the first 500 years of simulation. The remaining dataset encompasses 4780 years of daily mean and maximum 2-m air temperature (Tx), 500-hPa geopotential height (Z500), precipitation as well as monthly soil moisture (SOIL), shortwave downwelling radiation (SSR), and evaporative fraction (EF; i.e., the ratio of the latent heat flux to the sum of latent and sensible heat fluxes at the surface).

A transient climate scenario is performed with the same version of the fully coupled CESM1.2, based on an 84-member initial condition ensemble (CESM-IC) from 1950 to 2018. Twenty-one members start from different ocean initial conditions from the long preindustrial control simulations. From each of the 21 ocean initial condition members three members are branched off on 1 January 1940 by randomly perturbing their atmospheric initial condition fields with a magnitude of 10^{-14} K. This results in a total of 84 ensemble members analyzed here that sample unforced internal variability from 1950 to 2100 and are run freely until the end of the simulation, prescribing the same historical forcing until 2005 and the RCP8.5 scenario from 2006.

The model performance is evaluated using the ECMWF reanalysis ERA5 over the period 1979–2018 (Hersbach et al. 2020). We compare the anomaly distributions and the summer climatology of daily mean temperature, Z500, precipitation, monthly mean surface solar radiation downward, and volumetric soil water (<https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview>) with the preindustrial control run to assess the climate variability. We particularly study the tail of temperature distribution by comparing only the hottest temperature extremes in the analyzed ERA5 period and the control run. From hourly mean 2-m temperature we derive the daily maximum 2-m temperature.

All presented simulation results of CESM1.2 or ERA5 are processed by the same steps: 1) ERA5 data are interpolated on the coarser longitude–latitude grid of CESM1.2. 2) To consider

primarily large-scale anomalies, all variables are averaged over the land area of our regions of interest, respectively: central North America (CNA; 36.9°–42.6°N, 82.5°–92.5°W), western Europe (WEU; 42.6°–50.2°N, 0°–7.5°E), and western Russia (WRU; 52.1°–57.8°N, 35°–45°E). 3) Afterward, 7-day running means of spatially averaged daily data are computed. 4) Then, only Northern Hemisphere summer months are selected, and the constant time mean over all summertime days (JJA) is subtracted and thereby, mean biases are removed, respectively. 5) Finally, all transient climate simulations, ERA5, and the CESM-IC runs are linearly detrended to avoid warming asymmetries in climatology. As a result, our analysis of temperature extremes is based on 7-day running mean daily maximum 2-m temperature (Tx7d) anomalies with respect to the constant mean value over all summer months.

We use the generalized extreme value (GEV) distribution as one approach to gain insight into the tails of temperature distributions, despite our limited sample of temperature extremes (Coles et al. 2001). Commonly, the fitted GEV distribution of temperature maxima has a negative GEV shape parameter, which characterizes the reversed Weibull distribution with an upper bound. We utilize the R package “extRemes” (Gilleland and Katz 2016) to fit the GEV distribution to regional means over our selected areas, using a block size of 20 years, which mostly avoids interannual dependencies of the climate system and ensures the analysis of rare extremes that are not reoccurring on an annual basis. Such large block sizes have scarcely been used in the analysis of meteorological variables, due to the lack of long climate data but are here feasible given the long simulation and the aim to sample the most extreme events only. Area means are selected to analyze regional-scale heat waves that are insensitive to local land surface conditions. The GEV parameters are calculated by the maximum likelihood estimation and the 95% confidence interval of the return levels is obtained by the profile likelihood method.

Since even several thousand years of simulation do not necessarily sample the very tails of climate distributions sufficiently, we introduce ensemble boosting as an alternative method to efficiently generate large samples of climate events in the CESM-PI simulation. Some of the most extreme regional Tx7d anomalies are rerun by writing out daily restart files in the weeks before the event. Rerunning the events is made possible by ensuring binary reproducibility of the numerical experiments on the high-performance computers. Using these daily restart files large reinitialization ensembles, hereafter referred to as boosted ensembles, are produced by adding roundoff perturbations to the atmospheric initial conditions for individual lead times (i.e., the number of time steps before the event). Selected heat waves are rerun with lead times of a few days to weeks before the day of maximum temperature anomaly. Fifty-member ensembles are randomly perturbed in their initial temperature fields with a magnitude of 10^{-14} K at each grid point. However, the simulation results cannot be interpreted in a probabilistic sense together with the samples of the multimillennial simulations, since the boosted ensembles are designed to generate many climate events from a specific event in the preindustrial control run. Thus,

FIG. 1. (a)–(c) Return period plots (blue lines) with 95% confidence intervals (dashed lines) for regional average T_{x7d} anomalies in the CESM-PI run. Black rectangles enclose the three selected areas in the map in (d). The 7-day running mean maximum 2-m temperature anomalies over central North America in 1995, western Europe in 2003, and western Russia in 2010 in ERA5 are displayed as horizontal lines. Yellow lines show the annual residual anomalies relative to a linearly detrended climatology. Red lines are the shifted regional mean temperature difference between CESM-PI and CESM-IC to estimate anomalies relative to preindustrial climate. (d) Difference of standard deviation of monthly 2-m temperatures (JJA) between historical CESM-IC runs and ERA5 (1979–2018).

very rare heat extremes in the boosted ensembles are not part of an independent sample.

3. Model evaluation

In the first part, we evaluate the representation of the Northern Hemisphere summer climate in the CESM1.2 simulations to assess the reliability of our analysis. To that end, the variability of summer mean temperature in CESM-PI and CESM-IC runs is compared to the ERA5 reanalysis. The midlatitudes, including

our regions of interest, show reasonably small differences of 2-m temperature standard deviation (SD) between the detrended present-day simulations CESM-IC and ERA5 (Fig. 1d). While the present-day (CESM-IC) is obviously warmer than the preindustrial period (Fig. S1 in the online supplemental material), there is similar interannual variability in CESM-PI and CESM-IC (Fig. S2) as well as ERA5 (Fig. 1d). Larger discrepancies occur in areas of well-known biases (e.g., in mountainous regions or underestimated cold upwelling currents off the west coasts). Additionally, the difference between

model and reanalysis may be partly affected by uncertainties in ERA5 such as in data-sparse regions, where ERA5 can assimilate few in situ observations. Moreover, the monthly variability of well-known heat wave drivers such as Z500 and precipitation is well captured by the model in the three regions of interest central North America (CNA), western Europe (WEU), and western Russia (WRU), although variabilities are generally somewhat lower in CESM-PI (Fig. S3). Figure 1d shows that this is specific for our selected regions. Fitting a GEV distribution to 2-yr block maxima of summertime Tx7d anomalies in CESM-PI and ERA5 reveals that our model captures the overall characteristics of the return levels reasonably well, but rather conservatively, being at the low end of the ERA5 confidence intervals (Fig. S4). Here, we compare short return periods as the reanalysis only covers a short period. The fact that the model performs quite well against ERA5 with regard to mean and variability in these regions and that the areas were recently affected by historically exceptional heat extremes motivates our choice of these areas of interest.

Overall, the long preindustrial and historical large ensemble simulations capture the summer climatology reasonably well for the three regions of interest. The following sections focus on methods to estimate the far tails of temperature distribution. Although these heat extremes cannot be rigorously evaluated, due to the short time series of ERA5, the good agreement increases our confidence in the realistic model representation of heat waves, including very rare heat waves.

