Earth Syst. Dynam., 15, 735–762, 2024
https://doi.org/10.5194/esd-15-735-2024

Earth System Dynamics
Open Access
EGU

Distribution-based pooling for combination and multi-model bias correction of climate simulations

Mathieu Vrac[1], Denis Allard[2], Grégoire Mariéthoz[3], Soulivanh Thao[1], and Lucas Schmutz[3]

[1]Laboratoire des Sciences du Climat et de l'Environnement (LSCE-IPSL), CEA/CNRS/UVSQ, Université Paris-Saclay, Centre d'Etudes de Saclay, Orme des Merisiers, 91191 Gif-sur-Yvette, France
[2]Biostatistics and Spatial Processes (BioSP), INRAE, 84914 Avignon, France
[3]University of Lausanne, Expertise Center for Climate Extremes (ECCE), Institute of Earth Surface Dynamics (IDYST), UNIL-Mouline, Geopolis, 1015 Lausanne, Switzerland

Correspondence: Mathieu Vrac (mathieu.vrac@lsce.ipsl.fr) and Grégoire Mariéthoz (gregoire.mariethoz@unil.ch)

Received: 13 December 2023 – Discussion started: 9 January 2024
Revised: 8 April 2024 – Accepted: 16 April 2024 – Published: 13 June 2024

Abstract. For investigating, assessing, and anticipating climate change, tens of global climate models (GCMs) have been designed, each modelling the Earth system slightly differently. To extract a robust signal from the diverse simulations and outputs, models are typically gathered into multi-model ensembles (MMEs). Those are then summarized in various ways, including (possibly weighted) multi-model means, medians, or quantiles. In this work, we introduce a new probability aggregation method termed "alpha pooling" which builds an aggregated cumulative distribution function (CDF) designed to be closer to a reference CDF over the calibration (historical) period. The aggregated CDFs can then be used to perform bias adjustment of the raw climate simulations, hence performing a "multi-model bias correction". In practice, each CDF is first transformed according to a non-linear transformation that depends on a parameter α. Then, a weight is assigned to each transformed CDF. This weight is an increasing function of the CDF closeness to the reference transformed CDF. Key to the α pooling is a parameter α that describes the type of transformation and hence the type of aggregation, generalizing both linear and log-linear pooling methods. We first establish that α pooling is a proper aggregation method by verifying some optimal properties. Then, focusing on climate model simulations of temperature and precipitation over western Europe, several experiments are run in order to assess the performance of α pooling against methods currently available, including multi-model means and weighted variants. A reanalysis-based evaluation as well as a perfect model experiment and a sensitivity analysis to the set of climate models are run. Our findings demonstrate the superiority of the proposed pooling method, indicating that α pooling presents a potent way to combine GCM CDFs. The results of this study also show that our unique concept of CDF pooling strategy for multi-model bias correction is a credible alternative to usual GCM-by-GCM bias correction methods by allowing handling and considering several climate models at once.

1 Introduction

Over the past century, the Earth's climate has been undergoing significant warming, with the rate of this change accelerating notably in the past 6 decades (IPCC, 2023; Wuebbles et al., 2017). Such warming is believed to be a catalyst not only for extreme events, but also for an alteration in societal and economic systems (Stott, 2016; Wuebbles et al., 2017). In this context, global climate models (GCMs) are seen as critical tools to simulate the future of our climate under different emissions scenarios and provide the scientific community and policy makers with essential climate information to guide adaptation to upcoming climatic changes (e.g. Arias et al., 2021; Eyring et al., 2016; IPCC, 2014).

In recent years, tens of GCMs have been designed, modelling the physical processes in the atmosphere, ocean, cryosphere, and land surface of the planet Earth differently, often by incorporating varied or uniquely modelled parameters (Eyring et al., 2016). However, the complexity of the processes represented means that these models are inevitably imperfect. They contain biases, meaning that, even over the historical period, they can fail to reproduce some statistics of the observed climate (e.g. François et al., 2020). To alleviate such errors, two distinct types of post-processing are typically applied to the models: bias correction and model combination. Bias correction methods aim at applying statistical corrections to climate model outputs, which can be as simple as a delta change (Xu, 1999) or a "simple scaling" of variance (e.g. Eden et al., 2012; Schmidli et al., 2006) or as advanced as multivariate methods adjusting dependencies (e.g. François et al., 2020) such as based on multivariate rank resampling (Vrac, 2018; Vrac and Thao, 2020) or machine learning techniques (e.g. François et al., 2021). Model combination aims to extract a robust signal from the diversity of existing GCM outputs. Models are typically gathered into multi-model ensembles (MMEs), which are synthesized into multi-model means (MMMs). This approach is grounded in the belief that members of the MMEs are "truth-centred". In other words, the various models act as independent samples from a distribution that gravitates around the truth, and as the ensemble expands, the MMM is expected to approach the truth (Ribes et al., 2017; Fragoso et al., 2018). The challenge of combining models lies not only in their inherent differences but also in the construction of the MME itself. While equal weighting of models is a common practice (e.g. Weigel et al., 2010), it does not account for possible redundancy of information between models. Indeed, climate models often share foundational assumptions, parameterizations, and codes, making their outputs redundant (Abramowitz et al., 2019; Knutti et al., 2017; Rougier et al., 2013). As a result, consensus among models does not necessarily result in reliable simulations. Advanced methods, such as Bayesian model averaging (Bhat et al., 2011; Kleiber et al., 2011; Olson et al., 2016) or weighted ensemble averaging (Strobach and Bel, 2020; Wanders and Wood, 2016), have been developed to refine model weights. However, Bukovsky et al. (2019) found that the weighting approach does not substantially change the multi-model mean (i.e. MMM) results.

Furthermore, the usual model combination approach is to apply a global weighting of the models, which can dilute the accuracy of regional predictions. For instance, a model that accurately represents European temperatures might be deemed subpar overall, thus not contributing significantly to the European temperature projection in the ensemble. This could result in a global weighting approach that inaccurately represents this region. To address this, some studies have adopted a regional focus, selecting an optimal set of models for specific regions (Ahmed et al., 2019; Brunner et al., 2020; Dembélé et al., 2020; Sanderson et al., 2017). Yet, such strategies are still suboptimal since they are only valid for a given study area, often of rectangular shape, and thus specific to the use case they have been developed for. Moreover, by construction, traditional model averaging techniques tend to homogenize the spatial patterns that are present in individual models, even though these patterns often stem from genuine physical processes. Approaches that consider per-grid-point model combinations have shown promise in enhancing performances in weather forecasting (Thorarinsdottir and Gneiting, 2010; Kleiber et al., 2011). Geostatistical methods, in particular, offer tools to characterize spatial structures and dependencies, providing a more nuanced approach to ensemble predictions (Gneiting and Katzfuss, 2014; Sain and Cressie, 2007). Recently, Thao et al. (2022) introduced a method that uses a graph cut technique stemming from computer vision (Kwatra et al., 2003) to combine climate models' outputs on a grid point basis. This approach aims to minimize biases and maintain local spatial dependencies, producing a cohesive "patchwork" of the most accurate models while preserving spatial consistency. However, one limit of the graph cut approach is that it only selects one single optimal model per grid point, whereas locally weighted averages of models might enable more subtle combinations that capitalize on the strengths of the ensemble of GCMs.

In addition, one limitation of all aforementioned model combination approaches is that they are all based on combining scalar quantities such as the decadal mean temperature produced by an ensemble of models. However, climate models' outputs are much richer than averages. They typically produce hourly or daily climate variables, from which entire probability distributions can be derived. It therefore makes intuitive sense to combine distributions to obtain an aggregated distribution that can borrow the most relevant aspects of all members of the MME. In statistics, the combination of distributions, or probability aggregation, has been studied for applications in decision science and information fusion. Comprehensive overviews of the different ways of aggregating probabilities and the hypotheses underlying each of them are provided in Allard et al. (2012) and Koliander et al. (2022), notably based on the foundational works of Bordley (1982).

In this study, we introduce an innovative probability aggregation method termed α pooling, which we apply to combine climate projections coming from several GCMs. It builds an aggregated cumulative distribution function (CDF) designed to be as close as possible to a reference CDF. During a calibration phase, an optimization procedure determines the parameters characterizing the transformation from a set of CDFs each representing a model to a reference CDF. This transformation includes weights that increase with the closeness to the reference CDF and a parameter α that characterizes how the transformation takes place. The optimization results in weights that are lower for models that are similar, i.e. that are redundant with each other. In that sense, α pooling combines models while addressing information re-

dundancy. In addition, as α pooling provides an aggregated CDF close to a reference one, corresponding time series can be obtained, for example via quantile–quantile-based techniques (e.g. Déqué, 2007; Gudmundsson et al., 2012) or its variants (e.g. Vrac et al., 2012; Cannon et al., 2015), hence providing bias-corrected values of the combined model simulations. Therefore, α pooling not only combines model CDFs but also corrects biases between the CDF of each model and the reference CDF. So, we bring together, in an original way, "bias correction" and "model combination", which are usually seen as different categories of methods employed by separate scientific communities. We stress that our proposed α-pooling method hinges on a unique concept that allows the simultaneous bias correction of multiple climate model simulations. This is accomplished through the innovative combination of model CDFs, which stands as an original concept in its own right.

Our application of the α-pooling method focuses on the simultaneous combination and bias correction (BC) of climate models over western Europe. Here, each member of the MME is perceived as an individual expert, whose cumulative distribution function (CDF) is used in the combination. We compare α pooling with other model combination and bias correction techniques, including multi-model mean (MMM), linear pooling, log-linear pooling, and CDF transform (CDF-transform, Vrac et al., 2012). Our analysis spans both short-term and extended projections of temperatures (T) and precipitation (PR), encompassed in three distinct experiments. In the first experiment, ERA5 serves as the reference, enabling performance evaluation against observational references. Subsequently, a perfect model experiment (PME) is employed, wherein each model is iteratively used as the reference. This PME approach offers insights into the stability of the alpha-pooling projections compared to other BC techniques, extending to the end of the century. A third experiment investigates the sensitivity of the aggregated CDFs to the choice of a specific subset of models to combine.

This paper is structured as follows. Section 2 describes the climate simulations and the reference used in this work. After some reminders on linear pooling and log-linear pooling, Sect. 3 presents the new α pooling. Section 4 describes the experiments carried out in this work and Sect. 5 describes the results obtained. In Sect. 6, we provide some conclusions and perspectives. Two appendices provide an approximate and faster alternative to the α-pooling method as well as optimal properties.

2 Climate simulations and reference

The reference data used in this study are daily temperature (hereafter T) and precipitation (PR) time series extracted from the ERA5 daily reanalysis (Hersbach et al., 2020) over the 1981–2020 period at a 0.25° horizontal spatial resolution. The western Europe domain, defined as $[10°\,\mathrm{W}, 30°\,\mathrm{E}] \times [30°\,\mathrm{N}, 70°\,\mathrm{N}]$, is considered.

The same variables (T and PR) are also extracted for the period 1981–2100 from 12 global climate models (GCMs) contributing to the sixth exercise of the Coupled Model Intercomparison Project (CMIP6, Eyring et al., 2016). This selection was dictated by the availability of T and PR fields at a daily timescale at the time of analysis: we have only selected models whose data were fully available for the whole period of 1981–2100. The list of GCMs used is provided in Table 1.

To ease the handling of the different simulated and reference datasets, all temperature and precipitation fields have been regridded to a common spatial resolution of $1° \times 1°$. Moreover, for the sake of simplicity, in the following, we only consider winter defined as December–January–February (DJF) and summer data (June–July–August, JJA) separately to investigate and test our α-pooling approach. Then, for each grid point and each dataset, the univariate CDFs of temperature and precipitation are calculated. Here, empirical distributions are employed (i.e. step functions via the "ecdf" R function) in order not to fix the distribution family and thus let the data "speak for themselves". Other parametric or nonparametric CDF modelling methods can be used if needed and appropriate.

3 Combining models via the CDF-pooling approach

The CDF of a random variable X is the function $F : \mathbb{R} \to [0, 1]$ defined as the probability that X is less than or equal to x, i.e. $F(x) = P(X \leq x)$. Combining CDFs thus essentially amounts to combining, or aggregating, probabilities for all values x in a way that makes the aggregated function a CDF, i.e. a non-decreasing function with $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

Allard et al. (2012) offer a review of probability aggregation methods in geoscience, with application in spatial statistics. Aggregation or pooling methods can be characterized according to their mathematical properties. Let us denote $p_1, \ldots, p_N$ as the probabilities to be pooled together and p_G as the resulting pooled probability. A pooling method verifying $p_G = p$ when $p_i = p$ for all $i = 1, \ldots, N$ is said to preserve unanimity. Furthermore, let us suppose that we are in the following case: at least one index i exists such that $p_i = 0$ ($p_i = 1$) with $0 < p_j < 1$ for $j \neq i$. A pooling method which returns $p_G = 0$ ($p_G = 1$) in this case is said to enforce a certainty effect, a property also called the 0/1 forcing property. Notice that for a pooling method verifying this property, deadlock situations are possible when $p_i = 0$ and $p_j = 1$ for $j \neq i$.

In the following, we will consider that there are N CDFs $F_i(x)$, with $i = 1, \ldots, N$. Pooling methods must be applied simultaneously to all probabilities $P(X \leq x) = F(x)$ and $P(X > x) = 1 - F(x)$. The aggregated (or pooled) CDF must verify all properties of a proper CDF recalled above.

Table 1. List of CMIP6 simulations used in this study, along with their run, approximate horizontal atmospheric resolution, and references. The models preceded by a "*" are the five models used in the ERA5 experiment (Sects. 4.1 and 5.1) and the perfect model experiment (Sects. 4.2 and 5.2). All 12 models are used in the sensitivity experiment (Sects. 4.3 and 5.3). See text for details.

Simulation name	Run	Atmospheric resolution	Data reference
*CNRM-CM6-1-HR	r1i1p1f2	∼ 100 km	Voldoire (2019)
*GFDL-CM4	r1i1p1f1	∼ 100 km	Held et al. (2019)
*IPSL-CM6A-LR	r14i1p1f1	∼ 250 km	Boucher et al. (2018)
*MRI-ESM2-0	r1i1p1f1	∼ 100 km	Yukimoto et al. (2019)
*UKESM1-0-LL	r1i1p1f2	∼ 250 km	Tang et al. (2019)
BCC-CSM2-MR	r1i1p1f1	∼ 100 km	Wu et al. (2018)
CanESM5	r10i1p1f1	∼ 500 km	Swart et al. (2019)
INM-CM4-8	r1i1p1f1	∼ 100 km	Volodin et al. (2019)
INM-CM5-0	r1i1p1f1	∼ 100 km	Volodin et al. (2019)
MIROC6	r1i1p1f1	∼ 250 km	Shiogama et al. (2019)
CESM2	r1i1p1f1	∼ 100 km	Danabasoglu et al. (2020)
CESM2-WACCM	r1i1p1f1	∼ 100 km	Danabasoglu et al. (2020)

3.1 Pre-processing: standardizing data

CDFs from climate model simulations can be very different from each other and from ERA5 CDFs. It is thus necessary to perform a preliminary standardization (i.e. basic adjustment) before pooling them. Note that the same operation is performed in many IPCC figures (IPCC WGI, 2021) when working on anomalies (instead of raw simulated or reference data). This allows easier comparison (and combination) of the different datasets. In the present study, temperature and precipitation are standardized differently. For temperature, the simulated data are rescaled such that the mean and standard deviation correspond to those of the reference data:

$$T_{\text{rescaled}} = \frac{T - m_{\text{mod}}}{\sigma_{\text{mod}}} \times \sigma_{\text{ref}} + m_{\text{ref}}, \tag{1}$$

where m_{mod} and σ_{mod} are the mean and standard deviation of the model data to rescale, and m_{ref} and σ_{ref} are those from ERA5. For precipitation, the data are rescaled to get the 90 % quantile similar to that of the reference precipitation:

$$\text{PR}_{\text{rescaled}} = \text{PR} \times \text{Q90}_{\text{ref}}/\text{Q90}_{\text{mod}}, \tag{2}$$

where Q90_{ref} and Q90_{mod} are respectively the 90 % quantiles from ERA5 and the model data to rescale. This choice of 90 % is a trade-off that enables having a robust quantile estimation and also a sufficient spread in the range of precipitation values (Vrac et al., 2016).

In the rest of this paper, all tested pooling methods are then applied to standardized data. As a preliminary step to our new α-pooling approach, we first briefly present the linear pooling and log-linear pooling with their main properties.

3.2 Linear pooling

The linear pooling, whose resulting pooled CDF is denoted as F_{L}, is simply a weighted average of all CDFs:

$$F_{\text{L}}(x) = \sum_{i=1}^{N} w_i F_i(x), \quad \forall x \in \mathbb{R}. \tag{3}$$

F_{L} is a proper CDF if and only if all w_i values are non-negative and $\sum_{i=1}^{N} w_i = 1$. Note that with linear pooling, the probabilities are weighted for a given value x, which is quite different than averaging the quantiles for a given probability, as done in a usual weighted MMM (e.g. Markiewicz et al., 2020). Indeed, in our linear pooling (Eq. 3), the weighted average is performed on the CDFs (i.e. probabilities $F_i(x)$) and not on quantiles (values) of the variable. While there is not an inherent problem with linear pooling, like any linear approach, the method may lack flexibility and thus fail to capture the necessary non-linearity required to adjust to the data and their CDF. That is why non-linear methods (e.g. log-linear pooling) have been developed.

