Multimodel Errors and Emergence Times in Climate Attribution Studies

Philippe Naveau, Soulivanh Thao

Multimodel Errors and Emergence Times in Climate Attribution Studies

PHILIPPE NAVEAU[a] AND SOULIVANH THAO[a]

[a] *Laboratoire des Sciences du Climat et de l'Environnement, EstimR Team, IPSL-CNRS-CEA-UVSQ, Gif-sur-Yvette, France*

(Manuscript received 30 April 2021, in final form 2 March 2022)

ABSTRACT: Global climate models, like any in silico numerical experiments, are affected by different types of bias. Uncertainty quantification remains a challenge in any climate detection and attribution analysis. A fundamental methodological question is to determine which statistical summaries, while bringing relevant signals, can be robust with respect to multimodel errors. In this paper, we propose a simple statistical framework that significantly improves signal detection in climate attribution studies. We show that the complex bias correction step can be entirely bypassed for models for which bias between the simulated and unobserved counterfactual worlds is the same as between the simulated and unobserved factual worlds. To illustrate our approach, we infer emergence times in precipitation from the CMIP5 and CMIP6 archives. The detected anthropogenic signal in yearly maxima of daily precipitation clearly emerges at the beginning of the twenty-first century. In addition, no CMIP model seems to outperform the others and a weighted linear combination of all improves the estimation of emergence times.

SIGNIFICANCE STATEMENT: We show that the bias in multimodel global climate simulations can be efficiently handled when the appropriate metric is chosen. This metric leads to an easy-to-implement statistical procedure based on a checkable assumption. This allows us to demonstrate that optimal convex combinations of CMIP outputs can improve the signal strength in finding emergence times. Our data analysis procedure is applied to yearly maximum of precipitation from CMIP5 and CMIP6 databases. The attribution of the anthropogenic forcing clearly emerges in extreme precipitation at the beginning of the twenty-first century.

KEYWORDS: Bias; Statistics; Uncertainty; General circulation models; Model comparison; Model errors; Model evaluation/performance

1. Introduction

Global climate model outputs like any numerical simulations correspond to an approximation of the true system under study, here the climate system. In the realm of detection and attribution (D&A), either in a transient setup or in the context of extreme event attribution (EEA), numerous review studies (see, e.g., Chen et al. 2018; Stott et al. 2016; Shepherd 2016) list different sources of variability, uncertainties, and errors. In particular, these reviews highlight that model error in numerical experiments like the Coupled Model Intercomparison Project (CMIP) can be large and has to be taken into account in any D&A statistical analysis (see, e.g., Knutti et al. 2019; National Academies of Sciences Engineering and Medicine 2016).

To address the issue of multimodel error in attribution studies, we need to go back to the origin of D&A. This research field aims to answer questions related to *relative* changes between two worlds. In EEA, a factual scenario of conditions that occurred around the time of a specific event[1] is compared to the probability of the same event but under a counterfactual scenario in which anthropogenic emissions had never occurred (see, e.g., Angélil et al. 2017). In D&A with transient runs, the two worlds correspond to global coupled climate runs with all forcings (ALL) and with only natural forcings (NAT), respectively (see, e.g., Hegerl and Zwiers 2011). To combine model error uncertainties, various authors (see, e.g., Lorenz et al. 2018) have noticed that giving equal weight to each available model projection may be suboptimal. In addition, model interdependencies have been identified as an important issue in uncertainties analysis (see, e.g., Abramowitz et al. 2019). To integrate multimodel error into the EEA, we leverage a hypothesis from the bias-correction community to propose an easy-to-implement strategy that, under well-identified conditions, has the main advantage of bypassing multimodel error. In addition, the problem of model interdependencies is handled by fixing a robust referential invariant for this issue. Our main application is the inference of emergence times in yearly maxima of daily precipitation and temperatures from the CMIP database.

To close this introduction on multimodel error in D&A, we note that various articles (e.g., van Oldenborgh et al. 2021) have proposed strategies to model extreme precipitation with multiplicative parametric models based on generalized extreme value and generalized Pareto distributions (see, e.g., Coles 2001). Bellprat et al. (2019) also promoted correction techniques to assess the probabilities of extreme event occurrences. These authors recalibrated an ensemble by fitting a Gaussian regression model based on well-chosen covariates and multiplicative correction factors [see their Eq. (2)]. In this work, we explore a different road, and we propose a statistical treatment of

[1] The word "event" can have different meanings. In this study, we follow the definition used in probability theory, i.e., an "event" is a set of outcomes of an experiment (a subset of the sample space) to which a probability can be assigned; for example, $\{X > u\}$ is an event. Here, the random variable X is represented by a capital letter while the constant scalar u is a non-capital letter. Also, we make the classical distinction between a "realization" (a draw), say x, and its random variable X.

Corresponding author: Philippe Naveau, philippe.naveau@lsce.ipsl.fr

DOI: 10.1175/JCLI-D-21-0332.1

multimodel bias with a nonparametric approach in a nonstationary context. By slightly changing the definition of the event of interest, we can completely bypass parametric modeling and avoid a nonstationary bias correction step. To explain in detail this new approach, clear notation, definitions, and assumptions need to be introduced.

2. Methods

a. Notation and assumptions

A common hypothesis in most D&A studies is that, although numerical models may not be able to exactly reproduce the true world, one can expect that any appropriate bias correction technique for a given numerical model should be applied in the same way to correct the factual and counterfactual worlds. For example, if factual temperature runs from a given model are too warm with respect to recorded measurements during a specific period and have to be corrected, say by 1 K in the factual world, then this bias of 1 K has to be corrected in the counterfactual world of this model during the same specific epoch. In practice, such a hypothesis can be challenged and needs to be assessed with caution (see, e.g., Maraun et al. 2017). Due to the lack of records in the counterfactual world, the assessment cannot be done without carefully designing the appropriate metric. To explain this point, we need some mathematical notation. Let Z_t and X_t be the same real-valued continuous random variable of interest for the year t but from the hypothetical true factual and counterfactual worlds with cumulative distribution functions (CDFs) $F_t(z) = P(Z_t \leq z)$ and $G_t(x) = P(X_t \leq x)$, respectively. In simulation studies, these two random variables are never available because perfect factual and counterfactual distributions cannot be exactly reproduced. Instead, imperfect ensemble outputs, say from M different numerical model experiments are available, and we denote this by $Z_t^{(m)}$ and $X_t^{(m)}$ the factual and counterfactual versions from model m and CDFs $F_t^{(m)}(z) = P(Z_t^{(m)} \leq z)$ and $G_t^{(m)}(x) = P(X_t^{(m)} \leq x)$, respectively.

A popular approach in the bias correction literature for univariate continuous random variables is the quantile mapping transform that matches two random variables with different distributions (e.g., Maraun et al. 2017; Cannon 2018). A positive aspect of quantile mapping is its theoretical basis. It can be viewed as the solution of an optimal transport problem (e.g., Robin et al. 2019, 2017) and it has been tested in various settings (e.g., for precipitation downscaling; Kallache et al. 2011). Quantile mapping can be adapted to our framework in the following way. As any continuous real-valued random variable, the climate model output $X_t^{(m)}$ can *always* be transformed, in a distributional sense, into the unobserved true counterfactual random variable X as follows:

$$X_t \overset{d}{=} \left(G_t^{\leftarrow} \circ G_t^{(m)}\right)\left(X_t^{(m)}\right),$$

where $\overset{d}{=}$ corresponds to an equality in distribution and $G_t^{\leftarrow}(\cdot)$ represents the inverse of G_t (i.e., its quantile function). The same type of operator can be implemented in the factual world, that is,

$$Z_t \overset{d}{=} \left(F_t^{\leftarrow} \circ F_t^{(m)}\right)\left(Z_t^{(m)}\right).$$

From these expressions of X_t and Z_t, it is natural to define the following hypothesis: assumption A holds for model m if

$$F_t^{\leftarrow} \circ F_t^{(m)} = G_t^{\leftarrow} \circ G_t^{(m)}, \quad \text{for } t \in 1, \ldots, T. \tag{1}$$

One of our main goals is to identify and work with CMIP models that satisfy assumption A. Equation (1) means that the bias between the simulated and unobserved counterfactual worlds is the same as between the simulated and unobserved factual worlds. Note that (1) allows for nonlinear bias correction. For example, heavy rainfall can have different tail behaviors in the observed and model runs [see, e.g., Coles (2001) for an introduction to upper tails modeling]. It is also important to notice that, although none of the available climate models may exactly satisfy assumption A, it makes sense to select models according to how closely they satisfy assumption A. Another key point is that assumption A is related to a specific variable. For the same climate model, assumption A can be valid for mean hemispheric temperatures, but incorrect for heavy rainfall over a specific region. Another feature of assumption A is the temporal indexing. The subscript t makes the notation complex, but it allows for having different bias corrections for different years, and brings flexibility. The cooling effect of volcanic forcing, like Pinatubo in 1991, can be included in our bias correction approach. The same could be said for slow changes due to solar forcing. So, under assumption A, the hypothesis of temporal stationarity is not needed in our framework.

b. Quantities of interest

In EEA studies, most researchers [see the bibliography in the review articles by Stott et al. (2016) and Naveau et al. (2020)] aim to contrast differences between two worlds (the factual and counterfactual worlds). In particular, making the distinction between the two survival functions

$$P(X_t > u) \quad \text{and} \quad P(Z_t > u) \tag{2}$$

has been a recurrent theme in EEA. To explain our statistical approach, we can start by asking a typical hydrological EEA question. For the current year t, are precipitation intensities in the factual world heavier than those produced in the counterfactual world? If two random precipitation intensities have the same distribution in the factual and counterfactual worlds, then the probability of observing the event $(Z_t > X_t)$ would be the same as that for $(Z_t < X_t)$ for any given year, and consequently $P(Z_t > X_t) = 0.5$ in this case.[2] If, instead, factual

[2] Independence between X_t and Z_t is not necessary to get $P(Z_t > X_t) = 0.5$. As the sum $P(Z_t > X_t) + P(Z_t < X_t)$ is always equal to one for continuous random variables, the only requirement is that $P(Z_t > X_t) = P(Z_t < X_t)$. This is always true whenever $(Z_t, X_t) \overset{d}{=} (X_t, Z_t)$, e.g., if (Z_t, X_t) follows a standardized and correlated bivariate Gaussian vector.

rainfalls are heavier than counterfactual ones, this will imply that the chance of Z_t being greater than X_t is greater than 0.5. To make the link with the classical EEA expressions defined by (2), we can look at the special case where the threshold u in (2) is chosen to be equal to a random draw from G_t. This choice leads to our definition of two simple probabilities:

$$q_0 = \frac{1}{2} \text{ and } q_t = P(Z_t > X_t). \tag{3}$$

These two probabilities have many advantages. They are invariant with respect to nondecreasing changes. For example, if both X and Z are simultaneously multiplied by two, then q_t remains the same. This is a critical feature when climate models are bias corrected. More precisely, under assumption A, we always have (see appendix B for a proof)

$$q_t = P\left(Z_t^{(m)} > X_t^{(m)}\right), \quad \text{for all} \quad t = 1, \ldots, T. \tag{4}$$

The fact that the left-hand part of this equation does not depend on m under assumption A is fundamental in this work. The remaining part of this article is to explain its consequences, its applicability, and its validity within the CMIP database.