4. Results

a. GEV-based estimates for unseen very rare heat extremes

The heat waves of Chicago in 1995, Europe in 2003, and Russia in 2010 are well known for their exceptional temperature extremes. This raises the question how much worse temperature anomalies could have been relative to the mean climate. One way to address this question is to fit a generalized extreme value (GEV) distribution to very rare Tx7d anomalies of the long CESM-PI run, using 20-yr block maxima (see section 2 for details). The temperature levels are presented in Figs. 1a–c as functions of return periods. The shapes of the fitted GEV distributions are insensitive to slight changes of the selected areas and consistent with the results of ERA5 for rather short return periods (Fig. S4). These results are also consistent with recent work (Huang et al. 2016; Ben Alaya et al. 2018; Hogan et al. 2019). We compare the GEV distributions of maximum daily Tx7d anomalies in the CESM-PI simulation against the linearly detrended values in ERA5 during the heat waves in Chicago in 1995, Europe in 2003, and Russia in 2010, averaged over the same regions, respectively (yellow lines).

To quantify how anomalous recent observed heat waves would have been relative to preindustrial climate, we have to account for the fact that the world has substantially warmed since the preindustrial period. Therefore, absolute temperature anomalies relative to the present-day climate in ERA5 are here shifted by the regional mean difference between the CESM-PI and CESM-IC simulations. Other variables remain unchanged. This generates a conservative first-order approximation of warming changes in the observed heat wave extremes (red horizontal lines)

to compare these temperature anomalies with the preindustrial climate. Note that the Chicago heat wave only lasted for a few days, whereas the European and Russian heat waves persisted for weeks (Black et al. 2004; Barriopedro et al. 2011). Therefore, the temperature peak, shifted by the regional warming over CNA, is strongly smoothed on a weekly time scale, only having a return period of about 28–33 years with respect to the preindustrial simulation, whereas the longer-lasting heat wave in 2003 is expected once every 800–3000 years and the one in 2010 once in 5000 or more years, in the preindustrial climate. In contrast to the findings by Vogel et al. (2019) and Imada et al. (2019), who concluded that recently observed large-scale seasonal heat extremes would have been nearly impossible under preindustrial climate conditions, we find that very rare extrapolated temperatures from the preindustrial long control run can exceed the 7-day maximum anomalies in the reanalysis.

In the simulations the observed heat anomalies have been exceeded over CNA and WEU even in the preindustrial period. However, even an almost 5000-yr GCM simulation does not perfectly sample the far tails of temperature. A comparison for rather small return levels between ERA5 and CESM-PI shows that our model captures the overall characteristics of 2-yr Tx7d block maxima reasonably well but tends to be at the low end of the estimated return levels in ERA5 (Fig. S4). Assuming also high return levels to be captured correctly by the control run, the fitted GEV distributions and first-order approximations suggest that the observed heat extremes could have been exceeded by up to several degrees for all three hot spot regions even in the cooler preindustrial climate. Note that the GEV fits are extrapolated here up to 100 million years for experimental purpose, where Tx7d anomalies asymptotically approach a potential upper limit of which even the lower confidence bound is higher than the observed events in WEU and CNA. These return levels far beyond the period length are obviously associated with very large uncertainty ranges but are used here as one approach to estimate potential levels of unseen very rare events, which are compared to alternative estimates below.

Given that model simulations suggest that 7-day running mean temperature anomalies much more intense than observed are plausible, we address the question how close the highest anomaly of ERA5 got relative to an estimated 1-in-2000-yr preindustrial anomaly. Figure 2 shows a map of the ratio between the highest Tx7d detrended anomalies in ERA5, shifted by the temperature difference between CESM-PI and CESM-IC from 1979 to 2018 and the 2000-yr return level in the CESM-PI run. Note that a temperature shift is only considered in Figs. 1a–c and 2, where we compare temperature anomalies relative to a common preindustrial baseline. Over land areas, ratios are mostly larger than one, suggesting that Tx7d anomalies much higher than the ones observed are possible. On the other hand, ratios smaller than one suggest that the highest observed anomaly even exceed a 1-in-2000-yr event in the preindustrial simulation. These can mainly be found over the oceans, potentially due to the warming since the preindustrial period, yielding a signal-to-noise ratio that is substantially higher for marine heat waves than heat waves on land (Froelicher et al. 2018). King et al. (2020) found corresponding patterns of temperature differences between transient and quasi-equilibrium

FIG. 2. Ratio of 2000-yr return level of 7-day running mean maximum 2-m temperature (Tx7d) anomalies in CESM-PI and the highest observed Tx7d anomaly in ERA5 (1979–2018) during JJA. Red colors mark areas where temperature substantially exceed those observed in recent heat waves. Black rectangles show the three selected regions of interest.

climates over the Atlantic Ocean, indicating that lower-latitude oceans warm faster in a transient climate and hence have a higher chance to exceed the 2000-yr temperature anomaly in the control run. Additionally, different horizontal resolutions may contribute to the land–ocean contrast and individual high values along coasts. The latter may relate to grid points that are partly land-covered in the high-resolution ERA5 but ocean-covered in CESM1.2, leading to lower variability and less intense extremes.

Fitting a GEV distribution to 7-day running mean temperature anomalies in the long preindustrial simulation suggests that temperatures asymptotically approach a theoretical upper limit. The exact temperature level is obviously highly uncertain. However, this analysis primarily focuses on the methodology on how to quantify such very rare extremes rather than providing precise numbers for real-world extremes. Note that the presented results are also highly model dependent and upper bounds for temperature are so far purely statistically derived. Thus, in the following, potential upper bounds for temperature are compared to alternative approaches of estimating very rare events.

b. Drivers of very rare heat extremes

The fact that CESM1.2 can simulate heat extremes more intense than the ones observed raises the question whether the drivers of those unseen events are the same as for more moderate ones or whether other factor come into play. Therefore, this section evaluates the physical drivers and large-scale context of very rare Tx7d anomalies in CESM-PI (>99th percentile) or ERA5 (>90th percentile) in comparison to the remaining more moderate events.

For all three selected regions, we find mean circulation patterns associated with very rare Tx7d anomalies that are consistent between the CESM-PI simulation and ERA5 (Fig. 3; see also Figs. S5–S9). The strongest Z500 anomalies found in the long preindustrial scenario are stronger than during the observed events and consistent with the respective temperature anomalies. This was to be expected as the CESM-PI simulations cover an about 100 times larger sample of heat waves than the reanalysis. The hot spots are related to collocated high Z500 anomalies above the 99th percentile of the summer

climatology in CESM-PI (90th percentile in ERA5), revealing persistent anticyclones as a primary physical driver. Particularly over WEU and WRU, these anticyclones are prolonged by the cyclonic circulation anomalies up- and/or downstream (Colucci 1985; Luo et al. 2014), whereas weaker zonal gradients of Z500 occur over CNA.

Considering various physical drivers, Fig. 4 and Figs. S10 and S11 show that very rare heat events in the preindustrial scenario are mostly related to preceding anomalies in well-known drivers, already starting in the winter–spring transition. Long-term moisture deficits are of particular importance, due to the long memory of soil moisture. Indeed, persistent precipitation deficits and high insolation yield continuing soil drying, accompanied by reduced evaporative fractions for all three selected areas. More than one-third of all very rare heat events over WEU and WRU are affected by strong monthly mean precipitation and soil moisture deficits (exceeding one standard deviation) over the preceding three months. However, in CNA only a few events are affected by strong moisture deficits. We suspect that these wetter conditions might be a reason for less pronounced temperature anomalies (Fig. 1a).