3.3 Log-linear pooling

The log-linear pooled CDF, denoted as F_{LL}, is found by considering that its logarithm is, up to a normalizing factor, a weighted average of the logarithm of the CDFs. Applying this to $F(x)$ and $1 - F(x)$ simultaneously, one gets

$$\ln F_{\text{LL}}(x) = K + \sum_{i=1}^{N} w_i \ln F_i(x) \text{ and}$$

$$\ln(1 - F_{\text{LL}}(x)) = K + \sum_{i=1}^{N} w_i \ln(1 - F_i(x)),$$

where $w_1, \ldots, w_N$ is a set of N non-negative weights and K is the normalizing factor. After some algebra, one finally

obtains

$$F_{\mathrm{LL}}(x) = \frac{\prod_{i=1}^{N} F_i(x)^{w_i}}{\prod_{i=1}^{N} F_i(x)^{w_i} + \prod_{i=1}^{N} (1 - F_i(x))^{w_i}}, \ \forall x \in \mathbb{R}, \quad (4)$$

which is a proper CDF for all non-negative weights w_i. The condition $S = \sum_{i=1}^{N} w_i = 1$ entails unanimity. In simulations, Allard et al. (2012) showed that log-linear pooling of probabilities consistently leads to the best validation scores among all other tested pooling methods. However, log-linear pooling verifies the 0/1 forcing property. This is not necessarily a desirable property since F_{LL} belongs to the interval $(0, 1)$ only for the restricted set of values x such that $0 < F_i(x) < 1$ for all $i = 1, \ldots, N$. Moreover, F_{LL} is undefined as soon as a pair i, j exists with $i \neq j$ such that $F_i(x) = 0$ and $F_j(x) = 1$.

3.4 α Pooling

In order to mitigate the problems faced with the log-linear pooling and the lack of flexibility of the linear pooling, we propose α pooling. Its theoretical expression is presented here. How the parameters are estimated from the models and the reference is shown in the next section. Our approach builds on the $A_{\alpha\text{-IT}}$ transformation proposed in Clarotto et al. (2022), which uses the less stringent power transformation instead of the log transformation used in the log-linear pooling approach. We first recall briefly that a D-part composition is a vector $(v_1, \ldots, v_D)^t$ of D non-negative values such that $\sum_{i=1}^{D} v_i = \kappa$, where κ is an arbitrary positive constant which can be set equal to 1 without loss of generality. In all generality, $A_{\alpha\text{-IT}}$ transforms a composition with D parts (constrained to belong to the simplex of dimension $D - 1$) to a vector with $D - 1$ unconstrained and well-defined coordinates, even when some parts are equal to 0 (Clarotto et al., 2022). For all $x \in \mathbb{R}$, the vector $\boldsymbol{F}(x) = (F(x), 1 - F(x))^t$ can be seen as a two-part composition. In this case, the $A_{\alpha\text{-IT}}$ transformation of $\boldsymbol{F}(x)$ results in a scalar:

$$z(x) = A_{\alpha\text{-IT}}(\boldsymbol{F}(x)) = \alpha^{-1} \boldsymbol{H}_2 \boldsymbol{F}(x)^{\alpha}, \quad (5)$$

where $\boldsymbol{H}_2$ is the $(1, 2)$ Helmert matrix $(\sqrt{2}, -\sqrt{2})$ and where $\boldsymbol{F}(x)^{\alpha}$ is the vector $(F(x)^{\alpha}, (1 - F(x))^{\alpha})^t$ with $\alpha > 0$. The α pooling postulates a linear aggregation of the scores $z_i(x)$ with

$$\begin{aligned} z_G(x) &= \sum_{i=1}^{N} w_i z_i(x) \\ &= \frac{\sqrt{2}}{\alpha} \sum_{i=1}^{N} w_i \left(F_i(x)^{\alpha} - (1 - F_i(x))^{\alpha} \right), \end{aligned} \quad (6)$$

where, as above, $w_1, \ldots, w_N$ is a set of N non-negative weights summing to 1, i.e. with $\sum_{i=1}^{N} w_i = 1$. The α-pooling aggregated CDF F_G is thus the CDF such that $z_G(x) = \frac{\sqrt{2}}{\alpha}(F_G(x)^{\alpha} - (1 - F_G(x))^{\alpha})$. Hence, for each x, $F_G(x)$ solves

$$\begin{aligned} F_G(x)^{\alpha} - (1 - F_G(x))^{\alpha} &= \alpha z_G(x) \\ &= \sum_{i=1}^{N} w_i \left(F_i(x)^{\alpha} - (1 - F_i(x))^{\alpha} \right). \end{aligned} \quad (7)$$

Let us define the function

$$G(y) = \alpha^{-1} \left[y^{\alpha} - (1 - y)^{\alpha} \right], \quad (8)$$

with $0 \leq y \leq 1$. $G(y)$ is an increasing one-to-one function on $[0, 1]$, with $G(0) = -\alpha^{-1}$, $G(1/2) = 0$ and $G(1) = \alpha^{-1}$. One thus gets $F_G(x) = G^{-1}(z_G(x))$, where G^{-1} is the inverse function of G, which exists and is unique. There is unfortunately no general closed-form solution to Eq. (7) for all values of α, but the aggregated probability can be found as

$$F_G(x) = G^{-1}(z_G(x)) = \arg \min_{y \in [0,1]} (G(y) - z_G(x))^2 \quad (9)$$

using numerical optimization. It is straightforward to check that when $\alpha = 1$, the solution to Eq. (7) is the linear pooling. Likewise, using $\lim_{\alpha \to 0} F_i(x)^{\alpha} = 1 + \alpha \ln F_i(x)$, it is easy to check that the α pooling tends to the log-linear pooling as $\alpha \to 0$. We can show the following.

Proposition 1. The function $F_G(x)$ defined in Eqs. (7) and (9) is a proper CDF.

Proof. The derivative of $z_G(x)$ with respect to x is $z_G(x)' = \sqrt{2} \sum_{i=1}^{N} w_i f_i(x)(F_i(x)^{\alpha-1} + (1 - F_i(x))^{\alpha-1}) \geq 0$. Hence, $z_G(x)$ is a non decreasing function of x. Since the derivative of the function $G(y)$ with respect to y is also positive, the function $F_G(x) = G^{-1}(z_G(x))$ is increasing because it is the composition of two increasing functions. In addition, using $\lim_{x \to -\infty} F_i(x) = 0$ and $\lim_{x \to \infty} F_i(x) = 1$ together with $\sum_{i=1}^{N} w_i = 1$, it is easy to check that $\lim_{x \to -\infty} F_G(x) = 0$ and $\lim_{x \to \infty} F_i(x) = 1$. Hence, F_G is a proper CDF. □

The α pooling presented in Eq. (7) mitigates the principal inconvenience of the log-linear pooling, since it eliminates the 0/1 forcing property and it is well defined for all values of $F_i(x)$. In addition it seamlessly accommodates the case $F_i(x) = 0$ and $F_j(x) = 1$ with $i \neq j$.

Remark 1. The constraint on the sum of the weights can be relaxed. In this case, if $S = \sum_{i=1}^{N} w_i > 1$, F_G will still be a proper CDF because y is constrained to belong to the interval $[0, 1]$ in Eq. (9). But if $S < 1$, the lower and upper limits of F_G will not be equal to 0 and 1, respectively, with $\lim_{x \to -\infty} F_G(x) = G^{-1}(-S/\alpha) = b > 0$ and $\lim_{x \to \infty} F_G(x) = G^{-1}(S/\alpha) = 1 - b < 1$.

In Appendix A, we present a closed-form expression which is a very good approximate solution to Eq. (7) in most cases, i.e. except when $S > 1$. Then in Appendix B, we present some optimal properties of the α pooling presented above related to the fact that α pooling derives from the general class of quasi-arithmetic pooling methods and corresponds to a proper scoring rule (Neyman and Roughgarden, 2023).

An illustration is provided in Fig. 1a for $N = 3$ distributions F_1, F_2, and F_3 to be combined, respectively corresponding to a lognormal CDF, a Gaussian one, and a Student's t distribution. A uniform CDF is arbitrarily fixed as a reference. Despite the fact that they belong to very different families, the four CDFs are constructed here such that they have the same mean and variance; i.e. they respect the constraints of our real-case application (see Sect. 3.1). For this example, the estimated α parameter tends to 0 and $w_1 = 0.06$, $w_2 = 0.79$, and $w_3 = 0$. The higher value for w_2 than for w_1 or w_3 indicates that the reference uniform CDF is closer to F_2 (i.e. the Gaussian distribution) than to the others, which was expected considering the behaviour of F_1 and F_3 in the lower tail. Overall, given the difficulty of the illustration (very different CDFs), the α-pooling pooled CDF (shown by the dashed black line in Fig. 1a) is able to approximate the reference CDF reasonably well (blue line), despite some larger errors on the upper tail. Notice that it performs significantly better than the linear pooling (red and green lines). In addition, Fig. 1b displays the z scores (i.e. G as a function of x in Eq. 8) for the three CDFs to be combined, the reference one and the resulting α-pooling CDF.

3.5 Estimating the parameters and computing the aggregated CDF

Given N CDFs F_i, $i = 1, \ldots, N$ and a reference CDF F_0, the parameters are estimated by minimizing the quadratic distance:

$$Q = \sum_{k=1}^{K} (x_k - x_{k-1})(F_0(x_k) - F_G(x_k))^2, \tag{10}$$

where $F_G(x)$ is obtained by solving Eq. (7) and where $x_0, \ldots, x_K$ is an increasing sequence discretizing the real line. The L-BFGS-B optimization algorithm (Byrd et al., 1995) is launched to minimize Eq. (10) and find the weights and the α parameter. This algorithm is a limited-memory extension of the BFGS quasi-Newton method and allows handling simple bound constraints on the variables. The parameter α and the weights must be positive, and the weights can be constrained to sum to 1 or not. In the following, the sum S of weights is let free, i.e. not necessarily equal to 1. Indeed, preliminary results indicated that this freedom gives more flexibility to the α pooling and thus better aggregated CDFs (not shown). When unconstrained, it was found that in most cases the optimal sum S was close to 1. There are two reasons for this: when $S < 1$, the pooled CDF varies from b to $1 - b$ (see Remark 1). As a consequence, b must be as close to 0 as possible and hence S as close to 1 as possible for the pooled CDF to be close to the reference; when $S > 1$, the inverse of all values $z_G < -1/\alpha$ (values $z_G > 1/\alpha$) leads to the same inverse equal to 0 (1). A value of S that is too high is therefore likely to lead to a lack of fit in the lower and upper tails. However, when $S < 1$, as the aggregated CDF goes from $b > 0$ to $1 - b < 1$, it is not a proper CDF per se. Hence, a "min–max" rescaling of the aggregated CDF F_G is performed such that the rescaled CDF F_{resc} is always in $[0, 1]$.

$$\begin{aligned} F_{\text{resc}}(x) &= \frac{F_G(x) - \min_x(F_G(x))}{\max_x(F_G(x)) - \min_x(F_G(x))} \\ &= \frac{F_G(x) - b}{(1-b) - b} = \frac{F_G(x) - b}{1 - 2b} \end{aligned} \tag{11}$$

In practice, this rescaling is only very slight as b is very often found to be extremely small, say less than 10^{-3}.

The weights are easily interpretable since, as a rule, the higher the weight w_i, the closer F_i is to the reference F_0. The parameter α has a less immediate interpretation. As shown in Clarotto et al. (2022), the $A_{\alpha\text{-IT}}$ transform can be seen as a difference between the Box–Cox transformation of $F(x)$ and that of $(1 - F(x))$ (see also Appendix A). The parameter α can thus be interpreted as the power necessary to deform all CDFs (reference and models) in order to get an optimal linear pooling for these deformed CDFs, from log transform ($\alpha \to 0$) to no transform ($\alpha = 1$), to quadratic transform ($\alpha = 2$).

3.6 Benchmarking α pooling: CDF multi-model mean (MMM) and linear pooling

As a benchmark for evaluating the α-pooling approach, two CDF pooling methods are also applied. The first one is the simplest and consists of defining a "mean" CDF based on the N CDFs to be combined. Let us consider, for example, $N = 2$ GCMs with CDFs F_1 and F_2, say of temperature, for a given grid cell. For any temperature value x, the mean CDF $F_{\text{MMM}}(x)$ corresponds to the average of $F_1(x)$ and $F_2(x)$. An example is given in Fig. 1a for the three distributions used to illustrate the α-pooling method. Here, F_{MMM} is shown as a dashed red line. Note that, for MMM, the reference CDF is not used at all, as the N CDFs are linearly averaged with weights all equal to $1/N$, whatever the quality of the different model CDFs with respect to that of the reanalysis. Hence, it is not surprising that α pooling better approximates the reference CDF over the calibration period.

The second CDF pooling method applied for comparison is the linear pooling described in Eq. (3). Here, contrary to MMM, the reference CDF is used to infer the weight parameters. By comparing the linear and α-pooling methods, we can assess the potential added value brought by the alpha parameter.

The same illustration as previously used is also given in Fig. 1a for linear pooling, with the dashed green line. Based on this illustrative – but difficult – example, it is clear that the introduction of the α parameter allows us to get closer to the reference CDF, at least over the calibration period. This is clear from the value of the L^2 norm computed between the resulting CDF (i.e. α pooling, linear pooling, or MMM) and the reference: α pooling has the smallest L^2 (0.003), and linear pooling's L^2 is doubled (0.006), while

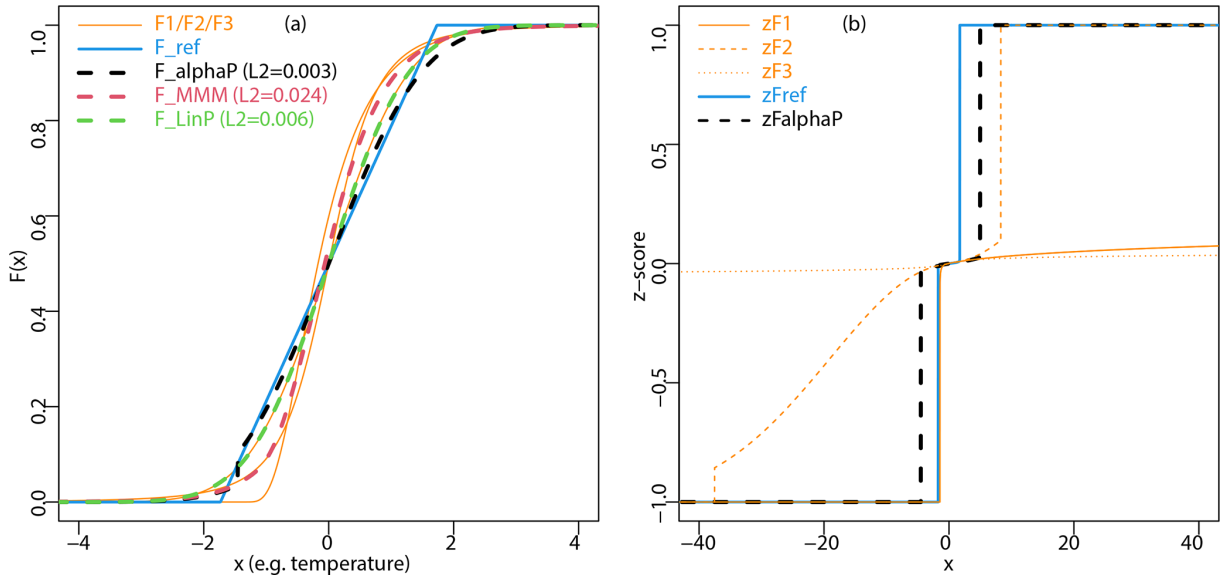

Figure 1. Illustration for $N = 3$ distributions F_1, F_2, and F_3 to be combined, respectively corresponding to a lognormal CDF, a Gaussian one, and a Student's t distribution. A uniform CDF is arbitrarily fixed as a reference. Note that the four CDFs are constructed here with the same mean and variance to respect the constraints of our real-case application. Panel **(a)** displays the three CDFs to combine (orange lines), the reference CDF (blue line), and the resulting α-pooling CDF (dashed black line), MMM CDF (dashed red line), and linear pooling CDF (dashed green line). For each pooling method, the value of the L^2 norm between the resulting CDF and the reference one (i.e. the quadratic distance Q in Eq. 10) is also indicated. Note that the reference is not used to perform MMM. Panel **(b)** shows the z scores (i.e. function G in Eq. 8, where $z = G(F(x))$ with $F(x)$ the CDF) for the three CDFs to be combined, the reference one, and the α-pooling CDF.

it is almost 10-fold for MMM (0.024). However, one major objective of this study is also to evaluate how MMM, linear pooling, and α pooling behave in a projection period wherein climate changes occurs. When driven only by model CDFs over a projection (future) period, are the three pooling methods able to capture the changes in reference (temperature or precipitation) CDFs?

3.7 Bias corrections from CDF pooling results

The aggregated CDF can be used within a CDF-based bias correction method applied to GCMs and, hence, to obtain corrected simulations in a way that preserves the temporal rank dynamics. Indeed, once $\hat{F}$ is estimated over a projection period, one can apply a quantile mapping technique (e.g. Gudmundsson et al., 2012, among many others) between $\hat{F}$ and the CDF F_m of a given model m over the same period: for any value x simulated by model m, it consists of finding the value y such that $\hat{F}(y) = F_m(x)$ which is equivalent to

$$y = \hat{F}^{-1}(F_m(x)), \tag{12}$$

where $\hat{F}^{-1}$ is the inverse CDF, allowing computing the quantile associated with a given probability. Therefore, by applying Eq. (12) successively to all simulations from model m, we can obtain bias corrections. Those have the same rank chronology as that of model m but their values follow distribution $\hat{F}$. By applying this bias correction technique to the different models employed within the MMM, linear pooling, or α-pooling methods, the N bias-corrected time series have the exact same distribution (i.e. $\hat{F}$) but their temporal dynamics are different, as stemming from the N models.