Under assumption A, we do not need to observe X_t and Z_t to compute q_t. This probability can be obtained directly from realizations of $X_t^{(m)}$ and $Z_t^{(m)}$. The most important consequence of assumption A is that biased models' outputs do not need to be corrected. Practically, this also implies that we do not need to know the CDFs, F_t and G_t, to estimate q_t. In addition, q_t in (3) is always equal to 0.5 whenever the true factual and counterfactual worlds are exchangeable. Consequently, no inference is required in this case. Last but not least, the value of 0.5 can be used a reference point when the factual and counterfactual start to differ.

c. *Inference in the transient case*

The temporal indexing t in the probability q_t defined by (2) can be interpreted in two different ways. One can freeze the time to the current year, and this leads to the so-called event attribution realm. Large ensembles of simulations of this given year are classically drawn from both factual and counterfactual worlds (see, e.g., Stott et al. 2016) to estimate probabilities like q_t. Another setup is to allow years to spread over a long time period, say from the preindustrial epoch to 2100. This so-called transient case corresponds to the framework of GCM experiments where two types of runs are compared, say NAT and ALL runs. In this paper, we focus on the transient case, but all statistical techniques developed here can be easily transferred to the EEA setup (see, e.g., Naveau et al. 2020). In terms of EEA terminology, our analysis belongs to the so-called *unconditional* class, a term found in Knutson et al. (2017). In particular we focus on unconditional events of the type $\{Z_t > X_t\}$. Knutson et al. (2017) defined this category to highlight the contrast with "conditional attribution." In other words, events like $\{X > u\}$, or like $\{X > u|C\}$ where the threshold u and the conditioning C (say a SST field) are chosen relatively to observed realizations, will not be treated in this paper.

The CMIP experiments are recognized worldwide as a valuable repository of climate simulations. This database contains numerous simulations and has the advantage of being global. The 16 model runs used in this work are listed in Table 1.[3] Their main drawbacks are the model uncertainties, small ensemble sample sizes, and spatial resolution, which can be too coarse for some applications. Statistically, a subtle point is the transient nature of these simulations. This implies that factual runs in CMIP contain some nonlinear trends that should be taken into account in the statistical analysis (see, e.g., Kharin and Zwiers 2000). As the ALL run $Z_t^{(m)}$ distribution may change over time, we estimate the time-varying $q_t^{(m)} = P(Z_t^{(m)} > X_t^{(m)})$ using a nonparametric regression approach. A classical kernel regression approach (Nadaraya 1964; Watson 1964) leads to the following estimator:

$$\hat{q}_t^{(m)} = \frac{1}{\sum K_h(t - t_j)} \sum_{j=1}^{J} K_h(t - t_j)\, \mathbb{G}_I^{(m)}\left(Z_{t_j}^{(m)}\right), \tag{5}$$

where the positive function $K_h(\cdot)$ corresponds to a weighting kernel with bandwidth h, and $\mathbb{G}_I^{(m)}(\cdot)$ represents any estimator of $G_t^{(m)}(\cdot)$. In appendix B, the choices of the kernel and $\mathbb{G}_I^{(m)}(\cdot)$ are discussed and the statistical arguments needed to build asymptotic confidence intervals for q_t are given.

At this stage, assumption A has not been used yet. If assumption A was satisfied, it would be straightforward to move, via (4), from $\hat{q}_t^{(m)}$ to a common estimator of q_t. To combine climate model outputs, the ability to satisfy assumption A for a given model m will be key.

3. Merging climate model simulations

The main roadblock for assessing the quality of simulated runs is that we will never observe draws from X_t, only measurements with observational errors at spatial scales different from those represented by climate models. This lack of data is even worse for Z_t, the hypothetical world of an unperturbed and never observed climate. To bypass this difficulty, we assume that there exists a time period, say $\mathcal{T}$, during which the ALL and NAT worlds were identical (in distribution). In practice, this corresponds to the preindustrial period for which we assume

$$F_t = G_t, \quad \text{for all years} \quad t \in \mathcal{T}. \tag{6}$$

Again, we do not need to assume that $F_t = F_{t+1}$ as natural forcings may change in time, even at the annual scale, such as after the 1883 Krakatoa eruption. In section 4, we define the preindustrial epoch $\mathcal{T}$ as $\mathcal{T} = \{1850, \ldots, 1900\}$. Then, we can always write

$$q_t = P(Z_t > X_t) = \frac{1}{2}, \quad \text{for any year} \quad t \in \mathcal{T}. \tag{7}$$

[3] After the year 2006, we analyze CMIP precipitation under their respective worst-case scenarios (RCP8.5 and SSP5–8.5).

TABLE 1. List of the 16 CMIP runs used in this study. For each run, the last column shows the percentage of grid points for which the climate modal is rejected according to the Anderson–Darling p value below 0.2 (see, e.g., Fig. 1). The fourth column represents the global average estimated weight for each model (see, e.g., Fig. C2).

Institute	Name	Runs	Global weights	Percent of p value < 0.2
CMIP6 climate models				
CCCma	CanESM5	r10i1p1f1	0.06	0.21
CNRM-CERFACS	CNRM-CM6-1	r1i1p1f2	0.07	0.21
IPSL	IPSL-CM6A-LR	r1i1p1f1	0.06	0.19
MRI	MRI-ESM2-0	r1i1p1f1	0.07	0.19
CMIP5 climate models				
CCCma	CanESM2	r1i1p1	0.06	0.20
CNRM-CERFACS	CNRM-CM5	r1i1p1	0.07	0.20
CSIRO-BOM	ACCESS1.3	r1i1p1	0.07	0.20
CSIRO-QCCCE	CSIRO-Mk3.6.0	r1i1p1	0.06	0.20
IPSL	IPSL-CM5A-LR	r1i1p1	0.06	0.21
IPSL	IPSL-CM5A-MR	r1i1p1	0.06	0.20
MIROC	MIROC-ESM	r1i1p1	0.06	0.20
MIROC	MIROC-ESM-CHEM	r1i1p1	0.06	0.24
MRI	MRI-CGCM3	r1i1p1	0.06	0.20
NCAR	CCSM4	r1i1p1	0.06	0.20
NCC	NorESM1-M	r1i1p1	0.07	0.19
NSF-DOE-NCAR	CESM1-CAM5	r1i1p1	0.07	0.19

This equality is parameter-free, and it does not depend on model m. There is nothing to estimate, so no inferential error needs to be taken into account. As already highlighted in section 2b, the bivariate vector (X_t, Z_t) does not have to be stationary in t; only exchangeability between X_t and Z_t is required. This last condition is always satisfied as X_t and Z_t are not computer simulated. They just represent *conceptual* independent draws from a thought experiment of two possible climate trajectories in preindustrial times. For simulated runs from model m, we also expect to have $P(Z_t^{(m)} > X_t^{(m)}) = 0.5$ for any year $t \in \mathcal{T}$ if the two continuous random variables $Z_t^{(m)}$ and $X_t^{(m)}$ are exchangeable (label free) during the preindustrial period. Exchangeability is a weak hypothesis with respect to the issue of model interdependencies studied by Abramowitz et al. (2019). Concerning the validity of $P(Z_t^{(m)} > X_t^{(m)}) = 0.5$ in regard to a given model, we leverage the following fact. During the time period $\mathcal{T}$, we have $F_t = G_t$, and consequently, the following equivalence is always true:

$$F_t^{(m)} = G_t^{(m)}, \forall t \in \mathcal{T} \Leftrightarrow \text{assumption A holds for any } t \in \mathcal{T}.$$

Hence, checking that $F_t^{(m)}$ is equal to $G_t^{(m)}$ for each model m during the preindustrial epoch appears as the appropriate step. A simple scatterplot between the ranked $X_t^{(m)}$ and the ranked $Z_t^{(m)}$ should be close to a straight line. As an example, yearly maxima of CMIP daily precipitation for two randomly chosen grid points near Oxford in Great Britain and Hohenpeissenberg in Germany are analyzed in Fig. 1 (see also the left panel of Fig. C1 in appendix C). Overall, most CMIP climate model runs behave appropriately with respect to simulated precipitation at the Oxford and Hohenpeissenberg grid points. The p values of the two-sample Anderson–Darling test (Pettitt 1976; Anderson and Darling 1952) are indicated in the white boxes (see also Fig. C3 for p values over the globe). As expected, precipitation ranges vary strongly among models; for example, compare the precipitation spread between IPSL-CM5A-LR and CERFACS-CNRM-CM6 for the Oxford grid point. By construction, this difference is not an issue because we always look at relative changes within a CMIP run. According to (7), we also expect $q_t^{(m)}$ to be close to half for $t \in \mathcal{T}$. This can be used to weight and merge our 16 models.