To understand the role of the different drivers, we use them as predictors in a linear regression model, estimating the explained variance in year-to-year variability of Tx7d anomalies. We take three predictors into account that were previously found to be the main drivers of the very rare heat waves in the preindustrial scenario: 1) 7-day running mean Z500 anomaly coinciding with maximum Tx7d anomaly, representing the circulation component of drivers, 2) 2-month mean surface solar radiation (SSR) as the radiation component, and 3) 2-month mean evaporative fraction (EF) that provides the contribution of land–atmosphere interactions. The components predictors 2 and 3 above are averaged over the month including the hottest day of the heat wave and the month before. The predictors are applied in a multiple linear regression on very rare or moderate Tx7d anomalies (Fig. 5; see also Fig. S12). Since the predictors are not independent, hierarchical partitioning is used to assign the individual contributions (Chevan and Sutherland 1991). Independent contribution describes the variability in a predictor independently from the

FIG. 3. Mean over very rare 7-day running mean maximum 2-m temperature (Tx7d) anomalies (color shading) and concurrent 7-day running mean Z500 anomalies (contour lines; 0.5 standard deviation spacing) over WEU, exceeding the (a) 99th percentile in the CESM-PI run or (b) 90th percentile in ERA5 from 1979 to 2018. Hatched areas indicate significance of Z500 at the 5% level (Student's t test).

other predictors. Joint contribution is the variability that can be explained by several predictors with positive interaction, showing additive effects, and negative interaction, showing suppressive effects. We find that the three drivers explain a large proportion of variance of moderate heat waves for all three selected regions. Z500 is by far the biggest contributor, but SSR and EF also explain a substantial fraction of the variance. In contrast, these drivers explain more variance among the very rare Tx7d anomalies, mainly due to a substantially smaller contribution of Z500. Therefore, we argue that the relative contribution of the heat wave drivers analyzed here changes for the hottest events in the long preindustrial simulation. The impact of circulation on maximum temperatures is reduced given that Z500 has already reached high percentiles and other driver components get more effective, such as when drier soils increase the fraction of sensible heating at the land surface. The latter example is confirmed by an increasing contribution of EF to Tx7d from moderate to rare events in WEU and WRU, emphasizing the growing impact of the soil moisture–temperature feedback on temperature anomalies for the hottest and driest heat waves. However, local Z500 does not perfectly reflect the impact of circulation anomalies, as the same Z500 anomalies can have different backward trajectories of air parcels, determining their properties. In contrast, only 39% variance of moderate annual maximum Tx7d anomalies can be explained by the selected drivers over CNA (Fig. S12), compared to WEU and WRU, indicating weaker or more short-lived heating periods.

Altogether, we have found that the well-known dynamic, thermodynamic, and radiative drivers of heat waves show reasonable

anomalies before and during the summer of very rare heat extremes, consistent with previous studies (Cassou et al. 2005; Ding and Wang 2005; Barriopedro et al. 2011; Jäger and Seneviratne 2011; Teng et al. 2019; Wehrli et al. 2019; Kornhuber et al. 2020). The analyzed physical drivers are consistent but more pronounced for very rare temperature extremes, compared to more moderate events. The contribution of drivers changes for the hottest and driest events. The fact that anomalies of some of the identified drivers are bounded by physical constraints (e.g., by completely desiccated soils or insolation time) is consistent with the idea of physical constraints bounding temperature maxima, as suggested by the GEV distribution. In the following section, we compare different approaches to quantify storylines of very rare events and their physical drivers.

c. Regression-based storylines of very rare heat waves

To further explore very rare extremes, we generate a regression model to estimate temperature anomalies under physically constrained limits of various drivers, representing heat wave scenarios. First, we repeat the multiple linear regression (MLR) of annual maximum Tx7d anomalies in CESM-PI and ERA5, using the following physical drivers as predictors, which show significant covariance with temperature in the previous sections: 7-day running mean Z500 concurrent with maximum Tx7d anomaly, 2-month mean SSR, and soil liquid water to a depth of 0.2 m (SOIL), both averaged over the month of the temperature peak and the month before. With each of the predictors, we consider another group of physical drivers, including dynamic drivers by Z500 anomalies,

FIG. 4. Monthly anomalies of (a) daily maximum temperature, (b) Z500, (c) precipitation, (d) soil liquid water to a depth of 0.2 m, (e) solar radiation, and (f) evaporative fraction at the surface, averaged over western Europe. Values are multiples of the standard deviation (SD), respectively. Each column represents a month. Only years with very rare Tx7d anomalies in JJA are considered and illustrated as anomalies relative to the corresponding monthly means in all years. Solid lines are the monthly medians, respectively.

thermodynamic drivers by SOIL, and radiation by SSR. We use SOIL instead of EF as both are highly correlated in the control run (0.78) and the physical limits of soil moisture are more straightforward to define for the purpose of heat wave scenarios. We further test to what extent a quadratic relation to SOIL can the systematic underestimation in the upper tail of temperature, which is of interest here. Adding higher complexity also increases the risk of overfitting. However, note that the sample size here is very large and that the overall

conclusion is insensitive to whether we use a linear or quadratic dependence on SOIL (not shown). The resulting regression model provides a reasonable prediction of annual maximum temperature anomalies, explaining a total variance of 73% over WEU (Fig. 6a). A little less variance of 70% is explained in WRU and only 39% of Tx7d variance is explained in CNA (Figs. S12 and S13). Since the explained variance is lowest for CNA, the estimated Tx7d maxima might be underestimated as other unconsidered drivers explain parts of the temperature

FIG. 5. Explained variance of moderate or very rare 7-day running mean maximum 2-m temperature (Tx7d) anomalies over western Europe, based on a multiple linear regression, using (a) 7-day running mean Z500 at the day of maximum temperature anomaly, (b) 2-month mean shortwave solar flux at the surface, and (c) 2-month mean evaporative fraction, averaged over the month of the heat wave and the previous month. (d) The total effect of all three variables. Independent (joint) effects are displayed in dark (light) gray.

FIG. 6. (a) Climate vs regression model based 7-day running mean maximum 2-m temperature (Tx7d) anomalies over western Europe, based on a regression model of annual maxima of Tx7d anomalies as a function of concurrent 7-day running mean Z500, 2-month soil liquid water, and solar radiation at the surface. Gray points show the annual anomalies in the CESM-PI run and orange points in ERA5 during 1979–2018. Solid gray lines mark the theoretical upper Tx7d bound and dashed lines the associated 95% confidence band, estimated by the GEV distribution (Fig. 1b). The thick black line denotes the bisector line, giving the perfect estimation. (b) Temperature anomalies estimated by the regression model and GEV distributions. The blue boxplot shows the estimate of the 4780-yr Tx7d anomaly, based on the GEV distribution. The remaining boxplots show the results of the regression function, assuming (from left to right) average values of all three considered drivers (empty boxplot), average values of Z500 and SSR but the minimum GEV bound of soil moisture (green boxplot), average of soil moisture but the maximum GEV bound of Z500 and SSR (yellow boxplot), and the GEV bounds of all three drivers (red boxplot).

variability. Despite a quadratic dependence on SOIL, very rare temperatures are still underestimated. We attributed this bias to a change of the impact of drivers from moderate to very rare heat waves as presented in Fig. 5, since the sample of our model fit is dominated by more moderate temperature anomalies.