3.8 Model-by-model bias correction via CDF-t

To evaluate the pros and cons of the bias corrections brought by the proposed pooling approaches, a more traditional "model-by-model" bias correction method is also applied for comparison: the "cumulative distribution function – transform" (CDF-t) method (Michelangeli et al., 2009; Vrac et al., 2012). It consists of a quantile mapping technique (e.g. Panofsky and Brier, 1968; Haddad and Rosenfeld, 1997; Déqué, 2007; Gudmundsson et al., 2012) allowing accounting for changes in the distributional properties of the climate simulations from the reference to the projection period. The reference CDF F_{Rp} over the projection period is first estimated as a composition of F_{Rc}, F_{Mc}, and F_{Mp}, as well as the reference CDF over the calibration period, the model CDF over the calibration period, and the projection period:

$$\hat{F}_{\mathrm{Rp}}(x) = F_{\mathrm{Rc}}\left(F_{\mathrm{Mc}}^{-1}(F_{\mathrm{Mp}}(x))\right), \tag{13}$$

where F_{Mc}^{-1} is the inverse CDF of F_{Mc}. See Vrac et al. (2012) or François et al. (2020) for more details. Based on the estimated projection reference CDF, a quantile mapping is then fitted between $\hat{F}_{\mathrm{Rp}}$ and F_{Mp} to bias-correct the simulations

from the model M. Hence, in the case of N climate models to adjust, N CDF-t bias corrections are defined and applied.

4 Design of experiments

In the following, three experiments are described to evaluate and compare α pooling, linear pooling, MMM, and CDF-t. For the sake of clarity and space, these experiments are carried out separately over two seasons only: winter (December, January, February – DJF) and summer (June, July, August – JJA). Only winter results are given in the following but summer results can be found in the Supplement.

4.1 ERA5 experiment

The first experiment considers ERA5 reanalysis as a reference. When considering linear and α-pooling methods, for each grid point and variable, we calibrate the approaches using N climate models with ERA5 data as a reference over the calibration period of 1981–2000. Then, we use the calibrated parameters (w_i and α) to combine the models CDFs over the projection period of 2001–2020. For CDF-t, the same calibration period (1981–2000) is used, and the corrections are made for each model independently for the projection period (2001–2020). For MMM, the CDFs of the climate models are directly averaged over 2001–2020. The results of each approach are then compared to the ERA5 data over 2001–2020.

In this experiment, only five GCMs are used. This is partly constrained by the α-pooling method that can have stability issues inferring the parameters when combining a large number of models. When a relatively high number of models (i.e. CDFs) are combined, such as 10, depending on the initialization values of the parameters in the inference algorithm, the "optimal" final parameters may vary. In essence, the optimized parameters are unstable in such a case. This is because many local minima attain undistinguishable L^2 distances. Indeed, while final parameters may differ between initializations, the minimized criterion values – specifically the quadratic distance in the CDF space outlined in Eq. (10) – remain relatively consistent, often converging to similar or nearly identical values. Although it has been tested with more than 10 models, the use of five GCMs appeared to be a good compromise in the sense that (i) it ensured not only stability in the quadratic criterion but also consistency in the final optimized parameters, (ii) it allows a reasonable computation time (e.g. no more than a few minutes of computations for each location and/or variable), and (iii) it employs a sufficient number of simulations to get robust results. These five GCMs (indicated with "*" in Table 1) were selected on the basis of a preliminary analysis showing that they approximately represent the spread of the future evolution of all 12 GCMs (not shown). Note that four models (IPSL-CM6A-LR, MRI-ESM2-0, UKESM1-0-LL, GFDL-CM4) out of the five selected ones are consistent with the choice made in the ISIMIP3 project (Lange and Büchner, 2021; Lange, 2021) for bias correction objectives.

The evaluations are performed in terms of biases of the obtained 2001–2020 temperature and precipitation with respect to ERA5. For each grid point, dataset, variable, and season (winter or summer), some statistics T are calculated. For temperature, statistics include the mean, standard deviation, and 99 % quantile (Q99). For precipitation, we consider the conditional mean given a wet state (Cm), probability of a dry day (P_1), and the 99 % quantile. A day with a PR value lower than 1 mm is considered dry (and thus >1 mm wet).

Then, absolute biases are calculated as

$$B(m,T) = T(m) - T(\text{ERA5}) \tag{14}$$

for temperature mean and Q99, while relative biases are calculated as

$$B(m,T) = \frac{T(m) - T(\text{ERA5})}{T(\text{ERA5})} \tag{15}$$

for temperature standard deviation and precipitation conditional mean, P_1, and Q99. m denotes the method (α pooling, linear pooling, MMM, or CDF-t) and $T(X)$ the statistics calculated from dataset X (ERA5 or method results).

4.2 Perfect model experiment (PME)

As the ERA5 experiment evaluates the methods for a projection period (2001–2020) very close to the calibration one (1981–2000), it does not allow understanding their quality in a strong climate change context. To perform such an assessment, we propose a "perfect model experiment" (e.g. de Elía et al., 2002; Vrac et al., 2007, 2022; Krinner and Flanner, 2018; Robin and Vrac, 2021; Thao et al., 2022, among many others). The main idea is that one model, among N, is taken as the reference. For the four methods, the procedure is the following.

- α Pooling and linear pooling are calibrated to combine the other $N-1$ models over 1981–2000. The obtained parameters (i.e. w_i and α for α pooling or w_i only for linear pooling) are next used to combine the $N-1$ models over five different future 20-year periods: 2001–2020, 2021–2040, 2041–2060, 2061–2080, and 2081–2100.

- The same approach is followed for CDF-t: one model serves as a reference over 1981–2000 to calibrate CDF-t – here separately for each of the $N-1$ remaining models – which is then used to bias-correct each model simulation over the five future periods.

- As previously for the ERA5 experiment, MMM does not require any calibration. CDF averaging is directly applied to combine the $N-1$ models for each of the five periods.

Over each future period and each grid point, biases can then be evaluated with respect to the reference model. For temperature, it includes absolute biases (Eq. 14) of the mean, 1 % quantile, 99 % quantile, minimum, and maximum, as well as relative biases (Eq. 15) of standard deviation. For precipitation, relative biases are computed for the conditional mean a given wet state, probability of a dry (< 1 mm) day, standard deviation, conditional 99 % quantile a given wet state, unconditional 99 % quantile, and maximum.

Hence, no observational or reanalysis data are used as a reference in this experiment. Indeed, this PME is made under the "models are statistically indistinguishable from the truth" paradigm (e.g. Ribes et al., 2017), where "the truth and the models are supposed to be generated from the same underlying probability distribution" (Thao et al., 2022). Therefore, an evaluation framework based on this paradigm can consider any model as the reference. In practice in our PME, the same five models as in the ERA5 experiment (Sect. 4.1) are used and each model is used in turn as the reference. The four methods are thus tested on a diversity of possible references, encompassing cases where the truth can be either in the centre of the multi-model distribution or far in the tail.

4.3 Sensitivity of projected future CDFs to the choice of models

Finally, our third experiment aims to evaluate the uncertainty brought by the choice of the N models to combine and/or bias-correct. If this sensitivity is not very present over the calibration period – by construction, linear pooling, α pooling, and CDF-t are relatively close to the reference CDFs over this period – or over periods very close to the calibration, the results of the four methods applied to long-term future projections can be sensitive to the chosen N models. To evaluate this sensitivity, for each variable, linear pooling, α pooling, and CDF-t are calibrated with respect to ERA5 data over 1981–2000. Then, all methods are applied to 2081–2100 projections. However, in this experiment, linear pooling, α pooling, and MMM do not combine a unique set of five models (as in the ERA5 experiment). Instead, 100 different sets of $N = 5$ models among the 12 presented in Table 1 are randomly drawn. The resulting 100 samples have been checked to contain each model in a uniform proportion (not shown). The linear pooling, α-pooling, and MMM methods are then applied 100 times, each with five models to combine, while CDF-t is applied to the 12 models separately. The 2081–2100 results obtained from each method and set of models do not allow any evaluation per se, as there is no reference over the future period. However, the use of multiple sets of models allows quantifying and comparing the statistical uncertainty brought by the choice of models for each method. In this experiment, for both temperature and precipitation, only six grid points are considered, corresponding to major capitals of the geographical domain: Paris (France), London (UK), Rome (Italy), Madrid (Spain), Berlin (Germany), and Stockholm (Sweden).

5 Results

5.1 ERA5 experiment results

Before looking at the results of the ERA5 experiment, it is interesting to visually understand how the α-pooling parameters are spatially distributed over the geographical domain. Hence, Fig. 2 displays maps of the winter (Fig. 2a and c) and summer (Fig. 2b and d) for the α parameter, temperature (Fig. 2a and b), and precipitation (Fig. 2c and d). First, note that the range of α is not the same for T and PR. While for temperature most of the values are lower than 1 (no unit), the range goes up to 2.5 for precipitation. Moreover, for both seasons, more pronounced spatial structures appear for T than for PR, with the latter α maps appearing more "pixelated". This can be explained by the widely recognized spatial variability of precipitation, encompassing both occurrence and intensity, which is often challenging to accurately capture in climate models and thus reflected in the spatial diversity in the estimated alpha-pooling parameters. However, globally, even for PR, large regions share similar α values, indicating some spatial consistency of the parameters.

Regarding the weight parameters of α pooling, winter maps are provided in Figs. 3 and 4 for temperature and precipitation, respectively. The results for summer are given in Figs. S1 and S2 in the Supplement. The spatial structures of the weights are clearly visible (for both T and PR) and even more pronounced than for the α maps. This strongly indicates that α pooling identifies large zones where some models have a larger influence on the combination and, thus, whose CDFs are closer to that of ERA5. Note, however, that for both variables, none of the models has the highest weights for all grid points of the domain. In other words, over this European region, each of the five models brings some valuable contribution, although there is contrast depending on the subregion. For example, with temperature, UKESM (Fig. 3e) shows the strongest contributions over the Mediterranean Sea, while MRI-ESM2 (Fig. 3d) displays the largest weights over the northeast part of the domain. Interestingly, the spatial distributions of the weights are not the same for T and PR. Thus, there is no clear link between the contribution of each variable, confirming that results from one variable cannot be generalized to another.

A concentration index is displayed in panel (f) of Figs. 3 and 4, which is equal to the sum of the squares of the five normalized weights. It takes the value 1 when one single GCM takes all the weight and reaches a minimum of $1/N = 0.2$ when the five normalized weights are equally distributed. The concentration index can only be applied to weights summing to 1. In our implementation of α pooling, the sum of weights is let free and, thus, not constrained to 1. Although this sum remains quite close to 1 (mostly between 0.95 and

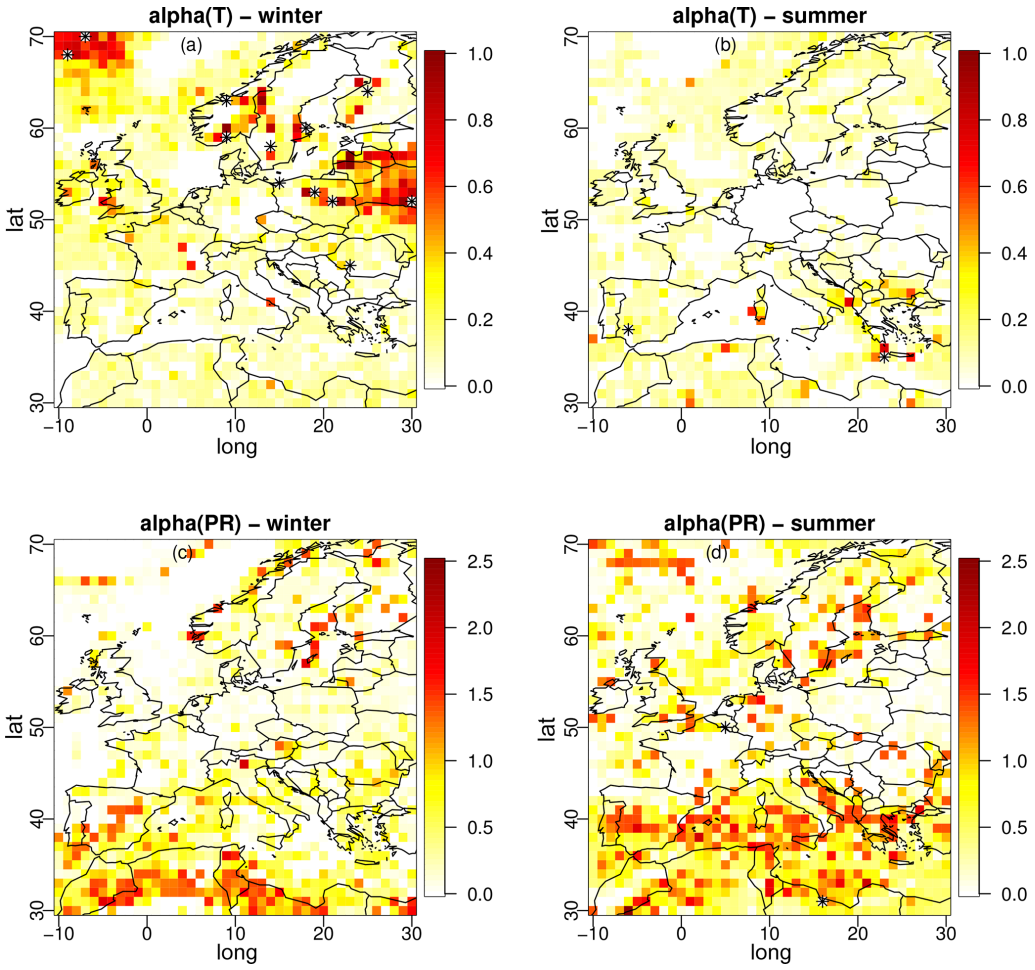

Figure 2. From α pooling, maps of the parameters α obtained within the ERA5 experiment for temperature **(a, b)** and precipitation **(c, d)** over winter **(a, c)** and summer **(b, d)** seasons.

1.05 for temperature and between 0.92 and 1.1 for precipitation, not shown), normalization is required, which is accomplished by dividing the weights by $S = \sum_{i=1}^{N} w_i$ before computing the concentration index. For temperature, Fig. 3f shows relatively well-distributed weights (most concentration indices between 0.2 and 0.7) despite two zones (close to Italy and close to Greece) strongly influenced by one single GCM (UKESM1, see Fig. 3e). For precipitation, more zones show a concentration index close to 1: for example, the northwestern part of the domain and northern France (MRI-ESM2, Fig. 4f), southern Norway and the northeastern part of the domain (CNRM-CM6, Fig. 4a), and the eastern Adriatic coast (UKESM1, Fig. 3e). Also note that the maps of weights obtained from linear pooling are given in Figs. S3 and S4 in the Supplement for temperature and Figs. S5 and S6 in the Supplement for precipitation. Interestingly, the spatial structures of the weights and concentration indices are very similar to those from α pooling. This confirms that the α parameter does not structurally modify the interpretation of the weights but brings additional flexibility.

The biases of the different methods with respect to 2001–2020 ERA5 are shown in terms of mean, standard deviation, and Q99 for winter temperature in Fig. 5 and in terms of conditional mean given a wet state, probability of a dry day (P_1), and Q99 for winter precipitation in Fig. 6. The equivalent figures for summer are provided in Figs. S7 and S8 in the Supplement for temperature and precipitation, respectively. In theses figures, the columns are associated with the different biases. The top row shows maps of biases for MMM: row 2 for α pooling, row 3 for CDF-t, and the fourth row for linear pooling. Note that, because CDF-t is applied separately for each GCM, the third row corresponds to the grid point median of the CDF-t biases. The fifth (bottom) row displays a more condensed view of the results via box plots of biases.

For temperature (Fig. 5), the differences between the maps of biases from the four methods are not very pronounced. This is especially true for the biases in mean temperature and standard deviation (SD). Some more differences appear for Q99. For instance, MMM (Fig. 5c) shows relatively high positive bias ($\sim 4\,^{\circ}$C) over the northeastern part of the

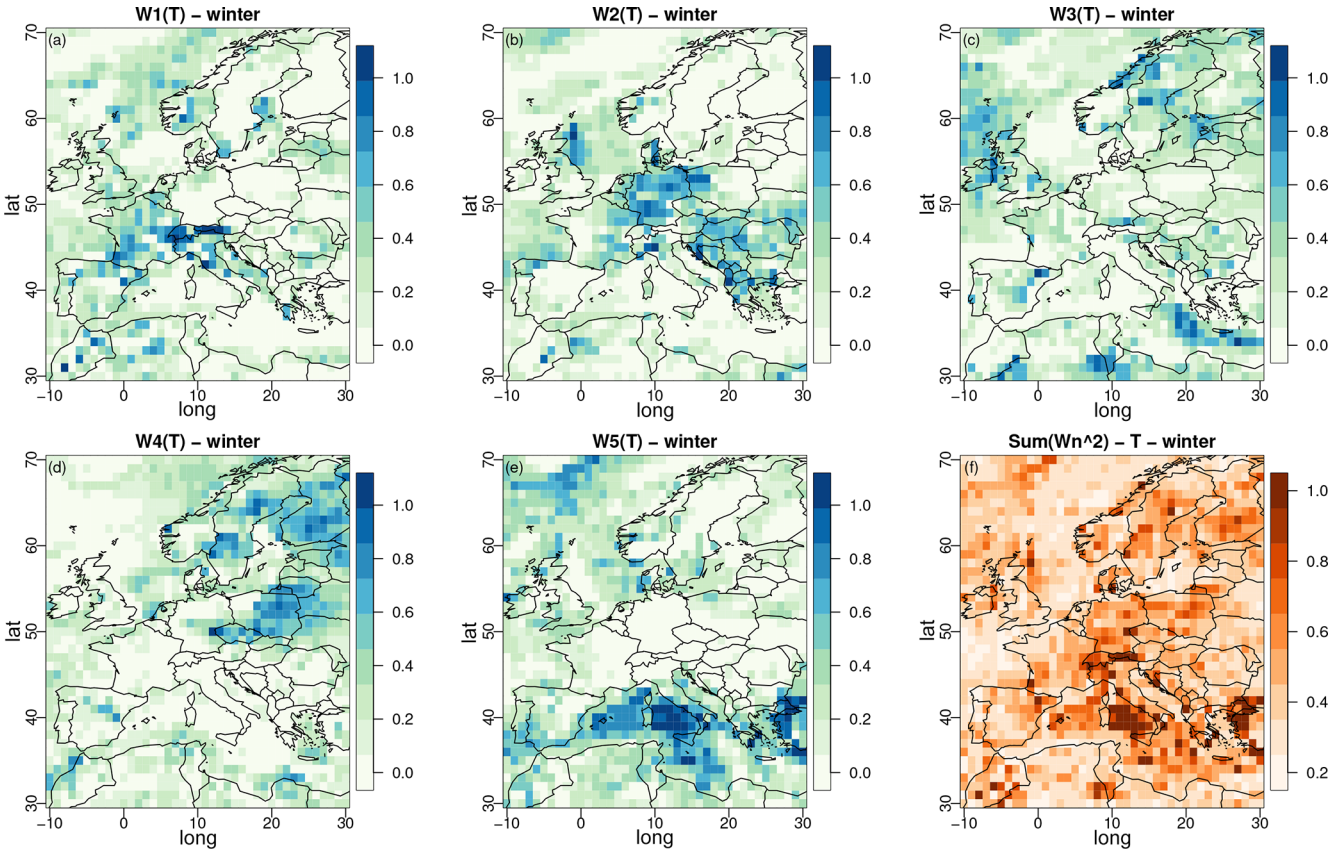

Figure 3. Maps of the weight parameters from α pooling for winter obtained with the ERA5 experiment for temperature over winter. Models 1 to 5 respectively correspond to CNRM-CM6-1-HR, GFDL-CM4, IPSL-CM6A-LR, MRI-ESM2-0, and UKESM1-0-LL. Panel **(f)** displays the concentration index, equal to sum of the squares of the five normalized weights. The results for summer are given in Fig. S1.

domain (Sweden and Finland), while biases for α-pooling Q99 (Fig. 5f), CDF-t (median) (Fig. 5i), and linear pooling (Fig. 5l) do not present this structure. Also, the CDF-t median Q99 (Fig. 5i) has a positive ($\sim$ 1–2 °C) bias pattern over the central domain (Germany, Italy, Poland, Hungary, Romania), while the three other methods show more nuanced and mixed structures. When looking at the more integrated box plot view (bottom row in Fig. 5), similar behaviour of α pooling, linear pooling, and MMM is visible for the three biases: the box plots are relatively equivalent from one method to another. However, even though this is also the case for the CDF-t median biases – at least for mean and SD and to some extent for Q99 – the individual CDF-t biases (i.e. GCM by GCM) show much larger variability, indicating that relying on a single GCM to perform the bias correction might lead to stronger errors within this ERA5 experiment.