Binary events like $\{Z_t^{(m)} > X_t^{(m)}\}$ during the preindustrial period are the building blocks of $q_t^{(m)} = P(Z_t^{(m)} > X_t^{(m)})$. In our setup, the reference distribution is the Bernoulli distribution with probability of success $q_t = 0.5$ and the competitor corresponds to a Bernoulli distribution with probability of success $q_t^{(m)}$. A statistical tool is needed to differentiate these two Bernoulli distributions. The Kullback–Leibler divergence (see, e.g., Burnham and Anderson 1998; Naveau et al. 2014) compares two distributions by calculating the expectation of the logarithmic difference between a target probability and a competitor, where the expectation is obtained with respect to the target density; see appendix D for a detailed discussion about Bernoulli modeling and the intermodel exchangeability assumption (see also Haughton et al. 2015; Knutti et al. 2009). As any Kullback–Leibler divergence is convex, it is natural to merge our estimates as a convex combination defined in the following way:

$$\hat{q}_t = \sum_{m=1}^{M} w_m \times \hat{q}_t^{(m)}, \quad \text{with} \quad w_i \geq 0 \text{ and } w_1 + \cdots + w_M = 1. \tag{8}$$

Concerning the Oxford and Hohenpeissenberg grid points, their associated $\hat{q}_t^{(m)}$ values are displayed in Fig. 2. Each panel represents one CMIP climate model from Table 1. In each

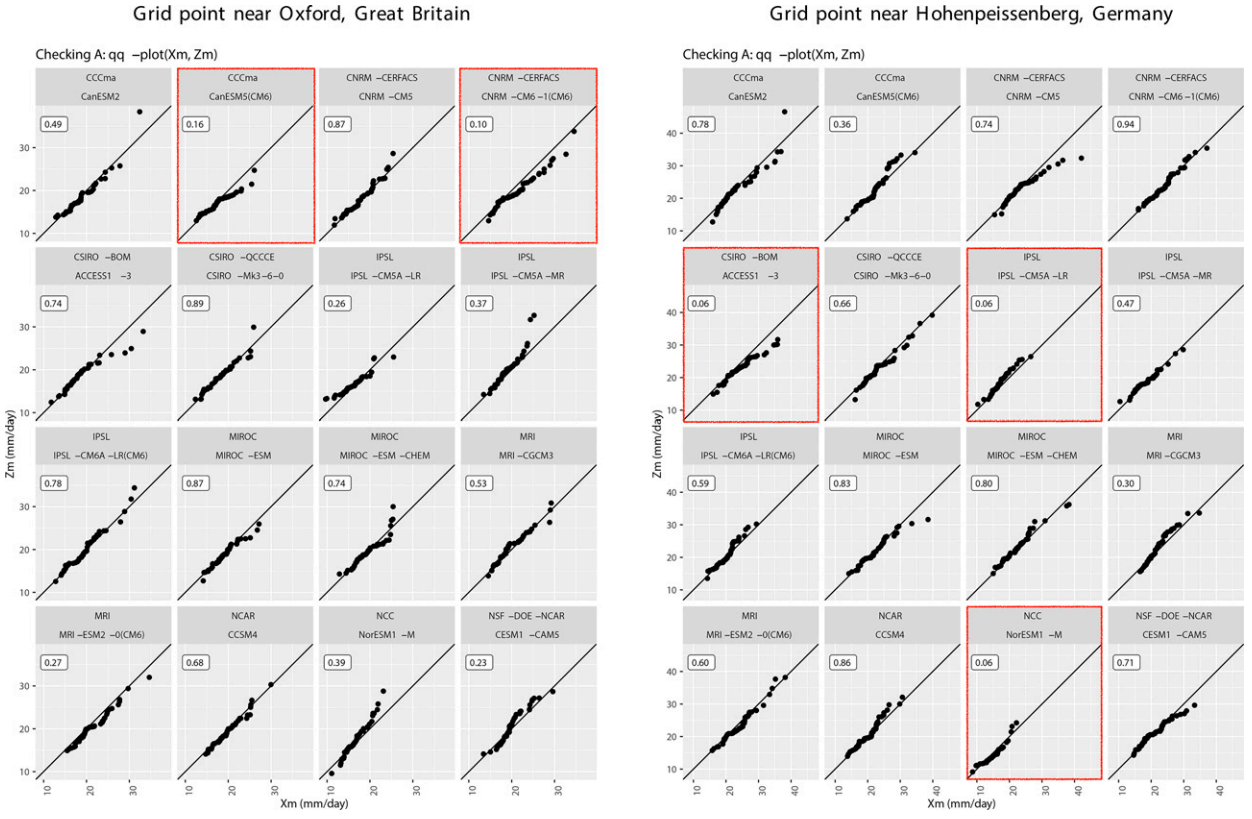

FIG. 1. Analysis of yearly maxima of daily precipitation from 16 CMIP climate runs, providing a visual check to know if, within each of the 16 models, NAT and ALL runs have similar distributions over the preindustrial period defined as years in $\mathcal{T} = \{1850, \ldots, 1900\}$ (left) at the grid point near Oxford and (right) at the grid point near Hohenpeissenberg. The x axis corresponds to the ranked yearly maxima of daily precipitation for the ALL run; the same quantity is displayed on the y axis, but for the NAT run. Each panel represents a CMIP model. If the points are aligned along the black diagonal, then NAT and ALL distributions can be considered similar. Each white box contains the p values of the two-sample Anderson–Darling test (Pettitt 1976; Anderson and Darling 1952). The red subpanels have a p value lower than 0.2 and the associated models are then discarded from the corresponding grid point.

panel, the x axis spans the year t = 1850 to t = 2100 and the y axis corresponds to the probability $\hat{q}_t^{(m)}$ obtained from (5). The departure from the horizontal line at $q_t = 0.5$ indicates a change between factual runs and counterfactual trajectories without anthropogenic forcing. All panels show a smooth increase in $\hat{q}_t^{(m)}$ over time, but each model appears to give a different speed and amplitude of change. Intermodel variability appears to be large and combining model errors is necessary.

To compute the weights $(w_1, \ldots, w_M)^{\mathrm{T}}$, we implement a two-step procedure. Small p values of the two-sample Anderson–Darling test in the previous section (see white boxes in Fig. 1) highlighted a poor fit. For this reason, our first step is to remove all models that have Anderson–Darling p values smaller than 20%, a very conservative rejection rate. The red boxes in Fig. 2 indicate these models, which will be removed from the merging. This first step allows us to treat the rare but possible case when all models are wrong. In such a case, the grid point is removed (in practice, this never occurs in our example). Our second step is to simply find the weights of selected models that minimize the Kullback–Leibler divergence between $\hat{q}_{t\in\mathcal{T}}$ and $q_{t\in\mathcal{T}} = (0.5,\ldots,0.5)^{\mathrm{T}}$, under the constraint $w_1 + \cdots + w_M = 1$.

Concerning CMIP yearly maxima of precipitation around the two grid points (Oxford and Hohenpeissenberg), the 90% green confidence band in Fig. 3 represents the estimate of the convex weighted combination (i.e., of $\hat{q}_t$) with the weights shown in Fig. 2. In Fig. 3, the 90% red confidence band corresponds to the model with the highest Anderson-Darling p value. As expected, combining estimates of q_t reduces the confidence bandwidth. By construction, $\hat{q}_t$ follows well the reference horizontal line centered at 0.5 during the preindustrial period. The departure from this horizontal gray line becomes statistically significant around the year 2000. Overall, the estimate $\hat{q}_t$ has a smooth trajectory over time and the detected signal is strong in 2100. A clear indication that anthropogenic forcing is projected to cause changes in precipitation intensities at this location.

4. Emergence times in yearly precipitation maxima

Our methodology can be applied to any type of real-valued continuous atmospheric variable: temperatures, wind speeds, precipitation intensities and others. As many EEA studies have already focused on temperature, we chose to study the

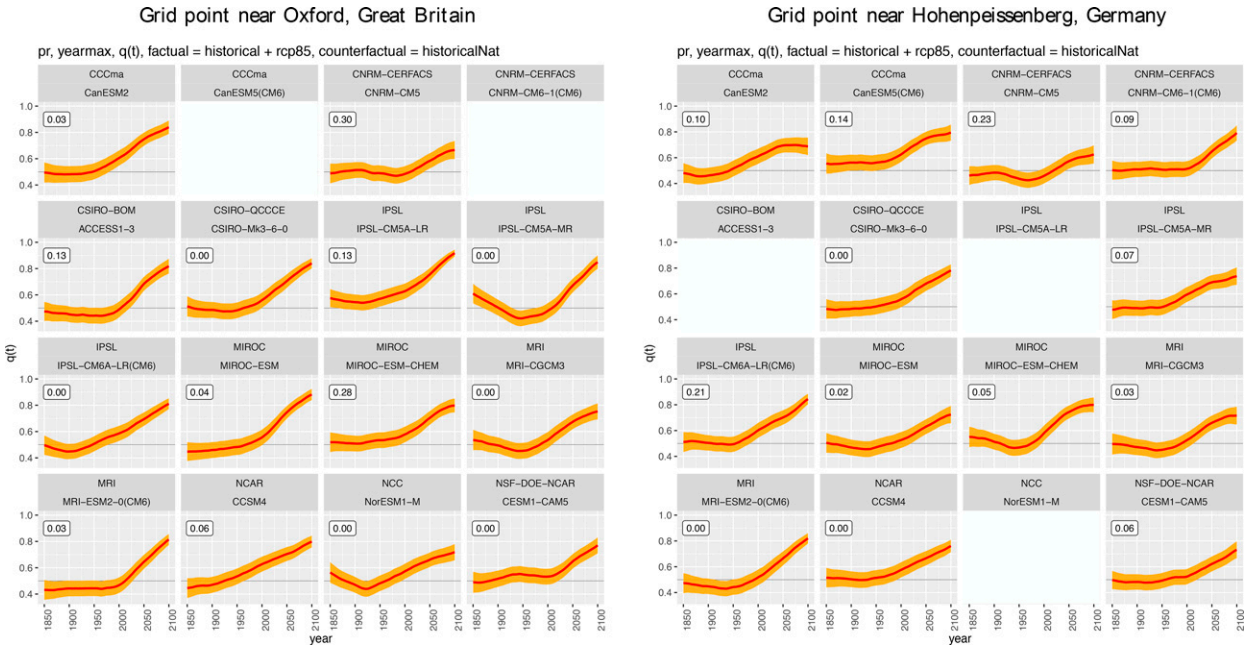

FIG. 2. As in Fig. 1, but in each panel the x axis corresponds to the years from $t = 1850$ to $t = 2100$ (RCP8.5 or SSP5–8.5 after 2006) and the y axis represents the estimate of q_t defined by (5). Each panel corresponds to a specific CMIP model setup (see Table 1). The shaded area denotes 90% confidence intervals around the mean estimate of q_t. The horizontal black line centered on 0.5 corresponds to the null hypothesis where the factual and counterfactual worlds are indistinguishable. Each white box contains the weight associated with each model.

annual maxima of daily precipitation (results about temperature are available upon request). Statistically, precipitation intensities are skewed and sometimes heavy tailed. So, this type of non-Gaussian random variable represents a challenging test bed for our approach. In addition, multi-error analysis with annual precipitation maxima can reveal key information for impact studies. Details about our selected CMIP models can be found in appendix A.