In a next step, we use the regression model to develop storylines of very rare temperature anomalies (Fig. 6b). We estimate how extreme Tx7d anomalies could get if the above identified drivers reach extremely high anomalies, assuming that the regression model still holds for very rare events as discussed above. The extremely high anomalies in the physical drivers are derived by fitting a GEV distribution to the 20-yr block maxima or minima of the physical drivers Z500, SSR, and SOIL, respectively. All fitted GEV have a negative shape parameter resulting in a bounded nature and an asymptotic limit. The derived limits of the physical drivers are then used as predictors in the regression function to estimate associated very rare Tx7d anomalies, again using bootstrapping to compute confidence intervals.

As a reference, the blue bar in Fig. 6b (first from the left) marks the 95% confidence interval of a 4780-yr Tx7d anomaly (i.e., a 7-day running mean temperature anomaly that is

statistically expected to occur once in our preindustrial control simulation period). The yellow bar (second from the right) assumes that the 7-day running mean Z500 and 2-month mean SSR reach their upper GEV bounds, accompanied by 2-month climatological mean SOIL with respect to the summer climate. Under such very extreme Z500 and SSR anomalies but average soil moisture conditions the regression model predicts a Tx7d anomaly of about 11.4°C (4.3 SD) over WEU. The green bar (third from the right) describes the opposite case of minimum SOIL, co-occurring with average Z500 and SSR. This case when SOIL is extremely dry but Z500 and SSR are average represents the least extreme cases with an anomaly of about 6.1°C (2.3 SD), which is consistent with the argument that very dry soils are not a sufficient condition for an extreme heat waves (Quesada et al. 2012). Finally, the red bar (first from the right) represents the most extreme case with about a 15.2°C (5.7 SD) Tx7d anomaly, when all three drivers reach extreme levels and are assigned to their GEV bounds.

The temperature estimates are taken from the previously fitted GEV distribution (Fig. 1b), using bootstrapping to calculate the uncertainty in the GEV bounds and the standard deviation of the model parameter to calculate the prediction

FIG. 7. Example of ensemble boosting for the most extreme heat wave over western Europe in the CESM-PI simulation. (a) Development of 7-day running mean maximum 2-m temperature anomalies in the ensemble, started with a lead time of 16 days before the temperature peak in the control run (thick black line). The light red shading indicates the 5%–95% range of the ensemble. (b) Columns display the range of maximum Tx7d anomalies in each ensemble with its lead time denoted below. Ensembles are sorted in decreasing lead time order from left to right. The red line is the maximum Tx7d anomaly of the underlying heat wave in the CESM-PI run. Gray lines indicate the theoretical temperature bound and associated 95% confidence band, estimated by the GEV distribution. The light red column represents the corresponding temperature range in (a).

intervals. These estimates should be interpreted with caution and may lead to large uncertainties. Note that multicollinearity between the drivers is not considered in this analysis. Furthermore, we extrapolate beyond the temperature range of the long pre-industrial simulation and merge statistically estimated driver anomalies that are uncertain. Given all these limitations, it is remarkable that the storylines based on our regression model are broadly coherent with the bounds of the GEV directly fitted to the Tx7d anomalies (Fig. 6a, gray solid lines) and fall in the prediction interval of this estimate (Fig. 6a, gray dashed lines) over the selected regions.

The fact that the confidence intervals estimated by the GEV fit and the regression model are overlapping increases our confidence in the estimated temperature extremes. Accounting for the underestimation in the upper temperature tail, estimated anomalies might be even higher. However, this study focuses on the methodological frameworks to develop storylines of very rare heat waves rather than giving exact quantities, as derived in this section. In the following, we compare the storylines with a novel ensemble boosting method in a physically consistent numerical modeling environment.

d. Storylines of very rare heat extremes based on model ensemble boosting

In previous sections, we statistically estimate storylines of very rare heat waves, which are never even nearly reached in the multimillennial preindustrial simulation over the three

selected regions. However, running climate simulations at time scales of more than 5000 years is usually infeasible in terms of computational resources. Therefore, we introduce the novel climate model-based ensemble boosting method to efficiently generate large samples of temperature extremes without violating the physics as implemented in the model. These temperature samples are produced by first rerunning selected heat waves with higher output frequency (i.e., writing out daily restart conditions). Rerunning these requires binary reproducibility of numerical simulations on the high-performance infrastructure, which we successfully achieved after extensive testing on our dedicated high-performance computer nodes. From these rerun extreme heat waves with more daily restart conditions, initial condition ensembles referred to as boosted ensembles are produced by adding round-off perturbations to the atmospheric initial conditions. In this analysis, the samples of boosted heat waves are used to test whether simulated temperature extremes correspond to the statistically estimated potential upper bounds and whether they are strictly uncrossable in the model.

Figure 7a illustrates a representative example of ensemble boosting, applied to the most extreme heat wave over WEU, regarding Tx7d anomalies. A 50-member ensemble is started 16 days before the temperature peak. During the first week, maximum Tx7d anomalies hardly differ in the ensemble (light red shading), as perturbations are still small. Thereafter, perturbations have grown so that atmospheric dynamics differ

FIG. 8. Synthesis of maximum 7-day running mean daily maximum 2-m temperature (Tx7d) anomalies in the summer, estimated by the three presented approaches and based on the CESM-PI simulation: the upper bound of the GEV distributions (GEVb; gray boxplots; cf. Figs. 1a–c), the regression model output for boundary values of the main physical drivers (RegrM; red boxplots; cf. Fig. 6b), and the 50 hottest boosted ensemble members of the three most extreme heat waves (EnsB; pink boxplots; cf. Fig. 7 and Figs. S14 and S15). The results are averaged over the three selected areas. The dark red lines show the highest Tx7d anomalies in the CESM-PI run. The yellow lines mark the maximum anomalies relative to a linearly detrended climatology in ERA5 over central North America in 1995, western Europe in 2003, and western Russia in 2010, respectively.

between the members and the temperature range of the ensemble increases. Most members have less intense Tx7d anomalies particularly before the peak of the event but in a series of members the event is longer lasting.

However, the temperature range of boosted ensembles depends on the respective lead time and the governing dynamics of the individual heat waves. Figure 7b visualizes the range of maximum Tx7d anomalies for 12 different lead times from 7 to 19 days before the peak anomaly in the most extreme heat wave over WEU (i.e., 12 columns display the range of maximum Tx7d anomaly in each boosted ensemble). Some members substantially exceed the reference Tx7d anomaly (red line) by up to 3.7°C (1.4 SD) but still fall well within the confidence interval of the GEV fit (gray dashed lines; see more examples in Fig. S14). For all three analyzed regions, the Tx7d anomaly ranges resulting from the boosted heat waves mostly overlap with the confidence intervals of the fitted GEV distributions and prediction intervals of the multilinear regression models (Fig. 8). The smaller temperature range and anomalies close to the respective control run maxima (dark red horizontal lines) in the boosted heat waves are caused by the similar initial conditions in the ensembles.

In general, the temperature range increases for growing lead times over the first few weeks before the event, which is to be expected, due to growing chaotic perturbation. Consequently, small lead times yield temperature maxima similar to those in

the unperturbed reference heat wave in the control simulation. At the same time, the median Tx7d anomaly decreases with increasing lead times. Presumably, the highly anomalous anticyclonic anomaly is too weak or lost with too large lead times in many realizations and thereby also the extreme heat anomaly is reduced. As a result, we have a trade-off between using 1) too short lead times, resulting in little spread across members and thus heat waves that at most only marginally exceed the corresponding reference case in CESM-PI, and 2) too long lead times, thereby losing the anticyclonic anomaly that is driving the record heat waves in the corresponding reference simulation. In the example of Fig. 7b, the extreme heat wave anomaly is lost at lead times larger than 16 days. Nevertheless, this time scale of lead time is much longer than in typical numerical weather prediction (NWP) predictability experiments, because we here minimize the initial perturbation to a roundoff error to ensure physical consistency of the simulations, whereas NWP ensemble systems use larger perturbation to represent the uncertainty in the initial conditions. These results are consistent with ensemble boosting of heat waves over WRU and CNA (Fig. S15).