For precipitation (Fig. 6), conclusions are somewhat similar, but some more differences between methods are now more visible. For example, in the Norwegian Sea, the relative biases of P_1 for MMM (Fig. 6b) have a large and strongly positive structure (~ 1) that does not appear in the other methods. Another example is the mostly negative bias (~ -1) in α-pooling Q99 (Fig. 6f) over the North African part of the domain, while MMM and (median) CDF-t show mostly highly positive biases and linear pooling more mixed patterns for this region. The box plot view for winter precipitation is similar to that for temperature: roughly equivalent box plots for the four methods, with more variability from the individual CDF-t results.

Note, however, that the ERA5 experiment results for summer (Figs. S7 and S8) show more differences between the four methods – especially in the box plots – slightly in favour of the linear pooling and α-pooling methods, which show box plots more centred around 0 for all biases and variables.

In the ERA5 experiment, the results are relatively similar for the four methods. This indicates that the added flexibility provided by α pooling may not be required over the 1981–2020 period of ERA5. This can nevertheless be different when considering other projection periods and reference datasets. Furthermore, the evaluation (2001–2020) and calibration (1981–2000) periods are quite close to each other, resulting in similar outcomes for both periods. These two results suggest that distinguishing between the different methods may be challenging in a climate that is relatively stable

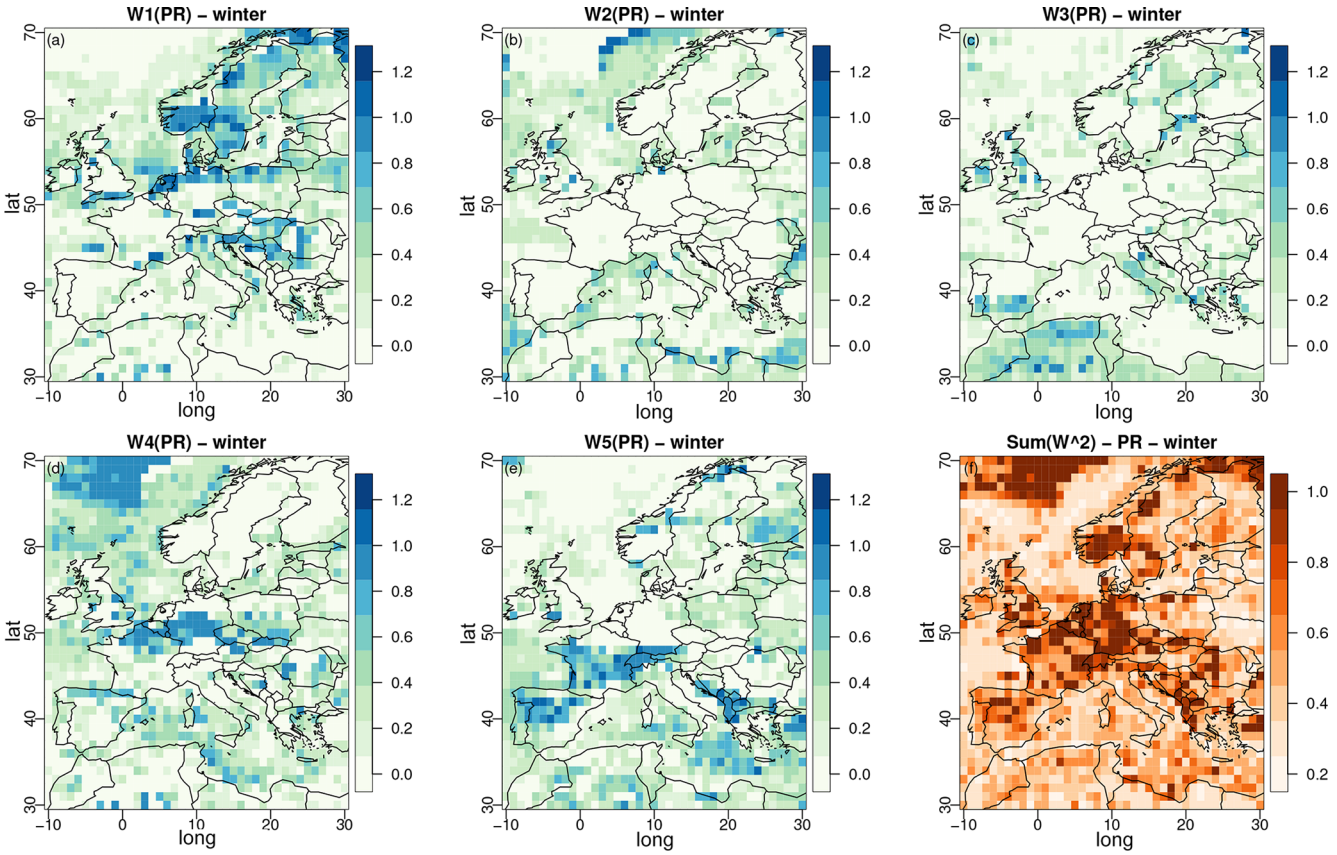

Figure 4. Same as Fig. 3 but for precipitation. The results for summer are given in Fig. S2.

or undergoing minimal change. However, our primary objective is to assess and compare our various pooling strategies in the context of significant climate change. Given that climate changes (in temperature and precipitation) from 1980 to 2100 in the SSP8.5 CMIP6 simulations are significantly more pronounced than what can be seen in the whole ERA5 reanalysis dataset over western Europe, the perfect model experiment (PME) will effectively and more clearly fulfil this purpose.

5.2 PME results

PME is first applied here to winter temperature, and summer results are in the Supplement. For each period and method, the box plots of the different biases, computed at each grid point, are provided in Fig. 7 (PME summer temperatures are in Fig. S9 in the Supplement). As expected, for all biases, the more distant the period, the larger the box plots, indicating an increase in possible statistical errors for periods further in the future. For brevity, we now focus on the last period (i.e. p6, 2081–2100), which results in the most pronounced differences between methods. For mean T bias (Fig. 7a), all four approaches show similar performance, although CDF-t has a wider box plot. The bias of minimum temperature (Fig. 7e) is roughly equivalent for MMM and the linear or α-pooling approaches, while CDF-t presents, on average, a negative bias. However, α pooling appears slightly better than MMM and linear pooling for the temperature 1 % quantile (Q01, Fig. 7c), with CDF-t having a median bias (i.e. box plot centre) equivalent to α pooling but with larger variability. For maximum temperature (Fig. 7f), CDF-t shows a strongly positive bias, while its biases look reasonable – at least more comparable to the other methods – for standard deviation (Fig. 7b) and 99 % quantile (Q99, Fig. 7d). Globally, for temperature standard deviation (Fig. 7b), Q99 (Fig. 7d), and maximum value (Fig. 7f), α pooling is more robust than the other methods since it clearly provides smaller biases over the 2081–2100 period.

Figure 8 shows the PME results for winter precipitation (summer results are in Fig. S10 in the Supplement). As was the case for temperature, the more distant the period, the wider the box plots, although this is less pronounced here. Over 2081–2100, CDF-t results are often the most biased, except for the probability of a dry day (P_1, Fig. 8b) where it is as good as the other methods. As in Fig. 7f, the maximum values of precipitation from CDF-t (green box plot in Fig. 8f) show strong biases with a high variability. Regarding MMM, linear pooling, and α-pooling methods, they give roughly similar biases in terms of conditional mean precip-

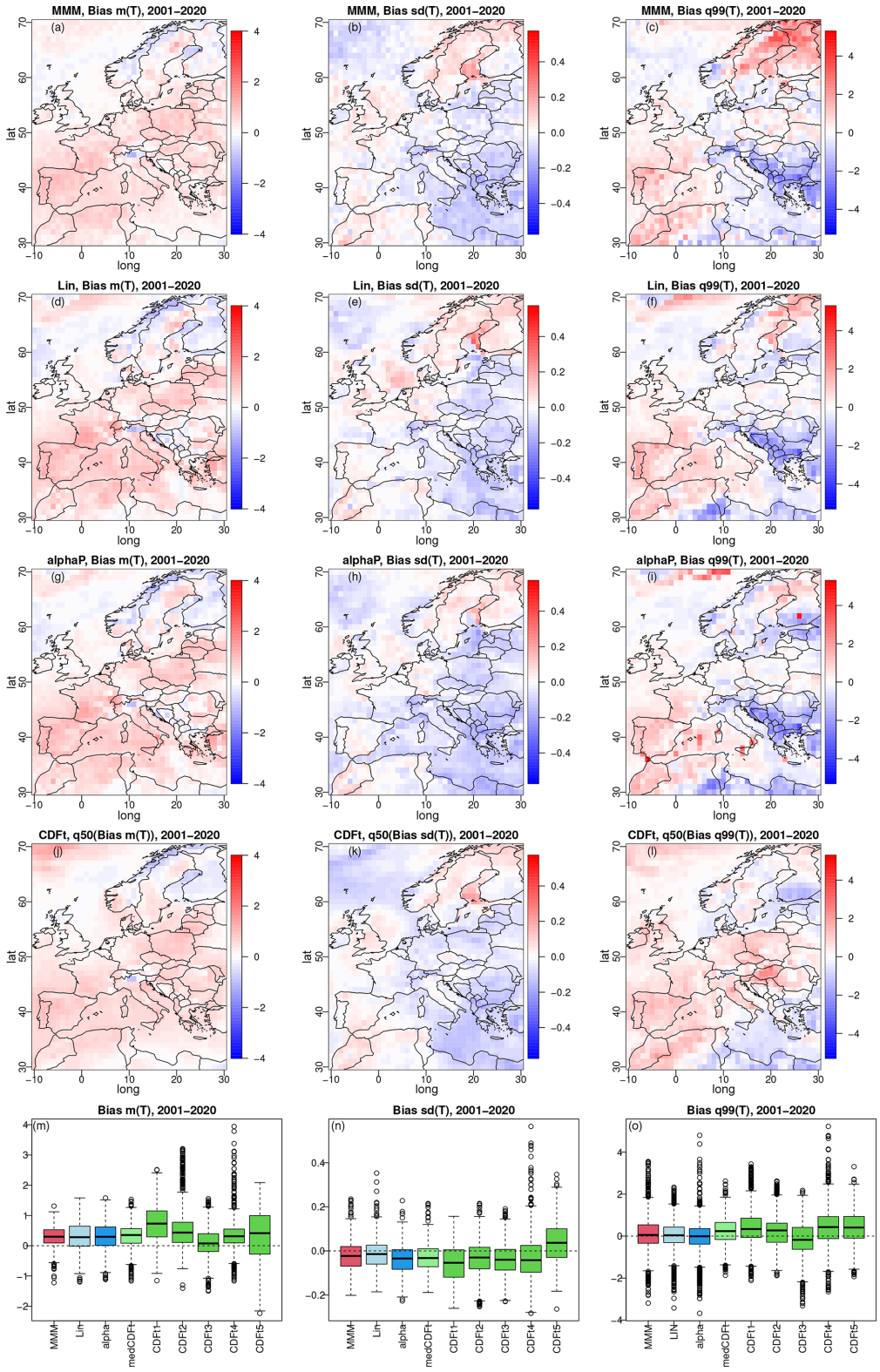

Figure 5. Biases in mean (left column: **a, d, g, j, m**), standard deviation (middle column: **b, e, h, k, n**), and 99 % quantile (right column: **c, f, i, l, o**) for winter temperature from MMM **(a–c)**, α pooling **(d–f)**, CDF-t **(g–i)**, and linear pooling **(j–l)** under the 2001–2020 (projection) time period of the ERA5 experiment. Third row corresponds to the grid point median of the CDF-t biases. **(m–o)** Box plots of biases for MMM, linear pooling, α pooling, and the median CDF-t biases, as well as for each of the five CDF-t results. The results for summer temperature are given in Fig. S7.

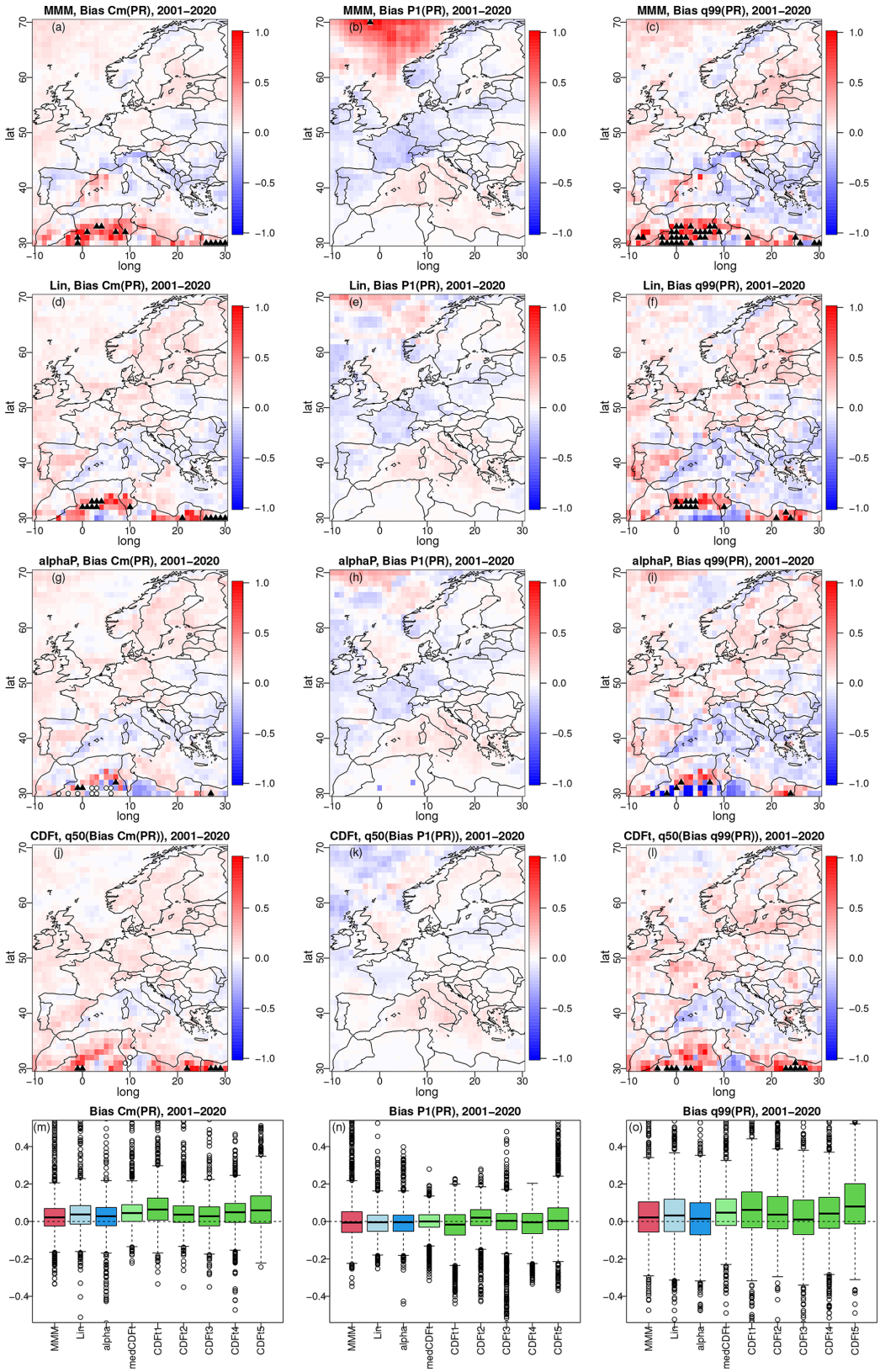

Figure 6. Same as Fig. 5 but for winter precipitation with biases in conditional mean given a wet state (left column: **a, d, g, j, m**), probability of a dry day (P_1, middle column: **b, e, h, k, n**), and 99 % quantile (right column: **c, f, i, l, o**). The results for summer are given in Fig. S8.