To illustrate our statistical method described in the previous sections, we focused on two grid points, one near Oxford (U.K.) and Hohenpeissenberg (Germany), see Figs. 1–3. The same type of analysis can be done for each individual grid point. The nine panels of Fig. 4 provide global snapshots of the estimates $\hat{q}_t$ for nine different years: 1850, 1900, 1940, 1970, 2000, 2020, 2030, 2050, and 2100. As already pointed out in Fig. 3, a signal starts to emerge in 2000 and, for some areas, becomes very clear in 2020. In 2050, $\hat{q}_t$ departs from the referential 0.5 in vast regions (red and orange), especially southern and northern, and there the annual precipitation maxima can be attributed to changes in anthropogenic forcing. This statement integrates larger zones in 2100. It is noteworthy that a few patches of green colors indicate that anthropogenic forcing has a reverse impact (i.e., a decrease in precipitation intensities). These spatial precipitation patterns are consistent with the results of previous studies showing that, under continued greenhouse gas emissions, heavy precipitation magnitude is expected to increase over much of the world, except in the subtropics where robust declines are projected (e.g., Pfahl et al. 2017; Tandon et al. 2018; Dong et al. 2021). Hence, the green areas of decreasing extreme precipitation are consistent with other studies (Collins et al. 2013) that reveal significant drying of precipitation of all intensities (mean and extreme). Significant decreases remain confined to ocean regions and barely, if at all, propagate to land areas.

Figures C2 and C3 show that no specific model appears to outperform other models. This complements our understanding at the Oxford grid point in Fig. 1 where four among the 16 models had strong weights. For other grid point locations, other models are chosen. Overall, if all models were equiprobable, then, on average, each model should have a weight of around 1/16 = 0.0625. This value basically corresponds to the fourth column of Table 1, which indicates the global average of the estimated weights for each model. Concerning the first step of our weight procedure during which we only kept models that have an Anderson–Darling p value above 0.2 at a given grid point, the last column of Table 1 confirms that, at the global scale, the average number of grid point rejected by each model is, as expected, around 20%. This points toward the fact that no model appears to be superior (less rejected) at the global scale.

Emergence times definition

As already mentioned, a consequence of our definition q_t is that, when there is no difference between the factual (ALL) and counterfactual (NAT) worlds, it has to be equal to 0.5. This robust reference allows us to define an emergence time in the following way:

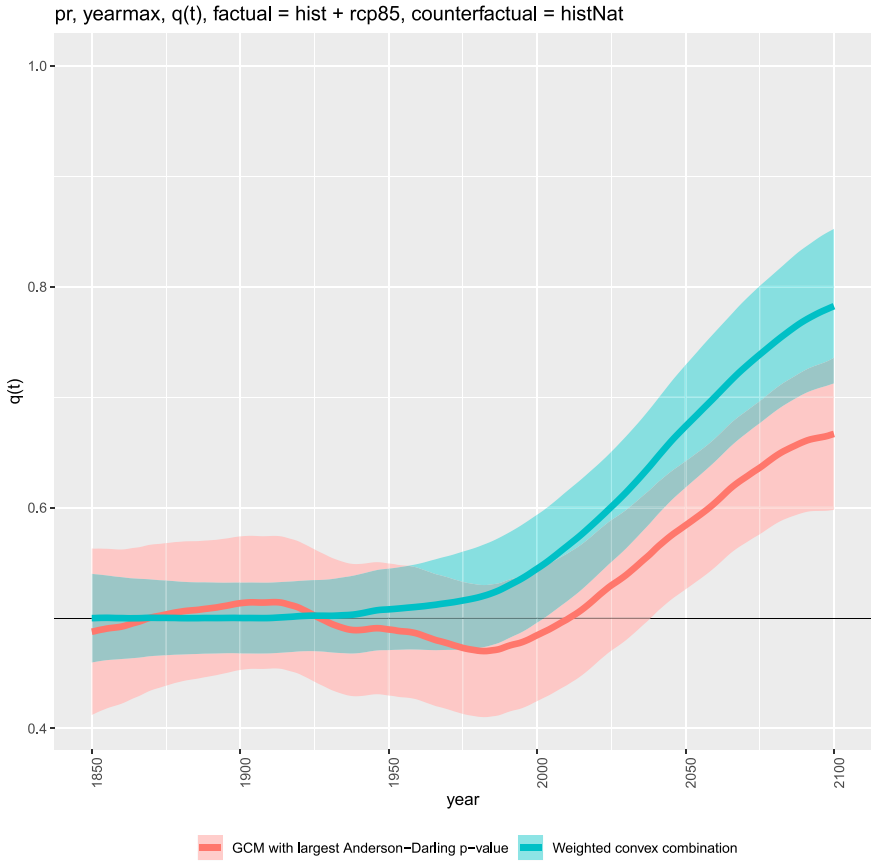

FIG. 3. Oxford grid point: comparison between the best individual model (red), with respect to the Anderson–Darling statistic, and the best convex combination (blue/green) $\hat{q}_t$ defined by (8) that merges the 16 estimates. The probability q_t should be equal to 0.5 (see the horizontal black line) over the preindustrial period when $t \in \mathcal{T} = \{1850, \ldots, 1900\}$.

$$\tau_p = \text{argmin}\{t \text{ for all } t' \geq t, \text{ we have } P(\hat{q}_{t'} > 0.5) > p\}.$$

This means that, at the $1 - p$ significance level, all years after the emergence time τ_p have a q_t value significantly higher than 0.5. That is, an increase in precipitation is detected at year τ_p and this signal remains present after this specific year. From Fig. 5, one can deduce that, in most regions, the anthropogenic signal becomes detectable in model-simulated annual maximum 1-day precipitation around the year 2000 at the level of 90%. In northern latitudes (below 50°N), the detection starts to emerge even as early 1950s. The gray areas do not mean that the signal is not attributable, but, instead of an increase, these regions correspond to a precipitation decrease (see the green areas in the bottom three panels of Fig. 4; e.g., Pfahl et al. 2017; Tandon et al. 2018; Dong et al. 2021).

5. Conclusions and discussion

In the introduction, we claimed that our approach has the key advantage of bypassing multimodel error in EEA analysis. Our treatment of annual maxima of daily rainfall from the CMIP repository provided a clear case study where none of the models were bias corrected. By removing this bias correction step, we avoid the inference of the distribution of extremes rainfall in both factual and counterfactual worlds and the use of bias correction like quantile–quantile mapping for each model (see, e.g., Maraun et al. 2017; Cannon 2018). In addition to providing a simpler setup, our approach also reduces the computational cost.

From a climatological point of view, our analysis clearly indicates changes in precipitation intensities. Emergence times and precipitation patterns are spatially coherent (see Figs. 4 and 5). Overall, our merging of models indicates significant anthropogenic influence in heavy rainfall over most of Earth's surface by 2030 [see Li et al. (2021) and Sun et al. (2022) for similar conclusions but with a parametric approach]. We also found that a convex combination of all models performs better than any single one.

The optimization of model weights is tuned during the preindustrial period. This calibration is assumed to be valid over other time periods and one can wonder if our results are robust with respect to weights changes. To explore this possibility, we also studied two other ways of setting weights. We

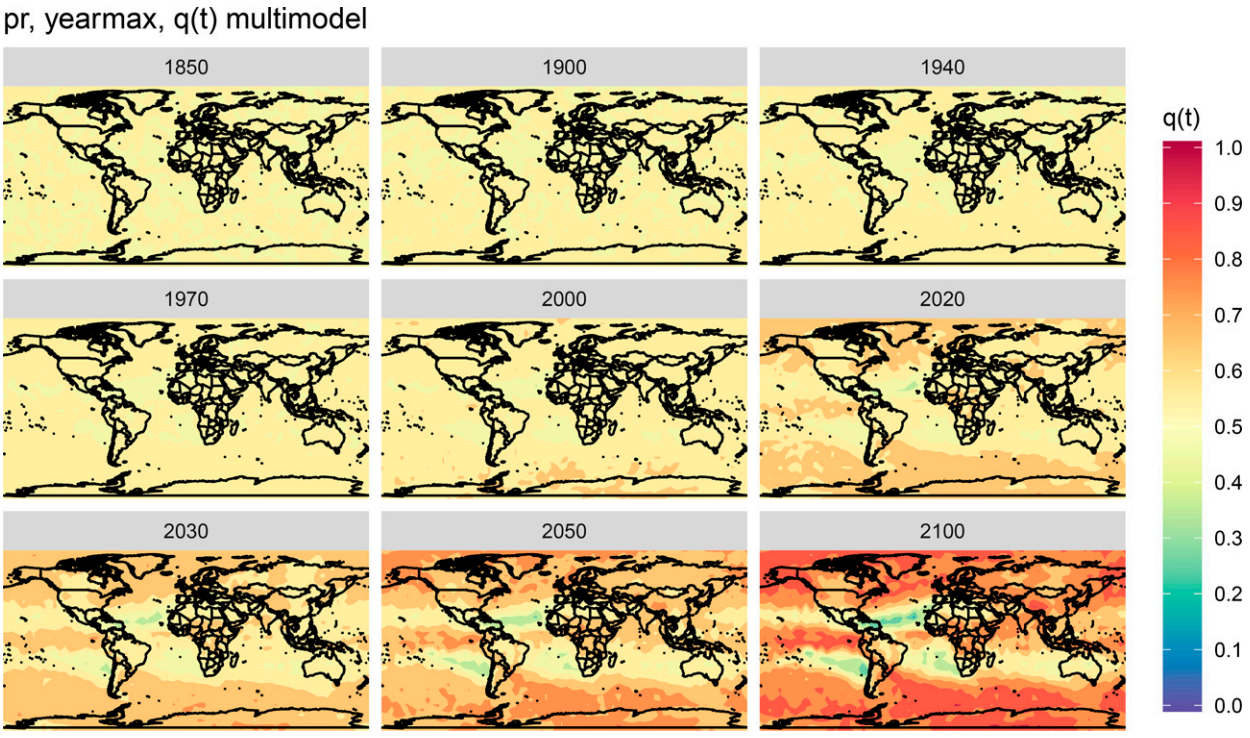

FIG. 4. Values of the multimodel estimates defined by (8) at each grid point for yearly maxima of daily precipitation from CMIP runs in Table 1. A red (green) color indicates a significant increase (decrease) in precipitation intensities due to anthropogenic forcing.

implemented the classical form of weights in variable selection problems (see, e.g., Burnham and Anderson 1998) and used in the climate literature (see, e.g., Lorenz et al. 2018). Another approach is a so-called expert aggregation approach (e.g., Gaillard and Goude 2015). Overall, all three approaches gave similar emergence times, with well-structured spatial patterns with the expert aggregation technique and our approach. For these two methods, estimates of $\hat{q}_t$ are robust to weight changes during the preindustrial period. For other time periods, the absence of counterfactual perfect observations makes it impossible to remove the assumption of stationary weights. Other research avenues could be explored to integrate observational data in our statistical approach (see, e.g., Sabourin et al. 2013). In this work, our analysis only relies on numerical simulation outputs from CMIP, and not on observations. Consequently, our conclusions have to be solely interpreted within numerical worlds [see Otto et al. (2020) for a discussion about the confidence assignment of EEA studies].