Here, we use the growing spread in the boosted ensembles to generate a large sample of very rare temperature anomalies at relatively low computational costs. The ensembles, however, cannot be easily interpreted in a probabilistic sense; that is, they are not part of an independent sample, since the boosted

members share very similar initial conditions. Nevertheless, ensemble boosting can provide storylines of very rare heat waves, explaining why some boosted members exceed even the highest temperature anomalies in the multimillennial control simulation. Therefore, we apply ensemble boosting to individual heat waves over WEU. We address the question of whether the previously identified drivers and their contribution to temperature variability of independent heat waves are the same as for temperature variability of related heat waves in the short term. First, we select 12 different heat waves, 6 out of the 10 hottest heat waves and 6 randomly chosen from the rest. We rerun these heat waves for 7 lead times every other day. Most boosted ensembles are started 7 days before, but due to technical reasons some are started only 6 days before the respective temperature peak (Fig. S14). As in Fig. 7, individual members exceed the reference temperature anomaly by up to 1–2 SD, but unsurprisingly most members are less pronounced. Interestingly, the highest temperature anomaly in the long control run is exceeded by multiple boosted heat waves. The respective lead times of the highest Tx7d anomaly vary across the different heat waves between 7 and 19 days, depending on the individual circulation field of the reference heat wave.

In a next step, we insert the boosted members into the former regression function (Fig. 6a) to test whether the three considered physical drivers, Z500, surface solar radiation (SSR), and soil liquid water to a depth of 0.2 m (SOIL) change their contribution to Tx7d anomalies in very rare heat waves, a few days before the temperature peak. Note that the same regression coefficients as above are used without recalibration. The predictors of 2-month mean SSR and SOIL are merged time series of the control run and the boosted members. In Fig. 9, the maximum Tx7d anomalies of all 12 boosted heat waves are plotted against their estimates from the regression model and added to the year-to-year maxima from Fig. 6a. The boosted ensembles cover large value ranges of Tx7d anomalies. The highest anomalies of each ensemble tend to be underestimated by the regression model. This can be caused either by unconsidered drivers or by a change of their contribution in the short term. We argue that the difference in the short-term temperature development between members is largely influenced by differences in circulation anomalies, because the 2-month mean anomalies of SSR and SOIL hardly change in the boosted ensembles (Fig. S16). Moreover, the Pearson correlation within each ensemble is stronger between maximum Tx7d and concurrent Z500 anomalies (0.58) than to SOIL (−0.12) or SSR anomalies (0.15). Note that these results do not contradict our previous findings about a decreasing impact of Z500 and increasing impact of land–atmosphere interactions on the most extreme heat waves (Fig. 5), as we here analyze the temperature development of related storylines.

To quantify the short-term impact of the physical drivers on heat waves over WEU, we conduct hierarchical partitioning on the 12 boosted heat waves, using mean Z500, SSR, and EF, averaged over the last 10 days before the temperature peak of the reference heat wave. Figure 10 illustrates the explained variance of maximum Tx7d anomaly for each heat wave, sorted by the maximum temperatures of the reference heat waves. The results reveal that variances of circulation and radiation explain the largest proportions of maximum temperature anomalies, whereas EF plays a minor role a few days before the

FIG. 9. As in Fig. 6a, but supplemented by maximum Tx7d anomalies of 12 heat waves in the CESM-PI simulation (colored points), used for ensemble boosting. Each boosted heat wave shows 350 points of the same color, giving the results of 50 members for seven different lead times every other day, starting one week before the reference temperature peak (black triangles). This results in a total of 4200 colored points of estimated maximum Tx7d anomalies, based on boosted heat wave ensembles.

temperature peak for the analyzed heat waves. However, absolute values of EF during very rare heat waves are already very low, undercutting the 10th percentile in the CESM-PI summer climatology. Comparing these results with previous findings, we conclude that the contributions of physical drivers to Tx7d anomalies differ between independent and related heat waves. Year-to-year variances within high Z500 levels barely impact very rare Tx7d anomalies, while low-level EF anomalies do. In other words, a few days to weeks before the temperature peak, when soil conditions hardly change, variances in Z500 can strongly impact Tx7d anomalies.

Altogether, the newly introduced ensemble boosting generates physically coherent storylines and a reasonable spread of 7-day running mean temperature anomalies for the different heat waves in the multimillennial CESM-PI simulation at low computational costs. Figure 8 contrasts the Tx7d anomalies estimated by all three presented storyline approaches. The fact that the temperature range of the boosted heat waves in WEU and WRU overlaps with the statistical estimates, based on GEV fits and the regression model of Tx7d anomalies, increases our confidence in the ensemble boosting method. Only for CNA the boosted temperature maxima are lower than the confidence interval of the fitted GEV distribution. Based on these storylines, we find that the contribution of heat wave drivers changes for the most extreme temperatures with circulation anomalies, becoming more impactful on temperature development as land–atmosphere interactions occur within a few days or weeks before the temperature peak.

5. Conclusions

In light of recent exceptional heat waves across the world, this study addresses the questions of how much more extreme

FIG. 10. As in Fig. 5, but column pairs show the explained variance of maximum Tx7d anomaly for all 12 boosted heat wave ensembles, sorted by the rank of their reference maximum Tx7d anomaly in the CESM-PI run, starting with the hottest event on the left. The contribution is separated into three driver components, (a) Z500, (b) surface solar radiation, and (c) evaporative fraction, in the previous 10 days before the maximum Tx7d anomaly is achieved. (d) The additive contribution of all three drivers.

the recent extreme events could have been relative to the corresponding climatology and what are their driving mechanisms. For this purpose, we apply three different approaches to develop storylines for very rare heat waves, based on a multimillennial preindustrial control simulation. We investigate three recently affected areas: central North America (CNA) including Chicago in 1995, western Europe (WEU) in 2003, and western Russia (WRU) in 2010.

The first approach fits a GEV distribution to Tx7d anomalies, revealing that return levels asymptotically approach an upper bound over the selected areas [e.g., 3.7°C (1.4 SD) in WEU]. Second, a regression model, using physical drivers such as 500-hPa geopotential height, shortwave radiation at the surface, and soil moisture anomalies as predictors is calibrated to the heat waves in the same simulations. Assuming a combination of idealized very extreme anomalies in all drivers, the model suggests temperature anomalies of similar magnitude. Finally, a novel ensemble boosting method is introduced to efficiently sample very rare heat extremes, providing again consistent anomalies. Thereby, most extreme heat wave scenarios suggest substantially higher heat wave anomalies than simulated in roughly 5000 years of control simulations.

All three approaches yield consistent storylines of very rare heat waves in the context of a preindustrial simulation, suggesting that 7-day running mean temperature anomalies in the preindustrial period could have exceeded the records of the historical outstanding heat waves for the three selected areas. The magnitude of the anomalies may still be model dependent and involves substantial uncertainties. However, the fact that the variability of the monthly summer temperatures, the GEV distribution for moderate heat waves, and the dependency of Tx7d anomalies on geopotential height in CESM-PI is in good agreement with ERA5 increases confidence in our results.