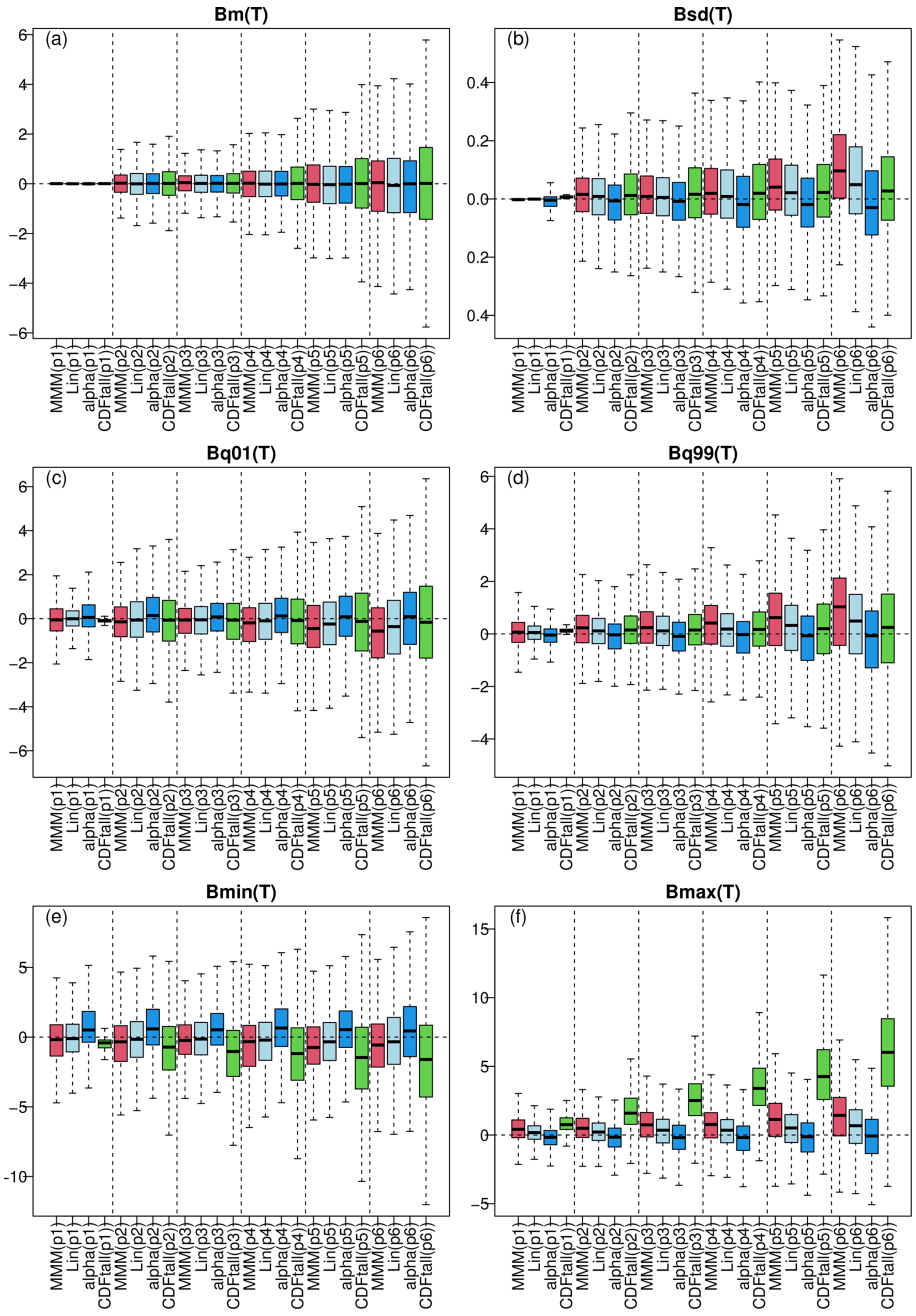

Figure 7. Results of the perfect model experiment for winter temperature: box plots of biases from the three methods (red: MMM, light blue: linear pooling, blue: α pooling, green: CDFt) for the six 20-year time periods (from p1 for 1981–2000 calibration to p6 for 2081–2100). The different panels display biases in **(a)** mean temperature, **(b)** standard deviation, **(c)** 1 % quantile, **(d)** 99 % quantile, and **(e)** minimum and **(f)** maximum temperature. Note that, for CDF-t, the box plots are drawn from the concatenation of all the individual CDF-t biases. Results for summer are provided in Fig. S9.

itation given a wet state (Cm, Fig. 8a) and $P-1$ (Fig. 8b), but more differences are visible for all other types of bias in favour of α pooling. Indeed, for precipitation standard deviation(Fig. 8c), condition 99 % quantile (CQ99, Fig. 8d), unconditional 99 % quantile (Q99, Fig. 8e), and maximum value (Fig. 8f), the α-pooling biases (blue box plots) are always more centred around 0 and with a smaller variability than the linear pooling and MMM biases.

The results from this PME allow us to conclude that the proposed α-pooling method is robust in a climate change context for both temperature and precipitation. In addition, it also indicates that a bias correction technique based on

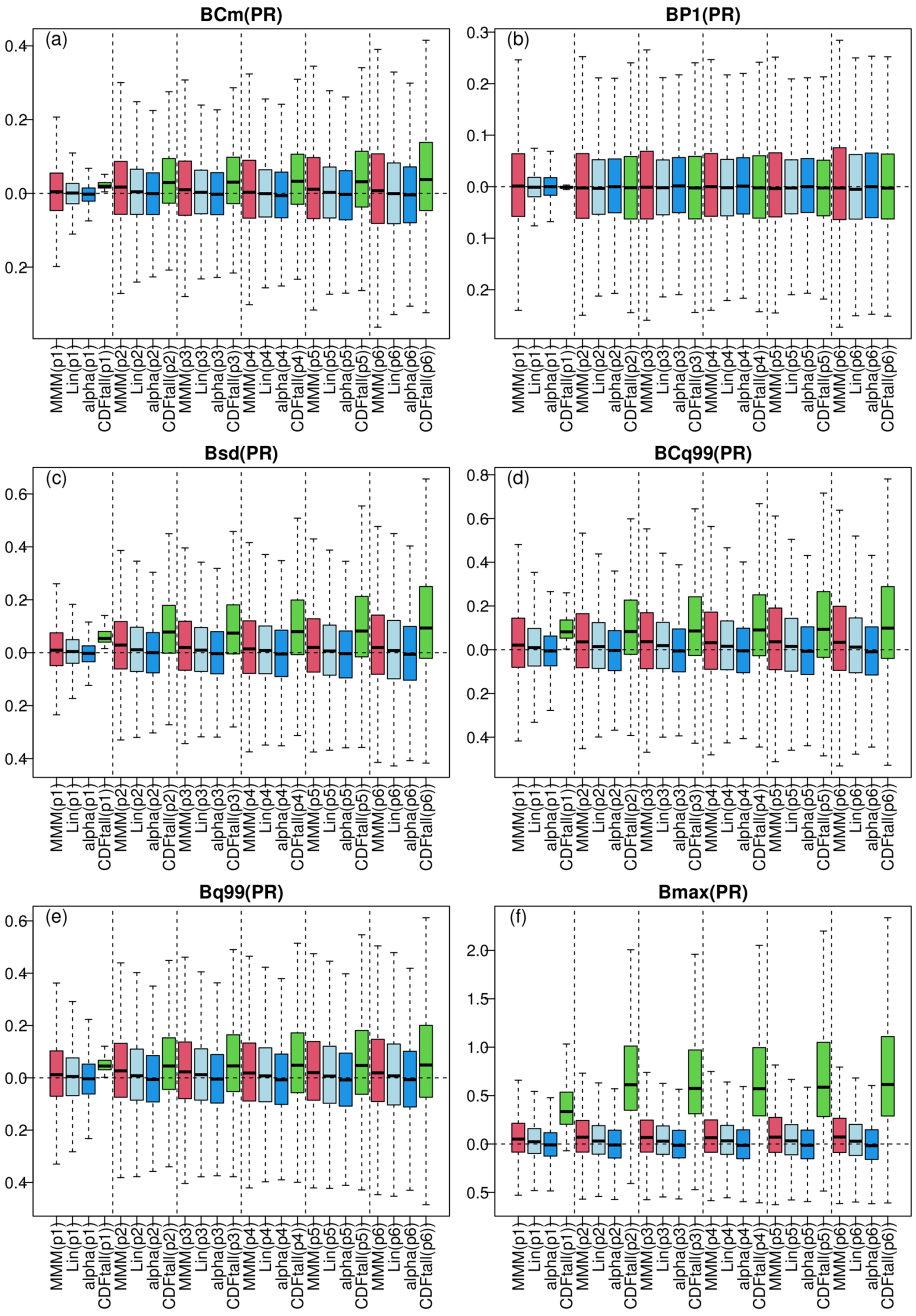

Figure 8. Results of the perfect model experiment for winter precipitation: same as Fig. 7 but for precipitation. The different panels display biases in **(a)** conditional mean precipitation given a wet state, **(b)** probability of a dry (< 1 mm) day, **(c)** standard deviation, **(d)** conditional 99 % quantile given wet conditions, **(e)** unconditional 99 % quantile, and **(f)** maximum precipitation. Results for summer are provided in Fig. S10.

an MMM (i.e. averaging) or linear combination of the GCM CDFs can be useful and robust, although the best results are achieved by the α-pooling technique.

5.3 Sensitivity experiment results

The conclusions brought by the perfect model experiment are based on the pooling and bias correction of five climate models, somewhat arbitrarily selected. One can wonder about the uncertainty or sensitivity of the resulting projected (i.e. future) CDFs of T and PR if other climate models were se-

lected. This is the reason why we perform the sensitivity experiment detailed in Sect. 4.3.

For each of the six selected cities over 2081–2100, Fig. 9 shows the 75 % confidence envelope of the 100 winter temperature CDFs obtained from MMM (red lines), α pooling (blue lines), and linear pooling (light blue lines), as well as the 75 % envelope from the 12 CDF-t results (green lines). Figure 10 show the 75 % confidence envelopes for winter precipitation CDFs. Summer CDF results are given in Figs. S11 and S12 in the Supplement.

All temperature corrections show a shift of the CDFs towards higher values for all six cities. All combination approaches (i.e. MMM, linear, and α pooling) have very similar 75 % envelopes for Paris (Fig. 9a) and are relatively close for Berlin (Fig. 9e) and Stockholm (Fig. 9f). The other cities present some more differences. The three combination-based methods show similar lower bounds for London but with a higher upper bound for the linear pooling and α-pooling techniques (depending on the quantiles). Rome and Madrid have an MMM envelope shifted towards lower temperature with respect to the other methods. CDF-t 75 % envelopes are generally larger and thus comprise most of the envelopes for any of the six cities. For precipitation (Fig. 10), as expected, the future projections – and thus their corrections – show varying trends depending on the cities. The combination-based methods give 75 % CDF envelopes showing more rain in Paris, London, Berlin, and, to some extent, Stockholm (Fig. 10a, b, e, and f), while they result in less rain in Rome (Fig. 10c). Madrid (Fig. 10d) appears to be the most uncertain for linear and α pooling – whose CDF envelope contains the ERA5 precipitation CDF – while MMM shows more frequent low to medium rain but less frequent heavy rain. For most cities, CDF-t envelopes tend to have lower bounds showing a potential negative shift of the precipitation CDFs with respect to ERA5.

In addition to the position of these envelopes, their size is also important. Hence, the widths of the 75 % CDF confidence envelopes for the six cities over 2081–2100 in winter are given in Fig. 11 for temperature and Fig. 12 for precipitation. For temperature, it is clear that CDF-t has, by far, the largest envelopes widths, while MMM generally has the smallest ones. It was somewhat expected that linear pooling and α pooling would have larger uncertainty than MMM. Indeed, the use of weights means that models with higher weights will have a stronger influence on the resulting CDFs and bias corrections. Thus, even if these models do not closely align with reality during the projection period, their influence can lead to combined projections that can significantly deviate from the simple average performed by MMM. However, there is no such a systematic conclusion for precipitation, showing much more variable rankings, depending on the cities and on the probability values.

Globally, the combination-based bias correction methods (MMM, linear, and α pooling) show some robustness in their application to future projections, with uncertainties and sensitivities to the chosen models not being much different from those of the more usual CDF-t technique for precipitation and being even smaller for temperature.

6 Conclusions and perspectives

In this study, we propose a new approach to perform bias correction of climate simulations, taking advantage of combinations of climate models. Combinations are realized via mathematical pooling of cumulative distribution functions (CDFs) – characterizing the variable of interest as simulated by the climate models – to provide a new CDF designed to be more realistic, i.e. closer to a reference CDF over the calibration period. It is important to emphasize that the proposed approach differs from the averaging of quantiles for a given probability as in Markiewicz et al. (2020). It also differs from the usual probability density aggregation, also sometimes called probability fusion (Koliander et al., 2022). Indeed, in our approach, we aggregate cumulative probability distributions. Moreover, our aggregation is indirect in that we aggregate transformed scores instead of directly aggregating the probabilities. In the latter case, we would be restricted to weights summing to 1, whereas in our approach there is no such restriction.

Three pooling strategies have been tested: a CDF multi-model mean (MMM), a linear pooling, and a new approach named α pooling that allows more flexibility, as well as a more traditional bias correction method (CDF-t) applied separately model by model. These four methods have been compared with three different experiments relying on (i) an evaluation with respect to ERA5 reanalyses over a historical period, (ii) a perfect model experiment (PME) over future time periods, and (iii) a sensitivity analysis to the choice of the climate models to combine.

In a cross-validation framework over the historical period (experiment (i), Sect. 5.1), the four methods generally behave similarly, with most biases relatively well centred around 0 in both temperature and precipitation. However, the application of the "pure" bias correction method CDF-t to separate GCMs can generate more biases with more variability. This is because the change (in temperature or precipitation) simulated by a single climate model over the historical period may not correspond to the change present in the reanalyses. By combining CDFs coming from different GCMs, the pooling techniques also combine the evolution (i.e. changes) over time, resulting in bias-corrected projections that are more consistent with the reanalyses.

The results of the PME show a good robustness of the three pooling strategies, even for the MMM approach, with biases of most statistics (including extremes) around 0. Moreover, the biases in high quantiles, especially for maximum values, are much lower for pooling-based methods than for traditional BC methods represented here by CDF-t. Overall, a quasi-systematic ranking of the four methods is observed

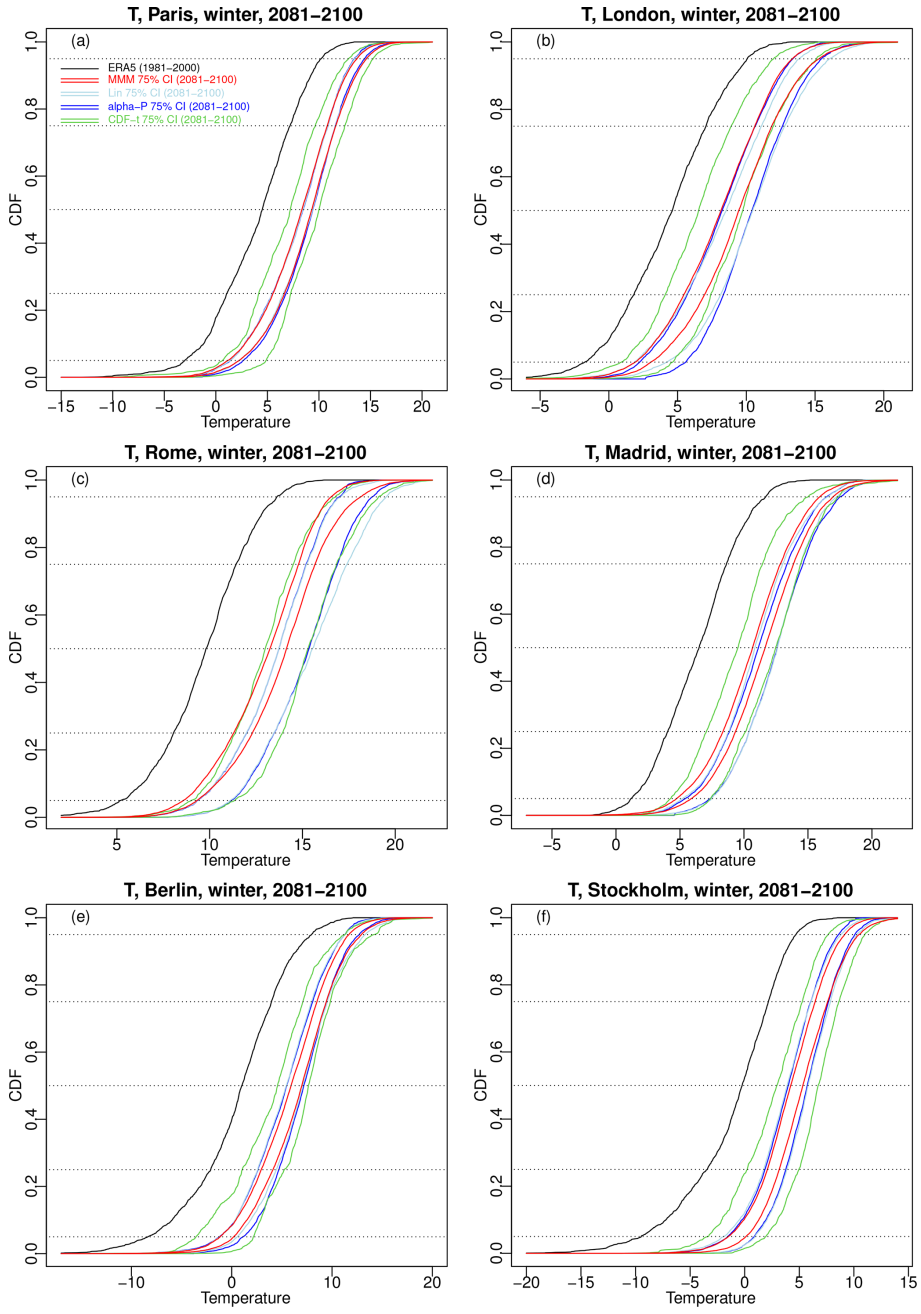

Figure 9. Results of the sensitivity experiment: for winter temperature over 2081–2100 and six major cities in western Europe, 75 % confidence intervals for α pooling (blue lines), linear pooling (light blue lines), MMM (red lines), and CDFt (green lines). The temperature ERA5 CDF (black line) over 1981–2000 is also displayed for visual evaluation of changes. Results for summer are provided in Fig. S11.

in this PME: while CDF-t can present some recurrent and pronounced biases – getting larger for further time periods – the MMM correction approach improves the results; the linear approach improves the results even more, and the best results are obtained with the α-pooling technique for both variables. This confirms the benefits of combining the information (here CDFs) from different models to perform bias correction, even in a strong climate change context. This is in agreement with results from Vrac et al. (2022), who showed, in a slightly different context, that accounting for the evolution of the mean temperature–precipitation correlation in an ensemble of climate models allows getting more robust estimates of future dependencies.