Another delicate issue, especially for precipitation, is to define the spatial scale of interest. In most CMIP-based D&A

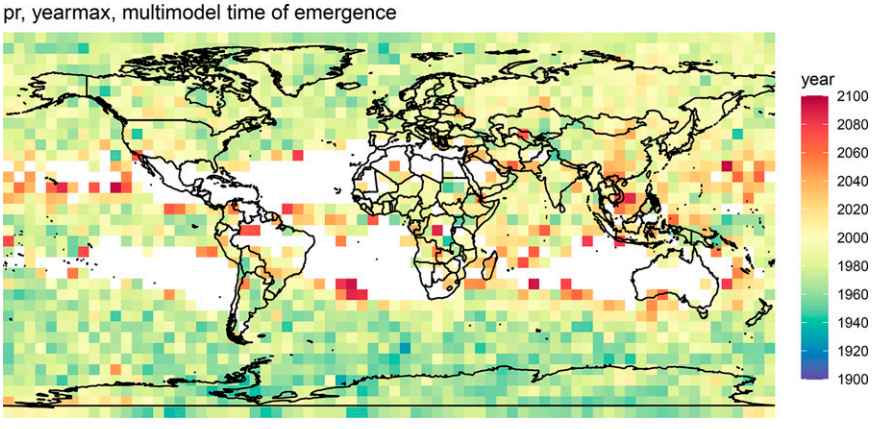

FIG. 5. Emergence times: years after which all q_t values are significantly higher the 0.5 at the significance level of 10%. The gray regions correspond to a detected decrease in precipitation; see green patches in Fig. 4.

studies, such as in Ribes et al. (2021), the analysis is done at global or regional scales. This facilitates the integration of observations and improve the signal-to-noise ratio (see, e.g., Hannart et al. 2014). But, gridded spatial features like the ones in Fig. 4 are lost. This also leads to the open question of how to find optimal regions that maximize the attribution power in a multivariate context (see, e.g., Le Gall et al. 2021; Kiriliouk and Naveau 2020; Yiou et al. 2017; Vannitsem and Naveau 2007). A related question is how to adapt our approach to a multivariate framework to attribute compounds events (see, e.g., Zscheischler et al. 2019). Coupling the field of counterfactual theory and multivariate analysis could help in the direction (see, e.g., Hannart et al. 2016). Our attribution approach can also be extended to the analysis of record events (e.g., to assess probabilities that the current realization is the largest ever recorded). In our approach, the level u was replaced by a random variable; see how we went from (2) to (3). As in Naveau et al. (2018) and Worms and Naveau (2020), it would be possible to replace u in (2) by the random variable $\max(X_1, \ldots, X_t)$. In this case, the set $\{X_t > \max(X_1, \ldots, X_{t-1})\}$ corresponds to the event that the record occurs at time t. Hence, it would be possible to use the same multimodel combination technique to analyze records. While conceptually possible, this extension needs future work in terms of implementation.

A last point is our use of the CMIP worst-case scenarios (RCP8.5 and SSP5–8.5). Our proposed technique can be easily applied to other scenarios. Still, most inferred emergence times of yearly maxima of precipitation span the period 2000–20 (see the blue to yellow regions in Fig. 5). This period is prior to any strong differences among scenarios. Hence, our estimated emergence times for these regions will remain valid with other scenarios.

Acknowledgments. Part of this work was supported by the DAMOCLES-COST-ACTION on compound events, the French national program (FRAISE-LEFE/INSU and 80 PRIME CNRS-INSU), and the European H2020 XAIDA (Grant agreement ID: 101003469). The authors also acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-20-CE40-0025-01 (T-REX project) and reference ANR-19-CE46-0011 (MELODY project). We also thank Dr. Tuel for fruitful discussions on global precipitation patterns and other related topics.

Data availability statement. The data used in this paper come from the CMIP repository and the guideline to use them can be found at https://pcmdi.llnl.gov/CMIP6/Guide/dataUsers.html. The interested readers can also contact the corresponding author.

APPENDIX A

Datasets

From the CMIP5 archive, we select 12 models for which we found a complete set of precipitation simulations for the historical (1850–2005), historicalNat (1850–2012), and RCP8.5 (2006–2100) experiments. Additionally, we also treat four CMIP6 models (see, e.g., Eyring et al. 2016) for which we found historical simulations in the CMIP6 deck and hist-nat simulations in DAMIP and SSP585 projection. The historical simulations combined with the RCP8.5 and SSP585 simulations represent the factual world whereas the historicalNat simulations correspond to the counterfactual world. All runs have been remapped to a common 5° × 5° HadCRUT grid (cdo rmapcon operator). Table 1 provides the list of the 16 CMIP models used in this study.

APPENDIX B

Methods

Proof of Eq. (4): To simplify the proofs, we have dropped the temporal indexing t whenever it was possible. We have, by definition of Z_t and X_t,

$$
\begin{aligned}
q_t &= P(Z_t > X_t),\\
&= P\left[F_t^{\leftarrow} \circ F_t^{(m)}\left(Z_t^{(m)}\right) > G_t^{\leftarrow} \circ G_t^{(m)}\left(X_t^{(m)}\right)\right],\\
&= P\left[G_t^{\leftarrow} \circ G_t^{(m)}\left(Z_t^{(m)}\right) > G_t^{\leftarrow} \circ G_t^{(m)}\left(X_t^{(m)}\right)\right], \text{ from assumption A,}\\
&= P\left[G_t^{(m)}\left(Z_t^{(m)}\right) > G_t^{(m)}\left(X_t^{(m)}\right)\right], \text{ as } G_t(\cdot) \text{ non-decreasing,}\\
&= P\left(Z_t^{(m)} > X_t^{(m)}\right), \text{ as } \left(G_t^{(m)}\right)^{\leftarrow}(\cdot) \text{ non-decreasing.}
\end{aligned}
$$

Computation of confidence intervals of $\hat{q}_t^{(m)}$

For each climate model m, the year-varying probability $q_t^{(m)} = P(Z_t^{(m)} > X_t^{(m)}) = \mathbb{E}[G_t^{(m)}(Z_t^{(m)})]$ can be inferred from the Nadaraya–Watson estimate $\hat{q}_t^{(m)}$ defined by Eq. (5). To simplify the inference process, we consider the CDF G_t of the NAT as a stationary process.[B1] In this case, G_t can be estimated by the classical empirical estimator $\mathbb{G}_I^{(m)}$ defined as

$$\mathbb{G}_I^{(m)}(x) = \frac{1}{I}\sum_{I=1}^{I} \mathbb{I}\left(X_{t_i}^{(m)} \le x\right),$$

where $\mathbb{I}(A)$ represents the indicator function equal to one if A is true and zero otherwise and $(X_{t_1},\ldots,X_{t_I})^{\mathrm{T}}$ corresponds to a NAT run trajectory of I time steps. To derive confidence intervals around this estimate, we need to introduce

$$\tilde{q}_t^{(m)} = \frac{1}{\sum K_h(t-t_j)} \sum_{j=1}^{J} K_h(t-t_j)\mathbb{G}_I^{(m)}\left(Z_{t_j}^{(m)}\right),$$

which corresponds to a simpler version of $\hat{q}_t^{(m)}$ in which the CDF $G^{(m)}$ is supposed to be known. In this case, the asymptotic behavior of the difference $A = \tilde{q}_t^{(m)} - q_t^{(m)}$ is well known (see, e.g., Härdle 1991). It converges toward a Gaussian limit law with known asymptotic mean and variances,

[B1] This assumption could easily be removed by adding a kernel estimator for $G_t^{(m)}$; however, this was not necessary for our precipitation application.

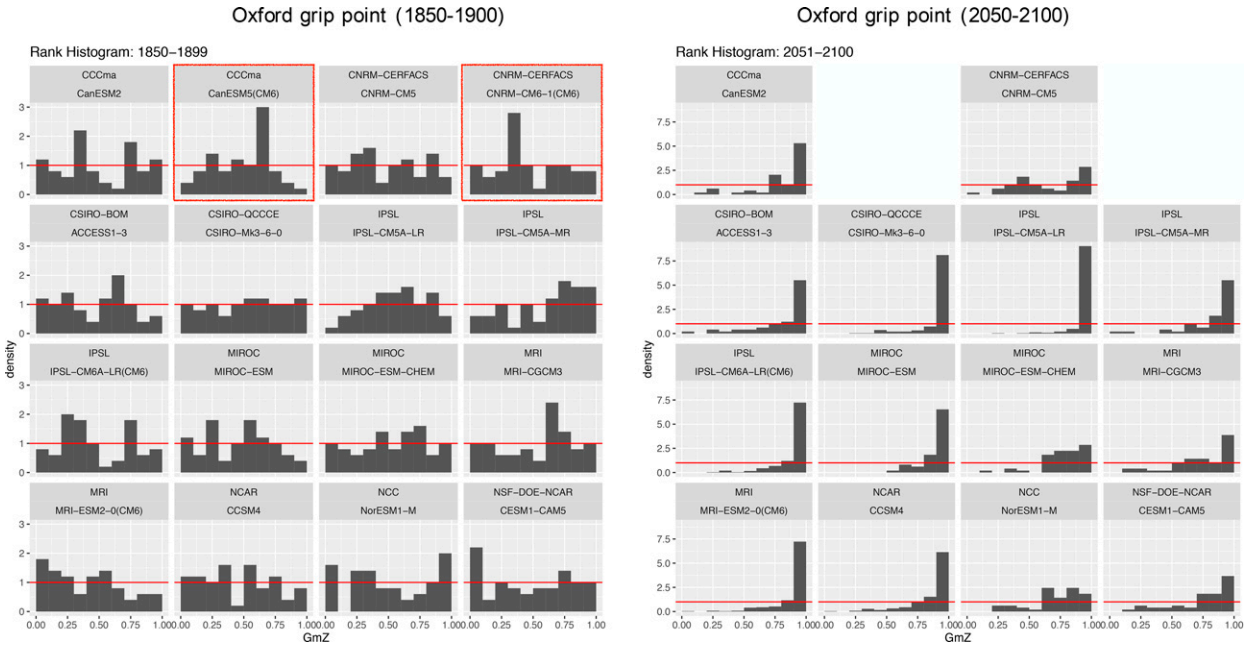

FIG. C1. Each graph represents the probability integral transform (PIT) of different CMIP model for the Oxford grid point (the two red boxes indicate the two models that are rejected by the Anderson–Darling test). The PIT is computed between the ranked yearly maxima of daily precipitation for the ALL run and the ones for the NAT run. The difference between the right and left panels is the period considered, 1850–1900 on the left and 2050–100 on the right.