While the evaluation on observations cannot be done for very rare events, we find that very rare heat waves unseen in the observations are driven by the same physical drivers as the more moderate observed heat waves, albeit the same physical mechanisms show substantially higher anomalies, including persistent anticyclones and increased surface heat fluxes, due

to successive soil dehydration on seasonal time scales. The contribution of the used heat wave driver changes for the hottest and driest events, caused by growing impact of land–atmosphere interactions, which depends however on the individual boundary conditions of the heat wave. Furthermore, the selected drivers and hence the methods also depend on the region. The variability in the drivers could explain more than 70% of the annual temperature peaks for WEU and WRU, but only 39% for CNA so that the estimated temperature maxima might be underestimated by the regression model.

Finally, the analysis of very rare extremes includes major uncertainties as we push into temperature levels unseen in the observational record, based on our understanding of heat waves. The results presented here are probably model dependent, since small-scale mechanisms, including land–atmosphere feedbacks, must be parameterized, given the limited numerical resolution. However, this study primarily focuses on the suitability of different methodologies to quantify very rare heat extremes. The approaches further do not isolate and only implicitly account for processes like the build-up of heat in the boundary layer (Miralles et al. 2014) or adiabatic warming through subsidence (Bieli et al. 2015), which may differ between events despite the same Z500 anomaly. Furthermore, subgrid-scale contributions to heat wave due to, for example, urban heat island effects or land management practices (Davin et al. 2014) are not directly accounted for. The consistency of all presented methods increases our confidence in our applied techniques.

On that account, the far tails of temperature distribution remain uncertain, but the presented methods applied to such large datasets provide good samples that can help to construct storylines of very rare heat waves with magnitudes beyond the observed temperature range. Such storylines are potentially useful for stress testing ecosystems or socioeconomic systems to increase resilience to extreme climate stressors.

Acknowledgments. This work was supported by the Swiss National Science Foundation SNSF (Grant 200020_178778). We thank all the scientists, software engineers, and administrators who contributed to the development of CESM.

High-performance computing was performed with the ETH cluster Euler located at the Swiss national supercomputing centre CSCS.

Data availability statement. The CESM-PI control simulations are available at the ETH Research Collection (<https://doi.org/10.3929/ethz-b-000448872>). ERA5 reanalysis data used in this study can be accessed from the European Center for Medium-Range Weather Forecasts (ECMWF; <https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5>).

REFERENCES

- Angélil, O., and Coauthors, 2016: Comparing regional precipitation and temperature extremes in climate model and reanalysis products. *Wea. Climate Extremes*, **13**, 35–43, <https://doi.org/10.1016/j.wace.2016.07.001>.
- Barriopedro, D., and Coauthors, 2011: The hot summer of 2010: Redrawing the temperature record map of Europe. *Science*, **332**, 220–224, <https://doi.org/10.1126/science.1201224>.
- Bastos, A., C. M. Trigo, R. M. Trigo, and S. Running, 2014: Analysing the spatio-temporal impacts of the 2003 and 2010 extreme heatwaves on plant productivity in Europe. *Biogeosciences*, **11**, 3241, <https://doi.org/10.5194/bg-11-3421-2014>.
- Ben Alaya, B., F. Zwiers, and X. Zhang, 2018: Probable maximum precipitation: Its estimation and uncertainty quantification using bivariate extreme value analysis. *J. Hydrometeorol.*, **19**, 679–694, <https://doi.org/10.1175/JHM-D-17-0110.1>.
- Bieli, M., S. Pfahl, and H. Wernli, 2015: A Lagrangian investigation of hot and cold temperature extremes in Europe. *Quart. J. Roy. Meteor. Soc.*, **141**, 98–108, <https://doi.org/10.1002/qj.2339>.
- Black, E., M. Blackburn, G. Harrison, B. Hoskins, and J. Methven, 2004: Factors contributing to the summer 2003 European heatwave. *Weather*, **59**, 217–223, <https://doi.org/10.1256/wea.74.04>.
- Cassou, C., L. Terray, and A. S. Phillips, 2005: Tropical Atlantic influence on European heat waves. *J. Climate*, **18**, 2805–2811, <https://doi.org/10.1175/JCLI3506.1>.
- Changnon, S. A., K. E. Kunkel, and B. C. Reinke, 1996: Impacts and responses to the 1995 heat wave: A call to action. *Bull. Amer. Meteor. Soc.*, **77**, 1497–1506, [https://doi.org/10.1175/1520-0477\(1996\)077<1497:IARTTH>2.0.CO;2](https://doi.org/10.1175/1520-0477(1996)077<1497:IARTTH>2.0.CO;2).
- Chevan, A., and M. Sutherland, 1991: Hierarchical partitioning. *Amer. Stat.*, **45**, 90–96.
- Christidis, N., and P. A. Stott, 2014: Change in the odds of warm years and seasons due to anthropogenic influence on the climate. *J. Climate*, **27**, 2607–2621, <https://doi.org/10.1175/JCLI-D-13-00563.1>.
- Ciais, P., and Coauthors, 2005: Europe-wide reduction in primary productivity caused by the heat and drought in 2003. *Nature*, **437**, 529–533, <https://doi.org/10.1038/nature03972>.
- Clark, R. T., S. J. Brown, and J. M. Murphy, 2006: Modeling Northern Hemisphere summer heat extreme changes and their uncertainties using a physical ensemble of climate sensitivity experiments. *J. Climate*, **19**, 4418–4435, <https://doi.org/10.1175/JCLI3877.1>.
- Coles, S., J. Bawa, L. Trenner, and P. Dorazio, 2001: *An Introduction to Statistical Modeling of Extreme Values*. Springer, 224 pp.
- Colucci, J., 1985: Explosive cyclogenesis and large-scale circulation changes: Implications for atmospheric blocking. *J. Atmos. Sci.*, **42**, 2701–2717, [https://doi.org/10.1175/1520-0469\(1985\)042<2701:ECALSC>2.0.CO;2](https://doi.org/10.1175/1520-0469(1985)042<2701:ECALSC>2.0.CO;2).
- Coumou, D., J. Lehmann, and J. Beckmann, 2015: The weakening summer circulation in the Northern Hemisphere mid-latitudes. *Science*, **348**, 324–327, <https://doi.org/10.1126/science.1261768>.
- Davin, E. L., and Coauthors, 2014: Preferential cooling of hot extremes from cropland albedo management. *Proc. Natl. Acad. Sci. USA*, **111**, 9757–9761, <https://doi.org/10.1073/pnas.1317323111>.
- Davini, P., C. Cagnazzo, S. Gualdi, and A. Navarra, 2012: Bidimensional diagnostics, variability and trends of Northern Hemisphere blocking. *J. Climate*, **25**, 6496–6509, <https://doi.org/10.1175/JCLI-D-12-00032.1>.
- De Bono, A., P. Peduzzi, S. Kluser, and G. Giuliani, 2004: Impacts of summer 2003 heat wave in Europe. United Nations Environment Programme, 4 pp., <https://archive-ouverte.unige.ch/unige:32255>.
- Della-Marta, P. M., M. R. Haylock, J. Luterbacher, and H. Wanner, 2007: Doubled length of western European summer heat waves since 1880. *J. Geophys. Res.*, **112**, D15103, <https://doi.org/10.1029/2007JD008510>.
- De Sario, M., K. Katsouyanni, and P. Michelozzi, 2013: Climate change, extreme weather events, air pollution and respiratory health in Europe. *Eur. Respir. J.*, **42**, 826–843, <https://doi.org/10.1183/09031936.00074712>.
- Diffenbaugh, N. S., and M. Burke, 2019: Global warming has increased global economic inequality. *Proc. Natl. Acad. Sci. USA*, **116**, 9808–9813, <https://doi.org/10.1073/pnas.1816020116>.
- , and Coauthors, 2017: Quantifying the influence of global warming on unprecedented extreme climate events. *Proc. Natl. Acad. Sci. USA*, **114**, 4881–4886, <https://doi.org/10.1073/pnas.1618082114>.
- , D. Singh, and J. S. Mankin, 2018: Unprecedented climate events: Historical changes, aspirational targets, and national commitments. *Sci. Adv.*, **4**, eaao3354, <https://doi.org/10.1126/sciadv.aao3354>.
- Ding, Q., and B. Wang, 2005: Circumglobal teleconnection in the Northern Hemisphere summer. *J. Climate*, **18**, 3483–3505, <https://doi.org/10.1175/JCLI3473.1>.
- Dole, R., and Coauthors, 2011: Was there a basis for anticipating the 2010 Russian heat wave? *Geophys. Res. Lett.*, **38**, L06702, <https://doi.org/10.1029/2010GL046582>.
- Ehret, U., E. Zehe, V. Wulfmeyer, K. Warrach-Sagi, and J. Liebert, 2012: Should we apply bias correction to global and regional climate model data? *Hydrol. Earth Syst. Sci.*, **16**, 3391–3404, <https://doi.org/10.5194/hess-16-3391-2012>.
- Easterling, D. R., and Coauthors, 2000: Climate extremes: Observations, modeling, and impacts. *Science*, **289**, 2068–2074, <https://doi.org/10.1126/science.289.5487.2068>.
- Fischer, E. M., and C. Schär, 2010: Consistent geographical patterns of changes in high-impact European heatwaves. *Nat. Geosci.*, **3**, 398–403, <https://doi.org/10.1038/ngeo0866>.
- , and R. Knutti, 2015: Anthropogenic contribution to global occurrence of heavy-precipitation and high-temperature extremes. *Nat. Climate Change*, **5**, 560–564, <https://doi.org/10.1038/nclimate2617>.
- , S. I. Seneviratne, D. Lüthi, and C. Schär, 2007: Contribution of land-atmosphere coupling to recent European summer heat waves. *Geophys. Res. Lett.*, **34**, L06707, <https://doi.org/10.1029/2006GL029068>.
- Froelicher, T., E. M. Fischer, and N. Gruber, 2018: Marine heatwaves under global warming. *Nature*, **560**, 360–364, <https://doi.org/10.1038/s41586-018-0383-9>.
- García-Herrera, R., and Coauthors, 2010: A review of the European summer heat wave of 2003. *Crit. Rev. Environ. Sci. Technol.*, **40**, 267–306, <https://doi.org/10.1080/10643380802238137>.