However, the CDFs resulting from our linear or α-pooling approaches might depend on the selected ensemble of model CDFs to combine. Hence, the choice of the models to com-

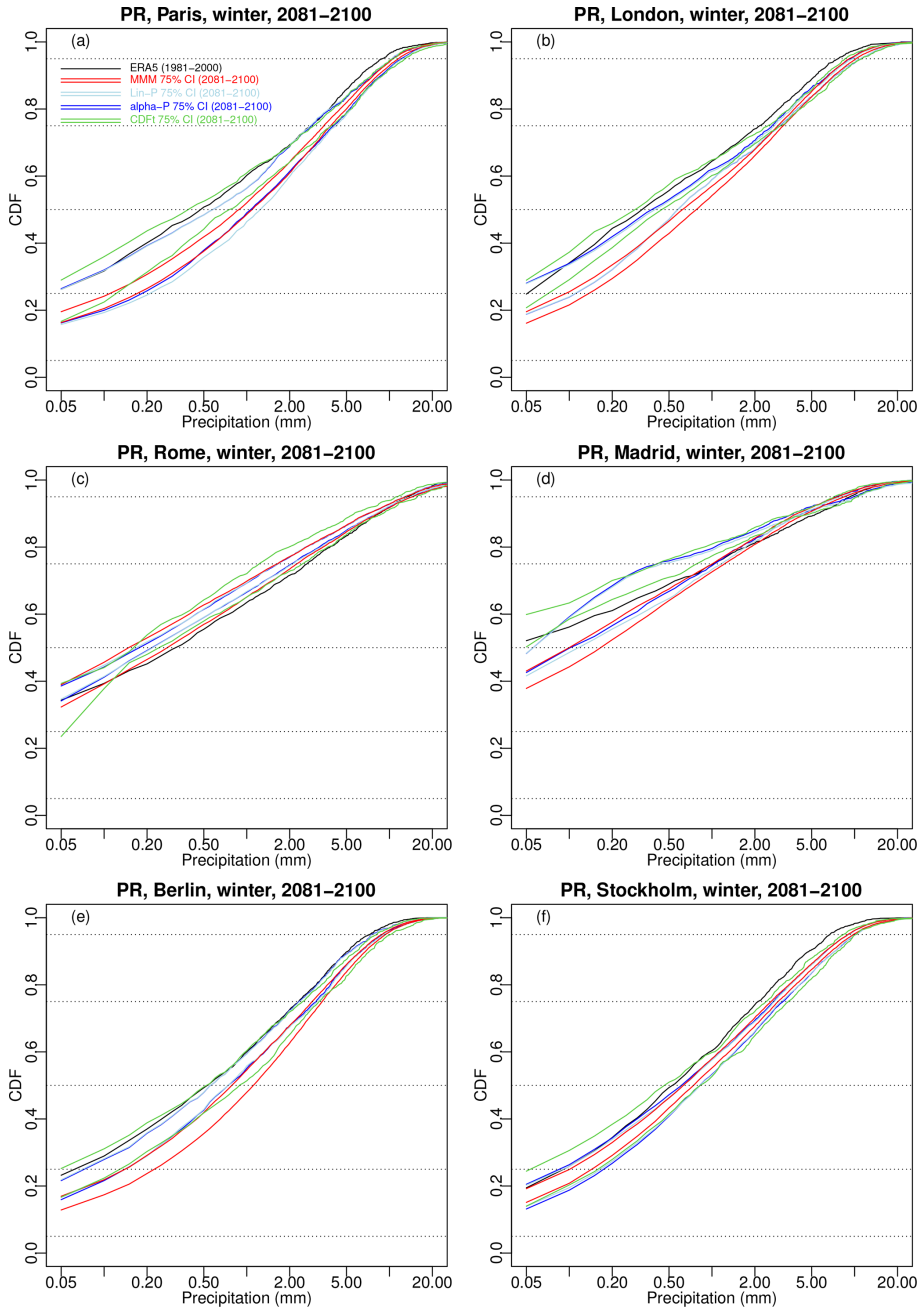

Figure 10. Results of the sensitivity experiment: same as Fig. 9 but for precipitation. Note that the x axis is displayed in log scale to ease evaluation. Results for summer are provided in Fig. S12.

bine remains key as it necessarily influences the results over the (future) projection periods. Note, nevertheless, that this is true for any combination strategy – i.e. not only our proposed pooling methods – or for any bias correction technique where the choice of the model simulation to correct will also necessarily affect the final results (e.g. time series, CDFs). We also note here that for the combination methods that include weights (i.e. linear and α pooling), the numerical optimization of the weights results in redundant CDFs receiving low weights. For instance, if two models result in the exact same CDF, the optimization will result in weights that will be shared between these identical CDFs and whose total would be the weight corresponding to this CDF not being duplicated. This is an important feature as it is known that some models are closely related and, thus, tend to provide similar forecasts.

The sensitivity analysis of the future (2081–2100) CDFs to the choice of the ensemble of models shows that the un-

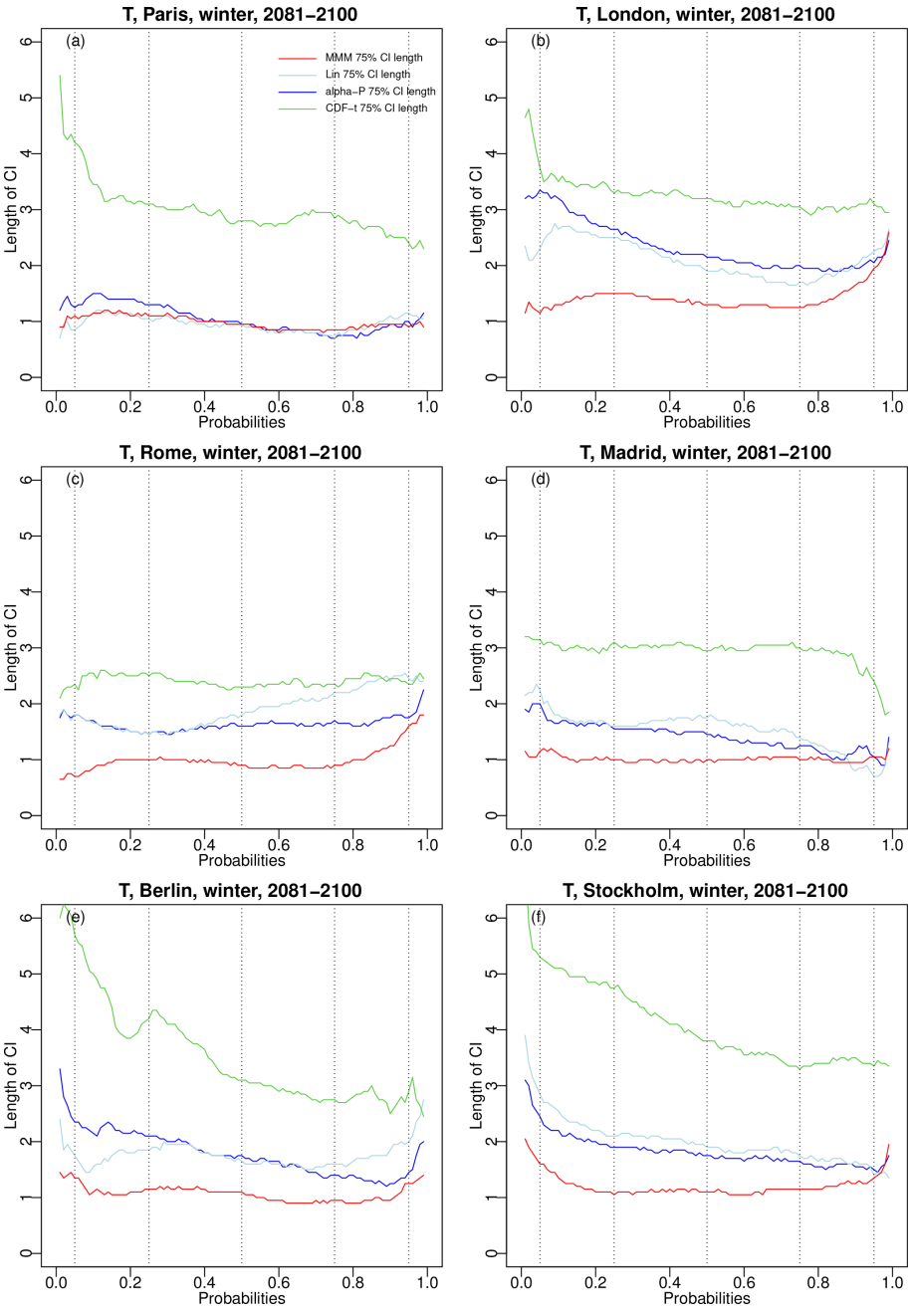

Figure 11. For winter temperature over 2081–2100 and six major cities in western Europe, width of the 75 % CDF confidence intervals for MMM (red line), linear pooling (light blue line), α pooling (blue line), and CDFt (green line). Results for summer are provided in Fig. S13 in the Supplement.

certainty in long-term projections was found to be globally comparable for the three pooling-based methods, although it is slightly higher for α pooling and slightly lower for MMM pooling. Indeed, as the α pooling and linear pooling associate non-uniform weights with the different CDFs, they pull the results towards the models with the highest weights, hence generating more variability depending on the selected ensemble of models to combine. Conversely, the MMM pooling corresponds to a linear pooling with weights forced to be uniform. Therefore, it provides smoother CDF results that are less sensitive to the choice of the ensemble. The opposite example is given by CDF-t that is applied model by model and thus shows a high sensitivity to the selected ensemble. While MMM pooling has the potential to lead to overly confident projections, our novel pooling method may offer a more realistic representation of scenario uncertainty.

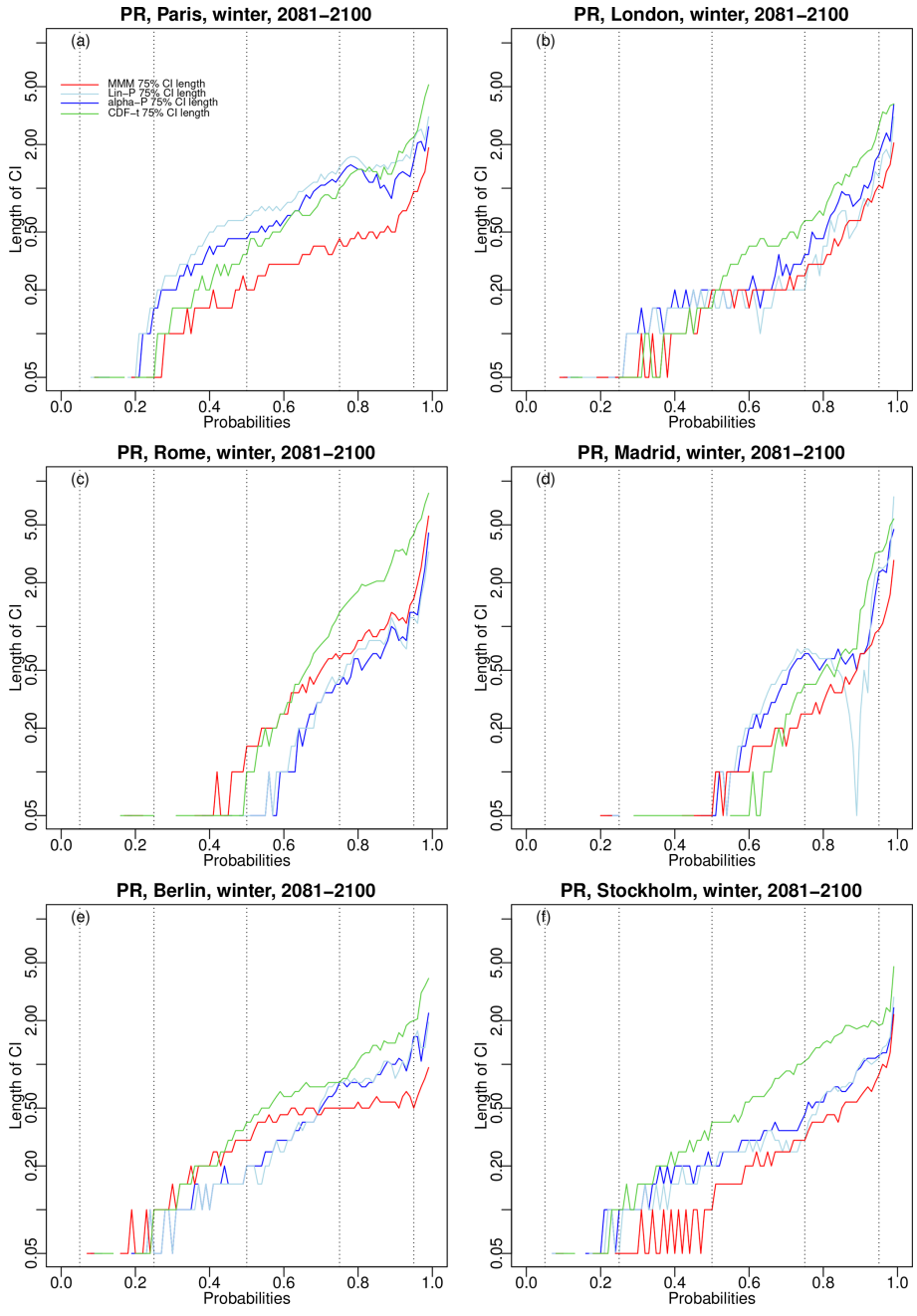

Figure 12. Same as Fig. 11 but for precipitation. Note that the y axis is displayed in log scale to ease evaluation. Results for summer are provided in Fig. S14 in the Supplement.

Nevertheless, it is crucial to acknowledge the potential for our α-pooling method to introduce unrealistic scenario uncertainty. This aspect warrants further investigation in future studies, especially for practical applications.

In terms of computation time, it is obvious that alpha pooling is more computationally demanding than linear or MMM pooling. This is in part due to the additional parameter α but mostly to the non-linearity induced by α pooling. However, for the combination of up to 10 climate models (i.e. CDFs), the computational time for each location and variable time series typically does not exceed a few minutes. Given the substantial computational demands associated with running individual climate models, the computational aspect of combining them is trivial by comparison. Moreover, considering that this post-processing of climate simulations does not need to be performed on a daily basis but rather once for all, we believe that this represents a reasonable computational cost,

ensuring the method's practical applicability without compromise.

As a conclusion, the α-pooling model appears to be a promising approach for pooling model CDFs. More generally, the results of this study show that the CDF pooling strategy for "multi-model bias correction" is a credible alternative to usual GCM-by-GCM correction methods by allowing handling and considering several climate models at once.

This work can be extended in various ways. First, even though only temperature and precipitation were considered in this study, many other climate variables – such as wind and humidity – can be handled with this CDF pooling strategy. Also, the proposed pooling method can be directly applied to regional climate model simulations, instead of GCM simulations, in order to get more regional views of climate changes.

In addition, some more technical and statistical developments could be made to improve the CDF pooling approach. For example, the present linear pooling and α-pooling methods are based on the L^2 norm to estimate the parameters. Other distances could be used, more specifically distances between distribution functions, e.g. the Hellinger distance, the total variation distance (Clarotto et al., 2022), the Wasserstein distance (e.g., Santambrogio, 2015; Robin et al., 2019), or the Kullback–Leibler divergence (Kullback and Leibler, 1951). Such distribution-based distances could potentially improve the quality of the fit and then provide more robust pooled CDFs.

Moreover, even though spatial patterns are visible in the parameters, there is variability between nearby grid cells that complicates the interpretation of the parameters (see Figs. 3 and 4). Such variability can be reduced by constraining the approach to provide more continuous and smoother spatial structures, presumably at the cost of longer computations.

Also note that it would be interesting to account for rainfall specificities when applying a CDF pooling strategy to precipitation. Indeed, in this study, the pooling was applied to all daily precipitation values. In practice, a distinction between dry day frequencies and distributions of wet intensity could be made by having two separate poolings. Although the α-pooling results for precipitation in this article were quite satisfying, such a rainfall-specific design could provide additional improvements and should be tested in the future.

Other modelling extensions could be considered. One interesting aspect could be to focus on extreme events. For example, α pooling could be applied to conditional CDFs above a high threshold related to the tail of the whole distribution or applied to the CDF of block maxima. Distributions stemming from the extreme value theory – such as the generalized Pareto distribution (GPD) or the generalized extreme value distribution (GEV) – would then have to be used. Besides the practical results that such an application could bring, the statistical properties of the resulting pooled (extreme) CDFs would also be worth studying from a theoretical point of view.