$$\left\{(Jh)^{1/2}\frac{\tilde{q}(t_j)-q(t_j)}{\left[\sigma^2(t_j)\|K\|_2^2/f(t_j)\right]^{1/2}}\right\}_{j=1}^{J}$$

converges into toward

$$N\left(\left\{[q''(t_j)+2q'(t_j)f'(t_j)/f(t_j)]\int s^2K(s)ds\right\}_{j=1}^{J},\ I\right)$$

where $\sigma^2(t)$ is the conditional variance of $\{G_t(Z_t\}$ and f is the density of the temporal variable T. Hence, we can write

$$\hat{q}_t^{(m)}-q_t^{(m)}=A+B,\quad \text{with}$$

$$B=\frac{1}{\sum K_h(t-t_j)}\sum_{j=1}^{J}K_h(t-t_j)\left[\mathbb{G}_I^{(m)}\left(Z_{t_j}^{(m)}\right)-G^{(m)}\left(Z_{t_j}^{(m)}\right)\right].$$

Following Naveau et al. (2018), we consider that $\sqrt{I}[\tilde{G}^{(m)}(Z_{t_j}^{(m)})-G^{(m)}(Z_{t_j}^{(m)})]$ behaves (almost) like a Brownian bridge:

$$B=\frac{1}{\sqrt{I}\sum K_h(t-t_j)}\sum_{j=1}^{J}K_h(t-t_j)\sqrt{I}\left[\mathbb{E}G_I^{(m)}\left(Z_{t_j}^{(m)}\right)-G^{(m)}\left(Z_{t_j}^{(m)}\right)\right]$$

$$\approx\frac{1}{\sqrt{I}\sum K_h(t-t_j)}\sum_{j=1}^{J}K_h(t-t_j)\sqrt{I}B\left[G^{(m)}\left(Z_{t_j}^{(m)}\right)\right]$$

where $B(u)$ represents a classical Brownian bridge on [0, 1] with $\mathbb{E}[B(u)]=0$ and covariance: $\mathrm{Cov}[B(u),B(v)]=\min(u,v)-uv$. Thus, $\mathbb{E}[B]=\mathbb{E}[\hat{q}_t^{(m)}-\tilde{q}_t^{(m)}]\approx 0$ and

$$\mathrm{Var}[B]\approx\frac{1}{I\left[\sum K_h(t-t_j)\right]^2}\sum_{j=1}^{J}\sum_{i=1}^{J}K_h(t-t_j)K_h(t-t_i)\times$$

$$\mathbb{E}\left\{\min\left[G^{(m)}\left(Z_{t_i}^{(m)}\right),G^{(m)}\left(Z_{t_j}^{(m)}\right)\right]-G^{(m)}\left(Z_{t_i}^{(m)}\right)G^{(m)}\left(Z_{t_j}^{(m)}\right)\right\}.$$

Assuming that the terms A and B are independent and that their mean is negligible, we build confidence intervals for $q_t^{(m)}$ assuming that asymptotically

$$\hat{q}_t^{(m)}\sim\mathcal{N}\left[q_t^{(m)},\mathrm{Var}(A)+\mathrm{Var}(B)\right].$$

To go one step further, we need to take into account that $q_t^{(m)}$ may not be centered around the real quantity of interest q_t. It is centered if assumption A is satisfied; otherwise, a bias exists between $q_t^{(m)}$ and q_t. To deal with this issue, we additionally assume that such a bias is randomly distributed between all our climate models in the following way:

$$q_t^{(m)}\sim\mathcal{N}\left(q_t,\sigma_t^2\right),$$

$$\hat{q}_t^{(m)}\left|q_t^{(m)}\right.\sim\mathcal{N}\left[q_t^{(m)},\left(\sigma_t^{(m)}\right)^2\right],$$

where $q_t^{(m)}$ and $(\sigma_t^{(m)})^2=\mathrm{Var}(A)+\mathrm{Var}(B)$ from the previous paragraph. The variance σ_t^2 captures the intermodel variability between the different climate runs. By marginalizing over $q_t^{(m)}$, we then obtain by Bayesian conjugation for normal laws that

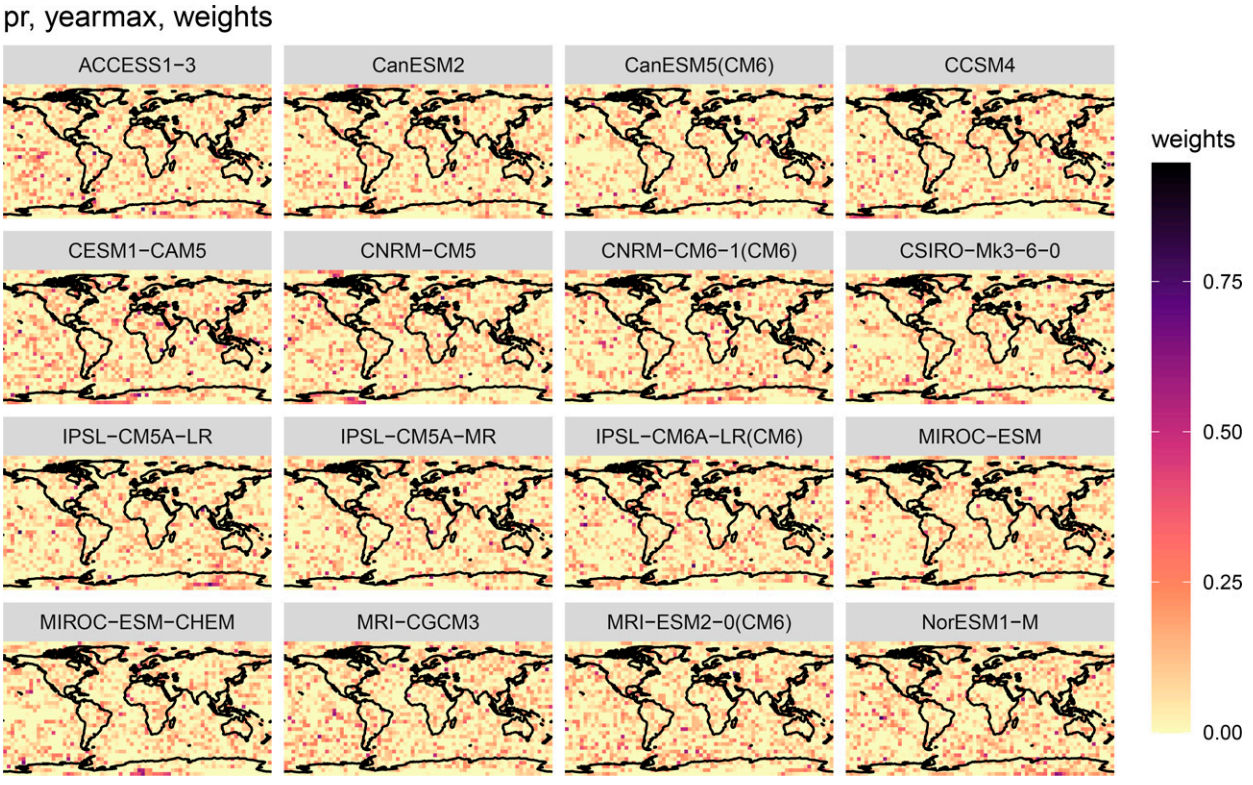

FIG. C2. Weights of each model in Eq. (8) to obtain Fig. 4.

$$\hat{q}_t^{(m)} \sim \mathcal{N}\left[q_t, \sigma_t^2 + \left(\sigma_t^{(m)}\right)^2\right]. \tag{B1}$$

This variance decomposition simply means that the variance of $\hat{q}_t^{(m)}$ can be divided into the intramodel variance $(\sigma_t^{(m)})^2$ and the intermodel variance σ_t^2 which can be estimated as

$$\hat{\sigma}_t^2 = \frac{1}{M}\sum_{m=1}^{M}\left(\hat{q}_t^{(m)} - \hat{q}_t\right)^2,$$

where

$$\hat{q}_t = \sum_{m=1}^{M} w_m \times \hat{q}_t^{(m)}.$$

The choice of the weights w_m depends on the climate model capability to satisfy during preindustrial period. Given the weights, the weighted estimator of q_t follows

$$\hat{q}_t \sim \mathcal{N}\left\{q_t, \sum_{m=1}^{M} w_m^2\left[\sigma_t^2 + \left(\sigma_t^{(m)}\right)^2\right]\right\}. \tag{B2}$$

Equations (B1) and (B2) were used to obtain the confidence intervals displayed in our figures. The kernel used in $\hat{q}_t^{(m)}$ [see Eq. (5)] is the classical Epanechnikov kernel (Epanechnikov 1969) with a bandwidth of 60.5 years. The bandwidth has been determined on using a leave-one-out cross-validation scheme to find out the bandwidth that minimizes the root-mean-square error (RMSE) between the estimated $\hat{q}_1(t_i)$ and the $\tilde{G}_{\mathbb{I}}(Z_{t_i})$. More precisely, the cross-validation has been performed for each model individually and then we select the median of bandwidths optimized for each model.