- Gill, J. C., and B. D. Malamud, 2014: Reviewing and visualizing the interactions of natural hazards. *Rev. Geophys.*, **52**, 680–722, <https://doi.org/10.1002/2013RG000445>.
- Gilleland, E., and R. W. Katz, 2016: extRemes 2.0: An extreme value analysis package in R. *J. Stat. Softw.*, **72**, 1–39, <https://doi.org/10.18637/jss.v072.i08>.
- Hempel, S., K. Frieler, L. Warszawski, J. Schewe, and F. Piontek, 2013: A trend-preserving bias correction—The ISI-MIP approach. *Earth Syst. Dyn.*, **4**, 219–236, <https://doi.org/10.5194/esd-4-219-2013>.
- Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, <https://doi.org/10.1002/qj.3803>.
- Hogan, E., and Coauthors, 2019: Representation of U.S. warm temperature extremes in global climate model ensembles. *J. Climate*, **32**, 2591–2603, <https://doi.org/10.1175/JCLI-D-18-0075.1>.
- Horton, R., and Coauthors, 2016: A review of recent advances in research on extreme heat events. *Curr. Climate Change Rep.*, **2**, 242–259, <https://doi.org/10.1007/s40641-016-0042-x>.
- Huang, W. K., and Coauthors, 2016: Estimating changes in temperature extremes from millennial-scale climate simulations using generalized extreme value (GEV) distributions. *Adv. Stat. Climatol. Meteor. Oceanogr.*, **2**, 79–103, <https://doi.org/10.5194/ascmo-2-79-2016>.
- Hurrell, J. W., and Coauthors, 2013: The Community Earth System Model: A framework for collaborative research. *Bull. Amer. Meteor. Soc.*, **94**, 1339–1360, <https://doi.org/10.1175/BAMS-D-12-00121.1>.
- Imada, Y., M. Watanabe, H. Kawase, H. Shiogama, and M. Arai, 2019: The July 2018 high temperature event in Japan could not have happened without human-induced global warming. *SOLA*, **15A**, 8–12, <https://doi.org/10.2151/sola.15A-002>.
- Jäger, E., and S. Seneviratne, 2011: Impact of soil moisture–atmosphere coupling on European climate extremes and trends in a regional climate model. *Climate Dyn.*, **36**, 1919–1939, <https://doi.org/10.1007/s00382-010-0780-8>.
- King, A. D., T. P. Lane, B. J. Henley, and J. R. Brown, 2020: Global and regional impacts differ between transient and equilibrium warmer worlds. *Nat. Climate Change*, **10**, 42–47, <https://doi.org/10.1038/s41558-019-0658-7>.
- Kornhuber, K., D. Coumou, E. Vogel, C. Lesk, J. F. Donges, J. Lehmann, and R. M. Horton, 2020: Amplified Rossby waves enhanced risk of concurrent heatwaves in major breadbasket regions. *Nat. Climate Change*, **10**, 48–53, <https://doi.org/10.1038/s41558-019-0637-z>.
- Lewis, S. C., and D. J. Karoly, 2013: Anthropogenic contributions to Australia's record summer temperatures of 2013. *Geophys. Res. Lett.*, **40**, 3705–3709, <https://doi.org/10.1002/grl.50673>.
- , and A. D. King, 2017: Evolution of mean, variance and extremes in 21st century temperatures. *Wea. Climate Extremes*, **15**, 1–10, <https://doi.org/10.1016/j.wace.2016.11.002>.
- Lin, N., and K. Emanuel, 2016: Grey swan tropical cyclones. *Nat. Climate Change*, **6**, 206–211, <https://doi.org/10.1038/nclimate2777>.
- Luber, G., and M. McGeehin, 2008: Climate change and extreme heat events. *Amer. J. Prev. Med.*, **35**, 429–435, <https://doi.org/10.1016/j.amepre.2008.08.021>.
- Luo, D., J. Cha, L. Zhong, and A. Dai, 2014: A nonlinear multiscale interaction model for atmospheric blocking: The eddy-blocking matching mechanism. *Quart. J. Roy. Meteor. Soc.*, **140**, 1785–1808, <https://doi.org/10.1002/qj.2337>.
- Merte, S., 2017: Estimating heat wave-related mortality in Europe using singular spectrum analysis. *Climatic Change*, **142**, 321–330, <https://doi.org/10.1007/s10584-017-1937-9>.
- Miralles, D. G., A. J. Teuling, C. C. van Heerwaarden, and J. V. de Arellano, 2014: Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation. *Nat. Geosci.*, **7**, 345–349, <https://doi.org/10.1038/ngeo2141>.
- Orlov, A., and Coauthors, 2019: Economic losses of heat-induced reductions in outdoor worker productivity: A case study of Europe. *Econ. Disaster. Clim. Chang.*, **3**, 191–211, <https://doi.org/10.1007/s41885-019-00044-0>.
- Parente, J., M. G. Pereira, M. Amraoui, and E. M. Fischer, 2018: Heat waves in Portugal: Current regime, changes in future climate and impacts on extreme wildfires. *Sci. Total Environ.*, **631**, 534–549, <https://doi.org/10.1016/j.scitotenv.2018.03.044>.
- Paté-Cornell, E., 2012: On “black swans” and “perfect storms”: Risk analysis and management when statistics are not enough. *Risk Anal.: Int. J.*, **32**, 1832–1833, <https://doi.org/10.1111/j.1539-6924.2011.01787.x>.
- Perkins, S. E., L. V. Alexander, and J. R. Nairn, 2012: Increasing frequency, intensity and duration of observed global heatwaves and warm spells. *Geophys. Res. Lett.*, **39**, L20714, <https://doi.org/10.1029/2012GL053361>.
- Perkins-Kirkpatrick, S. E., and P. B. Gibson, 2017: Changes in regional heatwave characteristics as a function of increasing global temperature. *Sci. Rep.*, **7**, 12256, <https://doi.org/10.1038/s41598-017-12520-2>.
- Petoukhov, V., S. Rahmstorf, S. Petri, and H. J. Schellnhuber, 2013: Quasiresonant amplification of planetary waves and recent Northern Hemisphere weather extremes. *Proc. Natl. Acad. Sci. USA*, **110**, 5336–5341, <https://doi.org/10.1073/pnas.1222000110>.
- Pfahl, S., and H. Wernli, 2012: Quantifying the relevance of atmospheric blocking for co-located temperature extremes in the Northern Hemisphere on (sub-) daily time scales. *Geophys. Res. Lett.*, **39**, L12807, <https://doi.org/10.1029/2012GL052261>.
- Pfleiderer, P., C. F. Schleussner, K. Kornhuber, and D. Coumou, 2019: Summer weather becomes more persistent in a 2°C world. *Nat. Climate Change*, **9**, 666–671, <https://doi.org/10.1038/s41558-019-0555-0>.
- Plotkin, D. A., and Coauthors, 2019: Maximizing simulated tropical cyclone intensity with action minimization. *J. Adv. Model. Earth Syst.*, **11**, 863–891, <https://doi.org/10.1029/2018MS001419>.
- Quesada, B., R. Vautard, P. Yiou, M. Hirschi, and S. I. Seneviratne, 2012: Asymmetric European summer heat predictability from wet and dry southern winters and springs. *Nat. Climate Change*, **2**, 736–741, <https://doi.org/10.1038/nclimate1536>.
- Ragone, F., J. Wouters, and F. Bouchet, 2018: Computation of extreme heat waves in climate models using a large deviation algorithm. *Proc. Natl. Acad. Sci. USA*, **115**, 24–29, <https://doi.org/10.1073/pnas.1712645115>.
- Robine, J., 2008: Death toll exceeded 70,000 in Europe during the summer of 2003. *C. R. Biol.*, **331**, 171–178, <https://doi.org/10.1016/j.crv.2007.12.001>.
- Schumacher, D. L., and Coauthors, 2019: Amplification of mega-heatwaves through heat torrents fuelled by upwind drought. *Nat. Geosci.*, **12**, 712–717, <https://doi.org/10.1038/s41561-019-0431-6>.
- Seneviratne, S. I., and Coauthors, 2010: Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Sci. Rev.*, **99**, 125–161, <https://doi.org/10.1016/j.earscirev.2010.02.004>.
- Shaposhnikov, D., and Coauthors, 2014: Mortality related to air pollution with the Moscow heat wave and wildfire of