Another interesting perspective, in terms of both practical and theoretical aspects, concerns the extension of the α pooling to the multivariate context. Indeed, so far, this pooling method has been developed and applied only in a univariate framework; i.e. different variables (temperature and precipitation) are handled, combined, and bias-corrected separately. An extension of α pooling allowing the combination of joint (i.e. multivariate) CDFs would allow improving the modelling of dependencies between the variables and, thus, providing more realistic inter-variable CDFs and bias-corrected projections. Such an extended α pooling should then be compared to other multivariate bias correction methods, such as those studied in François et al. (2020). It would then also allow investigating compound events (e.g. Zscheischler et al., 2018, 2020) and their potential future changes more robustly.

Finally, more generally, it is worth noting that combination and bias correction are not new questions or requirements. However, this is the first paper coupling methods from these two domains. This was made possible by our pooling strategy working on CDFs (and not on specific quantiles or statistical properties such as mean and max, as usually done), which is, in itself, an original contribution to the combination framework. This CDF pooling strategy and this hybrid combination–correction method deserve to be further explored, as do its potential applications beyond combination and bias correction.

Appendix A: An approximate solution to the α pooling

The well-known Box–Cox transformation $B(F)(x) = (1 - F(x)^\alpha)/\alpha$, with $\alpha > 0$, is well defined for all values $F(x) \in [0, 1]$, with $\lim_{\alpha\to 0} B(F)(x) = -\ln F(x)$ when $F(x) > 0$ and $\lim_{\alpha\to 0} B(1-F)(x) = -\ln(1-F(x))$ when $F(x) < 1$. Let us consider a pooling approach that consists of assuming that the Box–Cox transformation of the pooled CDF is, up to a normalizing factor K, the weighted average of the Box–Cox transformation, i.e.

$$B(K.F_B)(x) = \sum_{i=1}^{N} w_i B(F_i)(x) \text{ and}$$
$$B(K.(1-F_B))(x) = \sum_{i=1}^{N} w_i B(1-F_i)(x).$$

After multiplying by α and rearranging, one gets

$$K^\alpha F_B(x)^\alpha = 1 + \sum_{i=1}^{N} w_i \left(F_i(x)^\alpha - 1\right) \text{ and}$$
$$K^\alpha (1 - F_B(x))^\alpha = 1 + \sum_{i=1}^{N} w_i \left(1 - F_i(x)\right)^\alpha - 1\big).$$

From the fact that $F_B(x) + 1 - F_B(x) = 1$, one thus gets

$$F_B(x) = \frac{\left[1 - S + \sum_{i=1}^{N} w_i F_i(x)^\alpha\right]^{1/\alpha}}{\left\{\begin{array}{c}\left[1 - S + \sum_{i=1}^{N} w_i F_i(x)^\alpha\right]^{1/\alpha} + \\ \left[1 - S + \sum_{i=1}^{N} w_i (1 - F_i(x))^\alpha\right]^{1/\alpha}\end{array}\right\}}, \quad \forall x \in \mathbb{R}, \quad \text{(A1)}$$

with $S = \sum_{i=1}^{N} w_i$.

Let us now go back to the α-pooling approach described in Sect. 3.4. Inspired by Eq. (A1), let us plug into the α pooling in Eq. (7) a solution of the form $F_H(x)^\alpha = (\sum_{i=1}^{N} w_i F_i(x)^\alpha + A)/Z$ and $(1 - F_H(x))^\alpha = (\sum_{i=1}^{N} w_i (1 - F_i(x))^\alpha + A)/Z$, where Z is a normalizing factor. From $F_H(x) + 1 - F_H(x) = 1$ we find that $Z^{1/\alpha} = \left[\sum_{i=1}^{N} w_i F_i(x)^\alpha + A\right]^{1/\alpha} + \left[\sum_{i=1}^{N} w_i (1 - F_i(x))^\alpha + A\right]^{1/\alpha}$ and

$$F_H(x) = \frac{\left[\sum_{i=1}^{N} w_i F_i(x)^\alpha + A\right]^{1/\alpha}}{\left\{ \begin{array}{c} \left[\sum_{i=1}^{N} w_i F_i(x)^\alpha + A\right]^{1/\alpha} + \\ \left[\sum_{i=1}^{N} w_i (1 - F_i(x))^\alpha + A\right]^{1/\alpha} \end{array} \right\}},$$

which is nothing but Eq. (A1) with $A = 1 - S$. Hence $F_H = F_B$, and for the rest of this section, we will use the notation F_B for both constructions. F_B is well defined for all $\alpha > 0$ if $S \leq 1$. In this case, it can be shown that it is a non-decreasing function of x because its derivative with respect to x is non-negative. From

$$\begin{aligned} \lim_{x \to -\infty} F_B(x) &= \frac{(1 - S)^{1/\alpha}}{(1 - S)^{1/\alpha} + 1} \quad \text{and} \\ \lim_{x \to \infty} F_B(x) &= \frac{1}{(1 - S)^{1/\alpha} + 1}, \end{aligned} \tag{A2}$$

one finds that F_B in Eq. (A1) is a proper CDF if and only if the condition $S = 1$ is verified. In this case, F_B has the simpler expression

$$F_{B,1}(x) = \frac{\left[\sum_{i=1}^{N} w_i F_i(x)^\alpha\right]^{1/\alpha}}{\left\{ \begin{array}{c} \left[\sum_{i=1}^{N} w_i F_i(x)^\alpha\right]^{1/\alpha} + \\ \left[\sum_{i=1}^{N} w_i (1 - F_i(x))^\alpha\right]^{1/\alpha} \end{array} \right\}}. \tag{A3}$$

When $\alpha = 1$, the pooling formula (Eq. A3) reduces to the linear pooling. As $\alpha \to 0$, it is straightforward to check that it boils down to the log-linear pooling (Eq. 4). As was the case for the α pooling presented in Sect. 3.4, this pooling formula thus generalizes both the log-linear pooling and the linear pooling. It must be emphasized that replacing w_i by $K w_i$ with $K > 0$ in Eq. (A3) leads to the same value $F_{B,1}(x)$. Imposing $\sum_{i=1}^{N} w_i = 1$ or not in Eq. (A3) thus has no consequences for $F_{B,1}$.

The existence of two different pooling approaches, namely F_G and F_B, calls for some comments.

- In numerous tests, it was consistently found that the CDF F_G obtained by the α pooling (Eq. 7) and the CDF $F_{B,1}$ computed directly using Eq. (A3) are almost indistinguishable when imposing $S = 1$. In this case, the direct computation in Eq. (A3) is 5 to 10 times faster and should be preferred.

- However, as discussed in Sect. 3.4, F_G is a proper CDF even if $S > 1$. There is thus an extra parameter available for the α-pooling approach, allowing for a better fit between the models and the reference. The cost is increased computation time.

- When using the direct approach in Eq. (A1), $S \leq 1$ leads to well-defined values $F_B(x)$. It thus also offers an extra parameter for the pooling, but the CDF F_B varies between the limits in Eq. (A2) instead of $[0, 1]$. Strictly speaking, F_B is thus not a proper CDF. In practice, however, it was very often found that the quantity $((1 - S)^{1/\alpha} + 1)^{-1}$ was extremely small (say, less than 10^{-3}) and the min–max rescaling shown in Eq. (11) can be performed to get a proper CDF.

- In Eq. (A1) $S > 1$ must be avoided as it can lead to inconsistent results, such as non monotonic functions F_B.

Appendix B: Optimal properties of α pooling

We briefly report some optimal properties of the α pooling presented in Sect. 3.4. We refer to Neyman and Roughgarden (2023) for a complete presentation on proper scoring rules, quasi-arithmetic pooling, and min–max optimal properties. We first start with some generalities. For the sake of clarity, x is fixed and we write F for $F(x)$. We further define the vector $\boldsymbol{F} = (F, 1 - F)^t$. In what follows, vectors will be written in bold letters.

The accuracy of a pooling method for a probability distribution is assessed using a metric, called a scoring rule, which assigns a value (sometimes called a reward) when a probability $\boldsymbol{q}$ is reported and outcome j happens according to a reference probability $\boldsymbol{p}$. Among all possible scoring rules, we will restrict ourselves to *proper scoring rules*, i.e. a scoring rule that is maximized when the reported probability is $\boldsymbol{q} = \boldsymbol{p}$. Well-known examples of proper scoring rules are the Brier scoring rule (Brier et al., 1950) and the logarithmic scoring rule. As shown in Gneiting and Raftery (2007) and in Neyman and Roughgarden (2023, Theorem 3.1), proper scoring rules can be derived from a function $G(\boldsymbol{p})$, referred to as the *expected reward function*. According to this theorem a scoring rule is proper if and only if

$$s(\boldsymbol{p}; j) = G(\boldsymbol{p}) + \langle \boldsymbol{g}(\boldsymbol{p}), \delta_j - \boldsymbol{p} \rangle, \tag{B1}$$

where $\boldsymbol{g}(\boldsymbol{p})$ is the gradient of $G(\boldsymbol{p})$. Let $j = 1, \ldots, J$ be the possible outcomes with probabilities $\boldsymbol{p} = (p(1), \ldots, p(J))$. The Brier (also known as "quadratic") scoring rule corresponds to $G_{\text{Brier}}(\boldsymbol{p}) = \sum_j p(j)^2$ and the logarithmic scoring rule corresponds to $G_{\log}(\boldsymbol{p}) = \sum_j p(j) \ln p(j)$. A necessary condition on G is that it is a convex function with respect to $\boldsymbol{p}$. Neyman and Roughgarden (2023) call *quasi-arithmetic pooling* any pooling formula defined by

$$\begin{aligned} \boldsymbol{g}(\boldsymbol{F}_G) &= \sum_{i=1}^{N} w_i \boldsymbol{g}(\boldsymbol{F}_i), \\ w_i &\geq 0, \ i = 1, \ldots, N, \quad \sum_{i=1}^{N} w_i = 1, \end{aligned} \tag{B2}$$

where g is the gradient of a proper scoring rule G. They show (in Theorem 4.1) the following max–min property for quasi-arithmetic pooling formula. Let us define the following utility function,

$$u(\boldsymbol{F}; j) := s(\boldsymbol{F}; j) - \sum_{i=1}^{N} w_i s(\boldsymbol{F}_i; j) \tag{B3}$$

which corresponds to the expected difference between the scoring rule applied to $\boldsymbol{F}$ and the scoring rule applied to model i, chosen randomly according to $\boldsymbol{w}$. Then, the minimum $\min_F u(\boldsymbol{F}; j)$ is maximized by setting $\boldsymbol{F} = \boldsymbol{F}_G$ as given in Eq. (B2). In other words, the worst loss of scores (often interpreted as a reward) is maximized using quasi-arithmetic pooling.

In our case, for a given x, there are only two possible outcomes, $j \in \{0, 1\}$: being less than or equal to x, with probability $p(0) = F$, and being above s, with probability $p(1) = 1 - F$. We now consider the following convex function:

$$G(\boldsymbol{F}) = F^{1+\alpha} + (1-F)^{1+\alpha}, \tag{B4}$$

with the limit case $\lim_{\alpha \to 0} G(\boldsymbol{F}) = F \ln F + (1-F)\ln(1-F)$ corresponding to the logarithmic scoring rule. Notice that $\alpha = 1$ corresponds to the Brier scoring rule. The associated gradient is

$$\boldsymbol{g}(\boldsymbol{F}) = (1+\alpha)(F^{\alpha},\ (1-F)^{\alpha})^t, \tag{B5}$$

with $\lim_{\alpha \to 0} \boldsymbol{g}(\boldsymbol{F}) = (1 + \ln F,\ 1 + \ln(1-F))^t$. Since in Eq. (B4) the function G is convex, the scoring rule given by Eq. (B1) is proper and each component of the gradient is a continuous and injective function of F for all values $\alpha \geq 0$. The scoring rule associated with $G(\boldsymbol{F})$ in Eq. (B4) thus varies continuously from the logarithmic scoring rule to the Brier scoring rule as α varies from 0 to 1. Notice that α is also allowed to be larger than 1, but the scoring rule has no specific name in that case. The pooling formula defined by

$$\begin{aligned} \boldsymbol{H}_2 \boldsymbol{g}(\boldsymbol{F}_G) &= \textstyle\sum_{i=1}^{N} w_i \boldsymbol{H}_2 \boldsymbol{g}(\boldsymbol{F}_i), \\ w_i \geq 0,\ i = 1, \ldots, N, &\quad \textstyle\sum_{i=1}^{N} w_i = 1, \end{aligned} \tag{B6}$$

where $\boldsymbol{H}_2$ is the $(1, 2)$ Helmert matrix, corresponds exactly to the α pooling presented in Sect. 3.4, which thus inherits the optimal properties of quasi-arithmetic pooling.

Code and data availability. The CMIP6 model simulations can be downloaded through the Earth System Grid Federation portals. Instructions to access the data are available here: https://pcmdi.llnl.gov/mips/cmip6/data-access-getting-started.html (PCMDI, 2019). The ERA5 reanalysis data used as a reference in this study can be accessed via the Climate Data Store (CDS) web portal at https://cds.climate.copernicus.eu (Hersbach et al., 2017).

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/esd-15-735-2024-supplement.

Author contributions. MV and GM had the initial idea of the study, which has been completed and enriched by all co-authors. DA developed the initial mathematical framework and derived the main theoretical properties, helped by MV, ST, and GM. MV and DA developed the codes for inferring the α-pooling parameters. MV applied it to CMIP6 simulations for the different experiments and wrote the codes for the analyses and to plot the figures. All authors contributed to the methodology and the analyses. MV wrote the first draft of the article with inputs from all the co-authors.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modelling groups (listed in Table 1 of this paper) for producing and making available their model outputs. For CMIP, the US Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. We also thank the Copernicus Climate Change Services for making the ERA5 reanalyses available.

Financial support. This work has been supported by the COMBINE project funded by the Swiss National Science Foundation (grant no. 200021_200337/1), as well as by the COESION project funded by the French National programme LEFE (Les Enveloppes Fluides et l'Environnement). Mathieu Vrac's work also benefited from state aid managed by the National Research Agency under France 2030 bearing the references ANR-22-EXTR-0005 (TRACCS-PC4-EXTENDING project) and ANR-22EXTR-0011 (TRACCS-PC10-LOCALISING project). The authors also acknowledge the support of the INRAE/Mines Paris chair "Geolearning".

Review statement. This paper was edited by Olivia Martius and reviewed by two anonymous referees.

References

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., and Schmidt, G. A.: ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing, Earth Syst. Dynam., 10, 91–105, https://doi.org/10.5194/esd-10-91-2019, 2019.

Ahmed, K., Sachindra, D. A., Shahid, S., Demirel, M. C., and Chung, E.-S.: Selection of multi-model ensemble of general circulation models for the simulation of precipitation and maximum and minimum temperature based on spatial assessment metrics, Hydrol. Earth Syst. Sci., 23, 4803–4824, https://doi.org/10.5194/hess-23-4803-2019, 2019.

Allard, D., Comunian, A., and Renard, P.: Probability aggregation methods in geoscience, Math. Geosci., 44, 545–581, https://doi.org/10.1007/s11004-012-9396-3, 2012.

Arias, P., Bellouin, N., Coppola, E., Jones, R., Krinner, G., Marotzke, J., Naik, V., Palmer, M., Plattner, G.-K., Rogelj, J., Rojas, M., Sillmann, J., Storelvmo, T., Thorne, P., Trewin, B., Achuta Rao, K., Adhikary, B., Allan, R., Armour, K., Bala, G., Barimalala, R., Berger, S., Canadell, J., Cassou, C., Cherchi, A., Collins, W., Collins, W., Connors, S., Corti, S., Cruz, F., Dentener, F., Dereczynski, C., Di Luca, A., Diongue Niang, A., Doblas-Reyes, F., Dosio, A., Douville, H., Engelbrecht, F., Eyring, V., Fischer, E., Forster, P., Fox-Kemper, B., Fuglestvedt, J., Fyfe, J., Gillett, N., Goldfarb, L., Gorodetskaya, I., Gutierrez, J., Hamdi, R., Hawkins, E., Hewitt, H., Hope, P., Islam, A., Jones, C., Kaufman, D., Kopp, R., Kosaka, Y., Kossin, J., Krakovska, S., Lee, J.-Y., Li, J., Mauritsen, T., Maycock, T., Meinshausen, M., Min, S.-K., Monteiro, P., Ngo-Duc, T., Otto, F., Pinto, I., Pirani, A., Raghavan, K., Ranasinghe, R., Ruane, A., Ruiz, L., Sallée, J.-B., Samset, B., Sathyendranath, S., Seneviratne, S., Sörensson, A., Szopa, S., Takayabu, I., Tréguier, A.-M., van den Hurk, B., Vautard, R., von Schuckmann, K., Zaehle, S., Zhang, X., and Zickfeld, K.: Technical Summary, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J., Maycock, T., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., Cambridge University Press, 33–144, https://doi.org/10.1017/9781009157896.002, 2021.

Bhat, K. S., Haran, M., Terando, A., and Keller, K.: Climate Projections Using Bayesian Model Averaging and Space–Time Dependence, J. Agr. Biol. Environ. Stat., 16, 606–628, https://doi.org/10.1007/s13253-011-0069-3, 2011.

Bordley, R.: A multiplicative formula for aggregating probability assessments, Manage. Sci., 28, 1137–1148, 1982.

Boucher, O., Denvil, S., Levavasseur, G., Cozic, A., Caubel, A., Foujols, M.-A., Meurdesoif, Y., Cadule, P., Devilliers, M., Ghattas, J., Lebas, N., Lurton, T., Mellul, L., Musat, I., Mignot, J., and Cheruy, F.: IPSL IPSL-CM6A-LR model output prepared for CMIP6 CMIP, https://doi.org/10.22033/ESGF/CMIP6.1534, 2018.

Brier, G. W.: Verification of forecasts expressed in terms of probability, Mon. Weather Rev., 78, 1–3, 1950.