Note that the derivation of the confidence bands relies on the assumption of independence between the estimates of q_t. Although various studies (see, e.g., Knutti et al. 2009; Haughton et al. 2015) pointed out that climate models are not necessarily independent, this issue may not be too prevalent in our case for the following reasons. We do not require independence between raw atmospheric variables, but between events like $\{Z_t^{(i)} > X_t^{(i)}\}$ and $\{Z_t^{(j)} > X_t^{(j)}\}$ for climate models i and j. These events are based on increments like $Z_t^{(i)} - X_t^{(i)}$ and $Z_t^{(j)} - X_t^{(j)}$ and consequently, by removing additive error, increments are more likely independent than raw data. Concerning the latter, our main focus is gridded annual maxima of precipitation. Such variables have high variability and intermodel dependence of increments is secondary compared to the signal-to-noise ratio issue. In addition, Fig. 2 indicates that the differences within a research center (e.g., IPSL) appear to be as important than the ones between different research laboratories.

APPENDIX C

Appendix Figure

In Fig. C1, the probability integral transform of the preindustrial period (left) for the Oxford grid point is compared to the one (at right) obtained in the future (2050-2100).

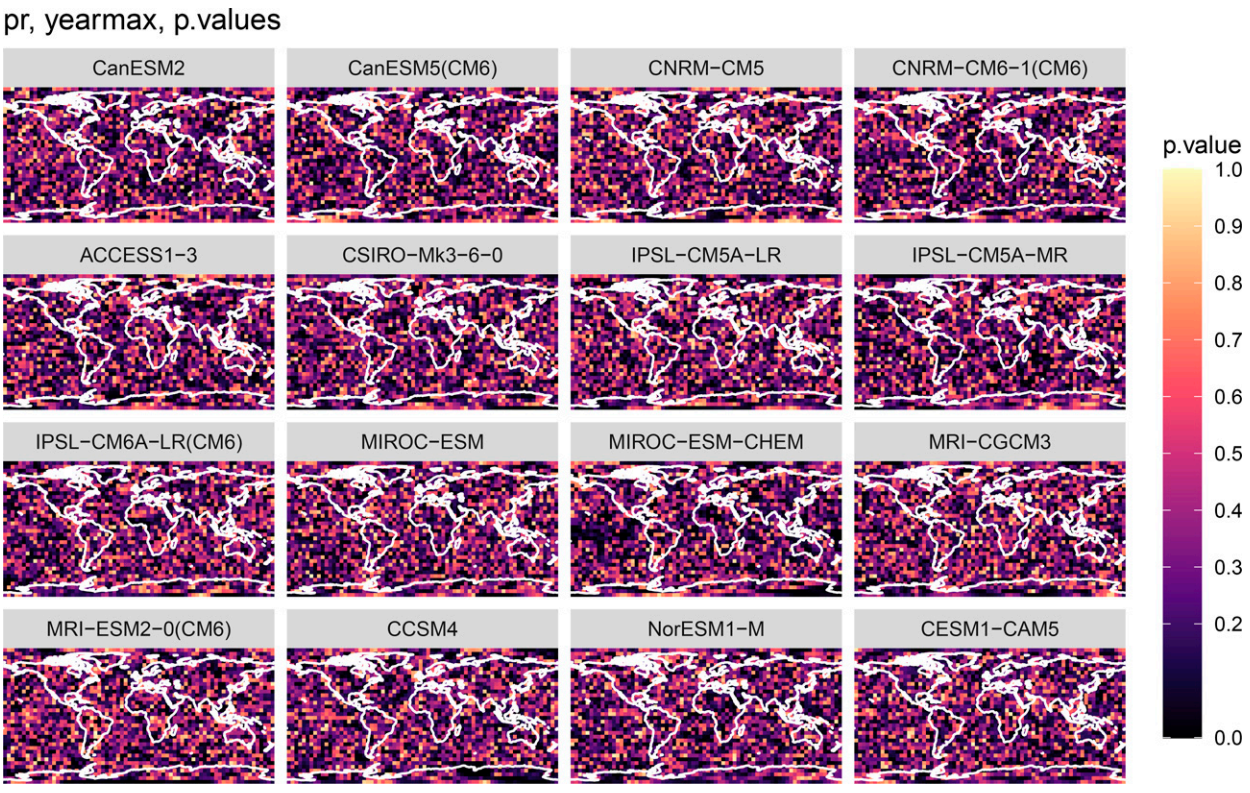

FIG. C3. The p values of the two-sample Anderson–Darling test (Pettitt 1976; Anderson and Darling 1952).

The weights used in the aggregated estimator of q_t defined by Eq. (8) are displayed in Fig. C2. The x axis corresponds to the labels of the 16 CMIP models used in the aggregation (see Table 1). The y axis represents grid points locations of each model. The black (yellow) color corresponds a weight near one (zero) in Eq. (8).

Figure C3 displays the p-value maps of each of the 12 CMIP models under study.

APPENDIX D

SI Kullback–Leibler Divergence Computation

The log-likelihood of independent[D1] Bernoulli sequences, $B_t^{(m)} = \{Z_t^{(m)} > X_t^{(m)}\}$, over $t \in \mathcal{T}$ can be written as

$$\sum_{t\in\mathcal{T}} \log\left[\left(1 - q_t^{(m)}\right)^{1-B_t^{(m)}} \left(q_t^{(m)}\right)^{B_t^{(m)}}\right].$$

For two binomial distributions with respective success rates p and q, the Kullback–Leibler divergence is equal to

$$D(q; \mathrm{p}) = q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}.$$

This leads to the following Kullback–Leibler divergence that compares the T-dimensional vector $q_{t\in\mathcal{T}}^{(m)}$ with $q_{t\in\mathcal{T}} = c(0.5, \ldots, 0.5)$:

$$D\left[c(0.5, \ldots, 0.5); q_{t\in\mathcal{T}}^{(m)}\right] = -T \times \log 2 - \frac{1}{2}\sum_{t\in\mathcal{T}} \log\left[q_t^{(m)}\left(1 - q_t^{(m)}\right)\right].$$

For each model m, we can measure the departure from the term $T \times \log 2$ with $T = 50$ years. A Kullback–Leibler divergence is always nonnegative and equals to zero if $q_t^{(m)} = c(0.5, \ldots, 0.5)$. The estimate of $q_t^{(m)}$ from Eq. (5) can be plugged in the expression $D(\cdot; \cdot)$, and consequently each model m can be evaluated with respect to condition Eq. (7). The optimization to find the weights in Eq. (8) is numerically done using the solnp function in the R package Rsolnp.

In the model worlds, we also have that

$$P\left(Z_t^{(m)} > X_t^{(m)}\right) = \frac{1}{2}$$

for all years t during which the two variables $Z_t^{(m)}$ and $X_t^{(m)}$ are exchangeable. The exchangeability assumption simply means that the labeling of the numerical model type (X and Z) is uninformative; that is, $P(Z_t^{(m)} > X_t^{(m)}) = P(X_t^{(m)} > Z_t^{(m)})$.

[D1] As we analyze yearly maxima of daily values in our application, the hypothesis of year-to-year independence is reasonable. If that is not the case, then the full times series $B_t^{(m)}$ with $t \in \mathcal{T}$ will have to be modeled as a multivariate binary random vector with a memory component (see, e.g., Tuel et al. 2017; Dai et al. 2013). This will lead to a more complex likelihood. Albeit with added complexity, the principles based on the Kullback–Leibler divergence exposed in this section will remain valid.

This implies that, although climate runs from the same laboratory may be dependent and share the same code (see, e.g., Knutti et al. 2019), they are likely to be exchangeable (labeling free) during time periods with similar forcing. Another aspect is that the bivariate vector $(Z_t^{(m)}, X_t^{(m)})$ does not have to be stationary in time to have $P(Z_t^{(m)} > X_t^{(m)}) = 1/2$. In particular, during the early period where the anthropogenic forcing was weak (preindustrial period), natural forcings were still a source of nonstationarity in our climate system, but $q_t = 0.5$ for t within this preindustrial period.

Interpreting weights in the Kullback–Leibler divergence

One can notice that, if $q_t^{(m)}$ is stationary over $\mathcal{T}$ for model m, then the divergence $D[q_{t\in\mathcal{T}}; q_{t\in\mathcal{T}}^{(m)}]$ can be expressed as a function of a variance ratio

$$D\left(q_{t\in\mathcal{T}}; q_{t\in\mathcal{T}}^{(m)}\right) = -\frac{T}{2} \times \log\left(\frac{\mathbb{V}B_t^{(m)}}{\mathbb{V}B_t}\right),$$

where $\mathbb{V}B_t^{(m)} = q_t^{(m)}(1 - q_t^{(m)})$ is always smaller that $\mathbb{V}B_t = 1/4$. Equivalently, we can write the variance ratio as a function of the divergence $\mathbb{V}B_t^{(m)}/\mathbb{V}B_t = \exp[-2 \times D(q_{t\in\mathcal{T}}; q_{t\in\mathcal{T}}^{(m)})/T] \leq 1$. If this variance ratio is close to one (zero), then $q_t^{(m)}$ is close (far) from $q_t = 0.5$.

REFERENCES

Abramowitz, G., and Coauthors, 2019: ESD reviews: Model dependence in multi-model climate ensembles: Weighting, sub-selection and out-of-sample testing. *Earth Syst. Dyn.*, **10**, 91–105, https://doi.org/10.5194/esd-10-91-2019.

Anderson, T. W., and D. A. Darling, 1952: Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Stat.*, **23**, 193–212, https://doi.org/10.1214/aoms/1177729437.

Angélil, O., D. Stone, and M. Wehner, 2017: An independent assessment of anthropogenic attribution statements for recent extreme temperature and rainfall events. *J. Climate*, **30**, 5–16, https://doi.org/10.1175/JCLI-D-16-0077.1.

Bellprat, O., V. Guemas, F. Doblas-Reyes, and M. G. Donat, 2019: Towards reliable extreme weather and climate event attribution. *Nat. Commun.*, **10**, 1732, https://doi.org/10.1038/s41467-019-09729-2.

Burnham, K. P., and D. R. Anderson, 1998: *Model Selection and Inference: A Practical Information-Theoretical Approach.* Springer, 353 pp.

Cannon, A. J., 2018: Multivariate quantile mapping bias correction: An *N*-dimensional probability density function transform for climate model simulations of multiple variables. *Climate Dyn.*, **50**, 31–49, https://doi.org/10.1007/s00382-017-3580-6.