2010. *Epidemiology*, **25**, 359–364, <https://doi.org/10.1097/EDE.0000000000000090>.
- Shepherd, T. G., 2014: Atmospheric circulation as a source of uncertainty in climate change projections. *Nat. Geosci.*, **7**, 703–708, <https://doi.org/10.1038/ngeo2253>.
- , 2019: Storyline approach to the construction of regional climate change information. *Proc. Roy. Soc. London*, **A475**, 20190013, <https://doi.org/10.1098/rspa.2019.0013>.
- , and Coauthors, 2018: Storylines: An alternative approach to representing uncertainty in physical aspects of climate change. *Climatic Change*, **151**, 555–571, <https://doi.org/10.1007/s10584-018-2317-9>.
- Sippel, S., and F. E. Otto, 2014: Beyond climatological extremes—Assessing how the odds of hydrometeorological extreme events in South-East Europe change in a warming climate. *Climatic Change*, **125**, 381–398, <https://doi.org/10.1007/s10584-014-1153-9>.
- Sriver, R. L., C. E. Forest, and K. Keller, 2015: Effects of initial conditions uncertainty on regional climate variability: An analysis using a low-resolution CESM ensemble. *Geophys. Res. Lett.*, **42**, 5468–5476, <https://doi.org/10.1002/2015GL064546>.
- Stein, J. L., and S. Stein, 2014: Gray swans: Comparison of natural and financial hazard assessment and mitigation. *Nat. Hazards*, **72**, 1279–1297, <https://doi.org/10.1007/s11069-012-0388-x>.
- Stott, P. A., D. A. Stone, and M. R. Allen, 2004: Human contribution to the European heatwave 2003. *Nature*, **432**, 610–614, <https://doi.org/10.1038/nature03089>.
- Teng, H., G. Branstator, A. B. Tawfik, and P. Callaghan, 2019: Circumglobal response to prescribed soil moisture over North America. *J. Climate*, **32**, 4252–4545, <https://doi.org/10.1175/JCLI-D-18-0823.1>.
- van der Velde, M., F. N. Tubiello, and A. Vrieling, 2012: Impacts of extreme weather on wheat and maize in France: Evaluating regional crop simulations against observed data. *Climatic Change*, **113**, 751–765, <https://doi.org/10.1007/s10584-011-0368-2>.
- Vogel, M. M., and Coauthors, 2019: Concurrent 2018 hot extremes across Northern Hemisphere due to human-induced climate change. *Earth's Future*, **7**, 692–703, <https://doi.org/10.1029/2019EF001189>.
- Wang, J., and Coauthors, 2016: Evaluation of dynamically downscaled extreme temperature using a spatially-aggregated generalized extreme value (GEV) model. *Climate Dyn.*, **47**, 2833–2849, <https://doi.org/10.1007/s00382-016-3000-3>.
- Webber, R. J., and Coauthors, 2019: Practical rare event sampling for extreme mesoscale weather. *Chaos*, **29**, 053109, <https://doi.org/10.1063/1.5081461>.
- Wehrli, K., B. P. Guillod, M. Hauser, M. Leclair, and S. I. Seneviratne, 2019: Identifying key driving processes of major recent heat waves. *J. Geophys. Res. Atmos.*, **124**, 11 746–11 765, <https://doi.org/10.1029/2019JD030635>.
- Woollings, T., B. Hoskins, M. Blackburn, and P. Berrisford, 2008: A new Rossby wave-breaking interpretation of the North Atlantic Oscillation. *J. Atmos. Sci.*, **65**, 609–626, <https://doi.org/10.1175/2007JAS2347.1>.
- Yiou, P., and A. Jézéquel, 2020: Simulation of extreme heat waves with empirical importance sampling. *Geosci. Model Dev.*, **13**, 763–781, <https://doi.org/10.5194/gmd-13-763-2020>.