Brunner, L., McSweeney, C., Ballinger, A. P., Befort, D. J., Benassi, M., Booth, B., Coppola, E., de Vries, H., Harris, G., Hegerl, G. C., Knutti, R., Lenderink, G., Lowe, J., Nogherotto, R., O'Reilly, C., Qasmi, S., Ribes, A., Stocchi, P., and Undorf, S.: Comparing Methods to Constrain Future European Climate Projections Using a Consistent Framework, J. Climate, 33, 8671–8692, https://doi.org/10.1175/JCLI-D-19-0953.1, 2020.

Bukovsky, M., Thompson, J., and Mearns, L. O.: Weighting a regional climate model ensemble: Does it make a difference? Can it make a difference?, 77, 23–43, https://doi.org/10.3354/cr01541, 2019.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C.: A Limited Memory Algorithm for Bound Constrained Optimization, SIAM J. Sci. Comput., 16, 1190–1208, https://doi.org/10.1137/0916069, 1995.

Cannon, A. J., Sobie, S. R., and Murdock, T. Q.: Bias Correction of GCM Precipitation by Quantile Mapping: How Well Do Methods Preserve Changes in Quantiles and Extremes?, J. Climate, 28, 6938–6959, https://doi.org/10.1175/JCLI-D-14-00754.1, 2015.

Clarotto, L., Allard, D., and Menafoglio, A.: A new class of α-transformations for the spatial analysis of Compositional Data, Spat. Stat.-Neth., 47, 100570, https://doi.org/10.1016/j.spasta.2021.100570, 2022.

PCMDI: PCMDI – CMIP6, PCMDI [code], https://pcmdi.llnl.gov/CMIP6/ (last access: 30 May 2024), 2019.

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., Emmons, L. K., Fasullo, J., Garcia, R., Gettelman, A., Hannay, C., Holland, M. M., Large, W. G., Lauritzen, P. H., Lawrence, D. M., Lenaerts, J. T. M., Lindsay, K., Lipscomb, W. H., Mills, M. J., Neale, R., Oleson, K. W., Otto-Bliesner, B., Phillips, A. S., Sacks, W., Tilmes, S., van Kampenhout, L., Vertenstein, M., Bertini, A., Dennis, J., Deser, C., Fischer, C., Fox-Kemper, B., Kay, J. E., Kinnison, D., Kushner, P. J., Larson, V. E., Long, M. C., Mickelson, S., Moore, J. K., Nienhouse, E., Polvani, L., Rasch, P. J., and Strand, W. G.: The Community Earth System Model Version 2 (CESM2), J. Adv. Model. Earth Sy., 12, e2019MS001916, https://doi.org/10.1029/2019MS001916, 2020.

de Elía, R., Laprise, R., and Denis, B.: Forecasting Skill Limits of Nested, Limited-Area Models: A Perfect-Model Approach, Mon. Weather Rev., 130, 2006–2023, https://doi.org/10.1175/1520-0493(2002)130<2006:FSLONL>2.0.CO;2, 2002.

Dembélé, M., Ceperley, N., Zwart, S. J., Salvadore, E., Mariethoz, G., and Schaefli, B.: Potential of satellite and reanalysis evaporation datasets for hydrological modelling under various model calibration strategies, Adv. Water Res., 143, 103667, https://doi.org/10.1016/j.advwatres.2020.103667, 2020.

Déqué, M.: Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values, Global Planet. Change, 57, 16–26, 2007.

Eden, J., Widmann, M., Grawe, D., and Rast, S.: Skill, Correction, and Downscaling of GCM-Simulated Precipitation, J. Climate, 25, 3970–3984, https://doi.org/10.1175/JCLI-D-11-00254.1, 2012.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G.,

Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: Complete ERA5 from 1940: Fifth generation of ECMWF atmospheric reanalyses of the global climate, Copernicus Climate Change Service (C3S) Data Store (CDS) [data set], https://doi.org/10.24381/cds.143582cf, 2017.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: 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, 2016.

Fragoso, T. M., Bertoli, W., and Louzada, F.: Bayesian model averaging: A systematic review and conceptual classification, Int. Stat. Rev., 86, 1–28, 2018.

François, B., Vrac, M., Cannon, A. J., Robin, Y., and Allard, D.: Multivariate bias corrections of climate simulations: which benefits for which losses?, Earth Syst. Dynam., 11, 537–562, https://doi.org/10.5194/esd-11-537-2020, 2020.

François, B., Thao, S., and Vrac, M.: Adjusting spatial dependence of climate model outputs with Cycle-Consistent Adversarial Networks, Clim. Dynam., 57, 3323–3353, https://doi.org/10.1007/s00382-021-05869-8, 2021.

Gneiting, T. and Katzfuss, M.: Probabilistic Forecasting, Annu. Rev. Stat. Appl., 1, 125–151, https://doi.org/10.1146/annurev-statistics-062713-085831, 2014.

Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, J. Am. Stat. Assoc., 102, 359–378, 2007.

Gudmundsson, L., Bremnes, J. B., Haugen, J. E., and Engen-Skaugen, T.: Technical Note: Downscaling RCM precipitation to the station scale using statistical transformations – a comparison of methods, Hydrol. Earth Syst. Sci., 16, 3383–3390, https://doi.org/10.5194/hess-16-3383-2012, 2012.

Haddad, Z. and Rosenfeld, D.: Optimality of empirical z-r relations, Q. J. Roy. Meteor. Soc., 123, 1283–1293, 1997.

Held, I. M., Guo, H., Adcroft, A., Dunne, J. P., Horowitz, L. W., Krasting, J., Shevliakova, E., Winton, M., Zhao, M., Bushuk, M., Wittenberg, A. T., Wyman, B., Xiang, B., Zhang, R., Anderson, W., Balaji, V., Donner, L., Dunne, K., Durachta, J., Gauthier, P. P. G., Ginoux, P., Golaz, J.-C., Griffies, S. M., Hallberg, R., Harris, L., Harrison, M., Hurlin, W., John, J., Lin, P., Lin, S.-J., Malyshev, S., Menzel, R., Milly, P. C. D., Ming, Y., Naik, V., Paynter, D., Paulot, F., Ramaswamy, V., Reichl, B., Robinson, T., Rosati, A., Seman, C., Silvers, L. G., Underwood, S., and Zadeh, N.: Structure and Performance of GFDL's CM4.0 Climate Model, J. Adv. Model. Earth Sy., 11, 3691–3727, https://doi.org/10.1029/2019MS001829, 2019.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Q. J. Roy. Meteor. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

IPCC: Evaluation of Climate Models, in: Climate Change 2013 – The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Intergovernmental Panel on Climate Change (IPCC), Cambridge University Press, 741–866, https://doi.org/10.1017/CBO9781107415324.020, 2014.

IPCC: Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, 1st edn., Cambridge University Press, https://doi.org/10.1017/9781009157896, 2023.

IPCC WGI: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., Cambridge University Press, https://doi.org/10.1017/9781009157896, 2021.

Kleiber, W., Raftery, A. E., and Gneiting, T.: Geostatistical Model Averaging for Locally Calibrated Probabilistic Quantitative Precipitation Forecasting, J. Am. Stat. Assoc., 106, 1291–1303, https://doi.org/10.1198/jasa.2011.ap10433, 2011.

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., and Eyring, V.: A climate model projection weighting scheme accounting for performance and interdependence: Model Projection Weighting Scheme, Geopys. Res. Lett., 44, 1909–1918, https://doi.org/10.1002/2016GL072012, 2017.

Koliander, G., El-Laham, Y., Djurić, P. M., and Hlawatsch, F.: Fusion of probability density functions, P. IEEE, 110, 404–453, 2022.

Krinner, G. and Flanner, M. G.: Striking stationarity of large-scale climate model bias patterns under strong climate change, P. Natl. Acad. Sci. USA, 115, 9462–9466, https://doi.org/10.1073/pnas.1807912115, 2018.

Kullback, S. and Leibler, R. A.: On information and sufficiency, Ann. Math. Stat., 22, 79–86, 1951.

Kwatra, V., Schödl, A., Essa, I., Turk, G., and Bobick, A.: Graphcut Textures: Image and Video Synthesis Using Graph Cuts, ACM T. Graphic., 22, 277–286, https://doi.org/10.1145/882262.882264, 2003.

Lange, S.: ISIMIP3b bias adjustment fact sheet, Technical report, ISIMIP, https://www.isimip.org/documents/413/ISIMIP3b_bias_adjustment_fact_sheet_Gnsz7CO.pdf (last access: 30 May 2024), 2021.

Lange, S. and Büchner, M.: ISIMIP3b bias-adjusted atmospheric climate input data, https://doi.org/10.48364/ISIMIP.842396.1, 2021.

Markiewicz, I., Bogdanowicz, E., and Kochanek, K.: Quantile Mixture and Probability Mixture Models in a Multi-Model Approach to Flood Frequency Analysis, Water, 12, 2851, https://doi.org/10.3390/w12102851, 2020.

Michelangeli, P., Vrac, M., and Loukos, H.: Probabilistic downscaling approaches: application to wind cumulative distribution functions, Geophys. Res. Lett., 36, L11708, https://doi.org/10.1029/2009GL038401, 2009.

Neyman, E. and Roughgarden, T.: From Proper Scoring Rules to Max-Min Optimal Forecast Aggregation, Oper. Res., arXiv:2102.07081, https://doi.org/10.48550/arXiv.2102.07081, 2023.

Olson, R., Fan, Y., and Evans, J. P.: A simple method for Bayesian model averaging of regional climate model projections: Application to southeast Australian temperatures, Geophys. Res. Lett., 43, 7661–7669, https://doi.org/10.1002/2016GL069704, 2016.

Panofsky, H. and Brier, G.: Some applications of statistics to meteorology, Earth and Mineral Sciences Continuing Education, College of Earth and Mineral Sciences, 103 pp., 1968.

Ribes, A., Zwiers, F. W., Azaïs, J.-M., and Naveau, P.: A new statistical approach to climate change detection and attribution, Clim. Dynam., 48, 367–386, 2017.

Robin, Y. and Vrac, M.: Is time a variable like the others in multivariate statistical downscaling and bias correction?, Earth Syst. Dynam., 12, 1253–1273, https://doi.org/10.5194/esd-12-1253-2021, 2021.

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

Rougier, J., Goldstein, M., and House, L.: Second-Order Exchangeability Analysis for Multimodel Ensembles, J. Am. Stat. Assoc., 108, 852–863, https://doi.org/10.1080/01621459.2013.802963, 2013.

Sain, S. and Cressie, N.: A spatial model for multivariate lattice data, J. Econ., 140, 226–259, https://doi.org/10.1016/j.jeconom.2006.09.010, 2007.

Sanderson, B. M., Wehner, M., and Knutti, R.: Skill and independence weighting for multi-model assessments, Geosci. Model Dev., 10, 2379–2395, https://doi.org/10.5194/gmd-10-2379-2017, 2017.

Santambrogio, F.: Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling, Progress in Nonlinear Differential Equations and Their Applications, 1st edn., Birkhäuser, Cham, https://doi.org/10.1007/978-3-319-20828-2, 2015.

Schmidli, J., Frei, C., and Vidale, P.: Downscaling from GCM precipitation: a benchmark for dynamical and statistical downscaling methods, Int. J. Climatol., 26, 679–689, https://doi.org/10.1002/joc.1287, 2006.

Shiogama, H., Abe, M., and Tatebe, H.: MIROC MIROC6 model output prepared for CMIP6 ScenarioMIP, https://doi.org/10.22033/ESGF/CMIP6.898, 2019.

Stott, P.: How climate change affects extreme weather events, Science, 352, 1517–1518, https://doi.org/10.1126/science.aaf7271, 2016.

Strobach, E. and Bel, G.: Learning algorithms allow for improved reliability and accuracy of global mean surface temperature projections, Nat. Commun., 11, 451, https://doi.org/10.1038/s41467-020-14342-9, 2020.

Swart, N. C., Cole, J. N., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., Anstey, J., Arora, V., Christian, J. R., Jiao, Y., Lee, W. G., Majaess, F., Saenko, O. A., Seiler, C., Seinen, C., Shao, A., Solheim, L., von Salzen, K., Yang, D., Winter, B., and Sigmond, M.: CCCma CanESM5 model output prepared for CMIP6 ScenarioMIP, https://doi.org/10.22033/ESGF/CMIP6.1317, 2019.

Tang, Y., Rumbold, S., Ellis, R., Kelley, D., Mulcahy, J., Sellar, A., Walton, J., and Jones, C.: MOHC UKESM1.0-LL model output prepared for CMIP6 CMIP historical, https://doi.org/10.22033/ESGF/CMIP6.6113, 2019.

Thao, S., Garvik, M., Mariéthoz, G., and Vrac, M.: Combining Global Climate Models Using Graph Cuts, Clim. Dynam., 59, 2345–2361, https://doi.org/10.1007/s00382-022-06213-4, 2022.

Thorarinsdottir, T. L. and Gneiting, T.: Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression, J. Roy. Stat. Soc. Ser. A, 173, 371–388, https://doi.org/10.1111/j.1467-985X.2009.00616.x, 2010.

Voldoire, A.: CNRM-CERFACS CNRM-CM6-1-HR model output prepared for CMIP6 HighResMIP, https://doi.org/10.22033/ESGF/CMIP6.1387, 2019.

Volodin, E., Mortikov, E., Gritsun, A., Lykossov, V., Galin, V., Diansky, N., Gusev, A., Kostrykin, S., Iakovlev, N., Shestakova, A., and Emelina, S.: INM INM-CM5-0 model output prepared for CMIP6 CMIP abrupt-4xCO2, https://doi.org/10.22033/ESGF/CMIP6.4932, 2019.

Vrac, M.: Multivariate bias adjustment of high-dimensional climate simulations: the Rank Resampling for Distributions and Dependences (R^2D^2) bias correction, Hydrol. Earth Syst. Sci., 22, 3175–3196, https://doi.org/10.5194/hess-22-3175-2018, 2018.

Vrac, M. and Thao, S.: R^2D^2 v2.0: accounting for temporal dependences in multivariate bias correction via analogue rank resampling, Geosci. Model Dev., 13, 5367–5387, https://doi.org/10.5194/gmd-13-5367-2020, 2020.

Vrac, M., Stein, M. L., Hayhoe, K., and Liang, X.-Z.: A general method for validating statistical downscaling methods under future climate change, Geophys. Res. Lett., 34, L18701, https://doi.org/10.1029/2007GL030295, 2007.

Vrac, M., Drobinski, P., Merlo, A., Herrmann, M., Lavaysse, C., Li, L., and Somot, S.: Dynamical and statistical downscaling of the French Mediterranean climate: uncertainty assessment, Nat. Hazards Earth Syst. Sci., 12, 2769–2784, https://doi.org/10.5194/nhess-12-2769-2012, 2012.

Vrac, M., Noël, T., and Vautard, R.: Bias correction of precipitation through Singularity Stochastic Removal: Because occurrencesmatter, J. Geophys. Res.-Atmos., 121, 5237–5258, https://doi.org/10.1002/2015JD024511, 2016.

Vrac, M., Thao, S., and Yiou, P.: Should Multivariate Bias Corrections of Climate Simulations Account for Changes of Rank Correlation Over Time?, J. Geophys. Res.-Atmos., 127, e2022JD036562, https://doi.org/10.1029/2022JD036562, 2022.

Wanders, N. and Wood, E. F.: Improved sub-seasonal meteorological forecast skill using weighted multi-model ensemble simulations, Environ. Res. Lett., 11, 094007, https://doi.org/10.1088/1748-9326/11/9/094007, 2016.

Weigel, A. P., Knutti, R., Liniger, M. A., and Appenzeller, C.: Risks of Model Weighting in Multimodel Climate Projections, J. Climate, 23, 4175–4191, https://doi.org/10.1175/2010JCLI3594.1, 2010.

Wu, T., Chu, M., Dong, M., Fang, Y., Jie, W., Li, J., Li, W., Liu, Q., Shi, X., Xin, X., Yan, J., Zhang, F., Zhang, J., Zhang, L., and Zhang, Y.: BCC BCC-CSM2MR model output prepared for CMIP6 CMIP piControl, https://doi.org/10.22033/ESGF/CMIP6.3016, 2018.

Wuebbles, D., Easterling, D., Hayhoe, K., Knutson, T., Kopp, R., Kossin, J., Kunkel, K., LeGrande, A., Mears, C., Sweet, W., Taylor, P., Vose, R., Wehner, M., Wuebbles, D., Fahey, D., Hibbard, K., Dokken, D., Stewart, B., and Maycock, T.: Chap. 1: Our Globally Changing Climate, in: Climate Science Special Report: Fourth National Climate Assessment, Vol. I, https://doi.org/10.7930/J08S4N35, 2017.

Xu, C.-Y.: From GCMs to river flow: a review of downscaling methods and hydrologic modelling approaches, Prog. Phys. Geog., 23, 229–249, https://doi.org/10.1177/030913339902300204, 1999.

Yukimoto, S., Koshiro, T., Kawai, H., Oshima, N., Yoshida, K., Urakawa, S., Tsujino, H., Deushi, M., Tanaka, T., Hosaka, M., Yoshimura, H., Shindo, E., Mizuta, R., Ishii, M., Obata, A., and Adachi, Y.: MRI MRI-ESM2.0 model output prepared for CMIP6 CMIP, https://doi.org/10.22033/ESGF/CMIP6.621, 2019.

Zscheischler, J., Westra, S., van den Hurk, B., Seneviratne, S., Ward, P., Pitman, A., AghaKouchak, A., Bresch, D., Leonard, M., Wahl, T., and Zhang, X.: Future climate risk from compound events, Nat. Clim. Change, 8, 469–477, https://doi.org/10.1038/s41558-018-0156-3, 2018.

Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Raymond, C., Horton, R., van den Hurk, B., AghaKouchak, A., Jézéquel, A., Mahecha, M., Maraun, D., Ramos, A., Ridder, N., Thiery, W., and Vignotto, E.: A typology of compound weather and climate events, Nat. Rev. Earth Environ., 1, 333–347, https://doi.org/10.1038/s43017-020-0060-z, 2020.