Chen, Y., W. Moufouma-Okia, V. Masson-Delmotte, P. Zhai, and A. Pirani, 2018: Recent progress and emerging topics on weather and climate extremes since the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. *Ann. Rev. Environ. Resour.*, **43**, 35–59, https://doi.org/10.1146/annurev-environ-102017-030052.

Coles, S. G., 2001: *An Introduction to Statistical Modeling of Extreme Values.* Springer, 208 pp.

Collins, M., and Coauthors, 2013: Long-term climate change: Projections, commitments and irreversibility. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 1029–1136.

Dai, B., S. Ding, and G. Wahba, 2013: Multivariate Bernoulli distribution. *Bernoulli*, **19**, 1465–1483, https://doi.org/10.3150/12-BEJSP10.

Dong, S., Y. Sun, C. Li, X. Zhang, S.-K. Min, and Y.-H. Kim, 2021: Attribution of extreme precipitation with updated observations and CMIP6 simulations. *J. Climate*, **34**, 871–881, https://doi.org/10.1175/JCLI-D-19-1017.1.

Epanechnikov, V., 1969: Non-parametric estimation of a multivariate probability density. *Theory Probab. Appl.*, **14**, 153–158, https://doi.org/10.1137/1114019.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, **9**, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016.

Gaillard, P., and Y. Goude, 2015: Forecasting electricity consumption by aggregating experts; how to design a good set of experts. *Modeling and Stochastic Learning for Forecasting in High Dimensions*, A. Antoniadis, X. Brossat, and J.-M. Poggi, Eds., Lecture Notes in Statistics, Vol. 217, Springer, 95–115.

Hannart, A., A. Ribes, and P. Naveau, 2014: Optimal fingerprinting under multiple sources of uncertainty. *Geophys. Res. Lett.*, **41**, 1261–1268, https://doi.org/10.1002/2013GL058653.

——, J. Pearl, F. E. L. Otto, P. Naveau, and M. Ghil, 2016: Counterfactual causality theory for the attribution of weather and climate-related events. *Bull. Amer. Meteor. Soc.*, **97**, 99–110, https://doi.org/10.1175/BAMS-D-14-00034.1.

Härdle, W., 1991: *Smoothing Techniques: With Implementation in S.* Springer, 261 pp.

Haughton, N., G. Abramowitz, A. Pitman, and S. J. Phipps, 2015: Weighting climate model ensembles for mean and variance estimates. *Climate Dyn.*, **45**, 3169–3181, https://doi.org/10.1007/s00382-015-2531-3.

Hegerl, G., and F. W. Zwiers, 2011: Use of models in detection and attribution of climate change. *Wiley Interdiscip. Rev.: Climate Change*, **2**, 570–591, https://doi.org/10.1002/wcc.121.

Kallache, M., M. Vrac, P. Naveau, and P. A. Michelangeli, 2011: Nonstationary probabilistic downscaling of extreme precipitation. *J. Geophys. Res.*, **116**, D05113, https://doi.org/10.1029/2010JD014892.

Kharin, V. V., and F. W. Zwiers, 2000: Changes in the extremes in an ensemble of transient climate simulations with a coupled atmosphere–ocean GCM. *J. Climate*, **13**, 3760–3788, https://doi.org/10.1175/1520-0442(2000)013<3760:CITEIA>2.0.CO;2.

Kiriliouk, A., and P. Naveau, 2020: Climate extreme event attribution using multivariate peaks-over-thresholds modeling and counterfactual theory. *Ann. Appl. Stat.*, **14**, 1342–1358, https://doi.org/10.1214/20-AOAS1355.

Knutson, T., J. K. Kossin, C. Mears, J. Perlwitz, and M. Wehner, 2017: Detection and attribution of climate change. *Climate Science Special Report: Fourth National Climate Assessment*, D. J. Wuebbles et al., Eds., Vol. I, U.S. Global Change Research Program, 114–132.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl, 2009: Challenges in combining projections from multiple

climate models. *J. Climate*, **23**, 2739–2758, https://doi.org/10.1175/2009JCLI3361.1.

——, C. Baumberger, and G. H. Hadorn, 2019: Uncertainty quantification using multiple models—Prospects and challenges. *Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives*, C. Beisbart and N. J. Saam, Eds., Springer, 835–855.

Le Gall, P., A.-C. Favre, P. Naveau, and A. Tuel, 2021: Non-parametric multimodel regional frequency analysis applied to climate change detection and attribution. https://doi.org/10.48550/arXiv.2111.00798.

Li, C., F. Zwiers, X. Zhang, G. Li, Y. Sun, and M. Wehner, 2021: Changes in annual extremes of daily temperature and precipitation in CMIP6 models. *J. Climate*, **34**, 3441–3460, https://doi.org/10.1175/JCLI-D-19-1013.1.

Lorenz, R., N. Herger, J. Sedláček, V. Eyring, E. M. Fischer, and R. Knutti, 2018: Prospects and caveats of weighting climate models for summer maximum temperature projections over North America. *J. Geophys. Res. Atmos.*, **123**, 4509–4526, https://doi.org/10.1029/2017JD027992.

Maraun, D., and Coauthors, 2017: Towards process-informed bias correction of climate change simulations. *Nat. Climate Change*, **7**, 764–773, https://doi.org/10.1038/nclimate3418.

Nadaraya, E. A., 1964: On estimating regression. *Theory Probab. Appl.*, **9**, 141–142, https://doi.org/10.1137/1109020.

National Academies of Sciences, Engineering and Medicine, 2016: *Attribution of Extreme Weather Events in the Context of Climate Change.* National Academies Press, 186 pp.

Naveau, P., A. Guillou, and T. Rietsch, 2014: A non-parametric entropy-based approach to detect changes in climate extremes. *J. Roy. Stat. Soc.*, **B76**, 861–884, https://doi.org/10.1111/rssb.12058.

——, A. Ribes, F. W. Zwiers, A. Hannart, A. Tuel, and P. Yiou, 2018: Revising return periods for record events in a climate event attribution context. *J. Climate*, **31**, 3411–3422, https://doi.org/10.1175/JCLI-D-16-0752.1.

——, A. Hannart, and A. Ribes, 2020: Statistical methods for extreme event attribution in climate science. *Ann. Rev. Stat. Appl.*, **7**, 89–110, https://doi.org/10.1146/annurev-statistics-031219-041314.

Otto, F. E. L., and Coauthors, 2020: Toward an inventory of the impacts of human-induced climate change. *Bull. Amer. Meteor. Soc.*, **101**, E1972–E1979, https://doi.org/10.1175/BAMS-D-20-0027.1.

Pettitt, A. N., 1976: A two-sample Anderson–Darling rank statistic. *Biometrika*, **63**, 161–168, https://doi.org/10.2307/2335097.

Pfahl, S., P. A. O'Gorman, and E. M. Fischer, 2017: Understanding the regional pattern of projected future changes in extreme precipitation. *Nat. Climate Change*, **7**, 423–427, https://doi.org/10.1038/nclimate3287.

Ribes, A., S. Qasmi, and N. P. Gillett, 2021: Making climate projections conditional on historical observations. *Sci. Adv.*, **7**, eabc0671, https://doi.org/10.1126/sciadv.abc0671.

Robin, Y., P. Yiou, and P. Naveau, 2017: Detecting changes in forced climate attractors with Wasserstein distance. *Nonlinear Processes Geophys.*, **24**, 393–405, https://doi.org/10.5194/npg-2017-5.

——, M. Vrac, P. Naveau, and P. Yiou, 2019: Multivariate stochastic bias corrections with optimal transport. *Hydrol. Earth Syst. Sci.*, **23**, 773–786, https://doi.org/10.5194/hess-23-773-2019.

Sabourin, A., P. Naveau, and A. L. Fougères, 2013: Bayesian model averaging for multivariate extremes. *Extremes*, **16**, 325–350, https://doi.org/10.1007/s10687-012-0163-0.

Shepherd, T. G., 2016: A common framework for approaches to extreme event attribution. *Curr. Climate Change Rep.*, **2**, 28–38, https://doi.org/10.1007/s40641-016-0033-y.

Stott, P. A., and Coauthors, 2016: Attribution of extreme weather and climate-related events. *Wiley Interdiscip. Rev.: Climate Change*, **7**, 23–41, https://doi.org/10.1002/wcc.380.

Sun, Q., F. Zwiers, X. Zhang, and J. Yan, 2022: Quantifying the human influence on the intensity of extreme 1- and 5-day precipitation amounts at global, continental, and regional scales. *J. Climate*, **35**, 195–210, https://doi.org/10.1175/JCLI-D-21-0028.1.

Tandon, N. F., X. Zhang, and A. H. Sobel, 2018: Understanding the dynamics of future changes in extreme precipitation intensity. *Geophys. Res. Lett.*, **45**, 2870–2878, https://doi.org/10.1002/2017GL076361.

Tuel, A., P. Naveau, and C. Ammann, 2017: Skillful prediction of multidecadal variations in volcanic forcing. *Geophys. Res. Lett.*, **44**, 2868–2874, https://doi.org/10.1002/2016GL072234.

Vannitsem, S., and P. Naveau, 2007: Spatial dependences among precipitation maxima over Belgium. *Nonlinear Processes Geophys.*, **14**, 621–630, https://doi.org/10.5194/npg-14-621-2007.

van Oldenborgh, G. J., and Coauthors, 2021: Pathways and pitfalls in extreme event attribution. *Climatic Change*, **166**, 13, https://doi.org/10.1007/s10584-021-03071-7.

Watson, G. S., 1964: Smooth regression analysis. *Sankhyā*, **26**, 359–372.

Worms, J., and P. Naveau, 2020: Record events attribution in climate studies. https://hal.archives-ouvertes.fr/hal-02938596.

Yiou, P., A. Jezequel, P. Naveau, F. E. L. Otto, R. Vautard, and M. Vrac, 2017: A statistical framework for conditional extreme event attribution. *Adv. Stat. Climatol. Meteor. Oceanogr.*, **3**, 17–31, https://doi.org/10.5194/ascmo-3-17-2017.

Zscheischler, J., E. M. Fischer, and S. Lange, 2019: The effect of univariate bias adjustment on multivariate hazard estimates. *Earth Syst. Dyn.*, **10**, 31–43, https://doi.org/10.5194/esd-10-31-2019.