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

Mathieu Vrac, Denis Allard, Gregoire Mariethoz, Soulivanh Thao, Lucas Schmutz

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]INRAE, BioSP, 84914 Avignon, France
[3]Institute of Earth Surface Dynamics, University of Lausanne, Switzerland

Abstract. For the study of climate change, many General Circulation Models (GCM)s have been designed, modeling the climate on the planet Earth slightly differently either by emphasizing predictions in specific regions or by incorporating varied or uniquely modeled parameters. To extract a robust signal from the diverse outputs, models are typically combined into multimodel ensembles. Their results are summarized in various ways, including (possibly weighted) multimodel means, medians and other statistics, within a Bayesian framework or not. In this work, we introduce an new probability aggregation method termed "alpha-pooling" which builds an aggregated Cumulative Probability Function (CPF) designed to be closer to a reference CPF over the calibration period. α-pooling assigns a weight to each CPF, which is an increasing function of its closeness to the reference CPF. Key to the α-pooling is a parameter α that describes the type of aggregation, which includes linear aggregation and log-linear aggregation. We first establish that α-pooling is a proper aggregation method verifying some optimal properties. Then, focusing on climate models 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 perfect model experiment and a sensitivity analysis to the set of climate models are run. Our findings demonstrate the superiority of the proposed method, indicating that alpha-pooling presents a robust and efficient way to combine GCM's CPF. The results of this study show that the CDFs pooling strategy for "multi-model bias correction" is a credible alternative to usual GCM-by-GCM correction methods, by allowing to handle and consider several climate models at once.

1 Introduction

In recent years, many General Circulation Models (GCMs) have been designed, modeling the physical processes in the atmosphere, ocean, cryosphere and land surface of the planet Earth slightly differently either by emphasizing predictions in specific regions or by incorporating varied or uniquely modeled parameters (Eyring et al., 2016). To extract a robust signal from the diverse outputs, models are typically combined into multimodel ensembles (MMEs), and their results are synthesized into multimodel means (MMMs). This approach is grounded in the belief that members of the MMEs are "truth-centered". In other words, the various models act as independent samples from a distribution that gravitates towards the truth, and as the ensemble expands, the MMM is expected to approach the true average (Ribes et al., 2017).

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 individual model performance, nor their interdependencies. 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, ensuring they reflect both performance and interdependencies. However, climate models often share foundational assumptions, parameterizations, and codes, making their outputs interdependent (Abramowitz et al., 2019; Knutti et al., 2017; Rougier et al., 2013). This interdependence means that consensus among models does not necessarily result in a reliable projection. Furthermore, the global weighting of models 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 global regions (Ahmed et al., 2019; Dembélé et al., 2020). Yet, the potential for improved model combinations remains, especially if weights are optimized at the grid point level. Moreover, traditional model averaging techniques tend to homogenize the spatial patterns inherent in individual models, even though these patterns often stem from genuine physical processes. Approaches that consider per-grid point model combinations, as seen in meteorology, have shown promise in enhancing performance (Kleiber et al., 2011; Thorarinsdottir and Gneiting, 2010). 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 patchworking method, utilizing a graph cut technique from computer vision to combine climate model outputs. This approach aims to minimize biases and maintain local spatial dependencies, producing a cohesive "patchwork" of the most accurate models while preserving spatial consistency.

In this study, we introduce an innovative probability aggregation method termed "alpha-pooling". Grounded on the probabilistic approach to aggregating probabilities (Allard et al., 2012; Koliander et al., 2022), this technique assigns a weight to each probability distribution, proportional to its performance relative to the target distribution, and incorporates a regularization parameter, that we name alpha. As alpha approaches zero, the alpha-pooling converges to log-linear pooling, and as it nears

one, it aligns with linear pooling. The resultant probability distribution is precisely computed to be as close to a given reference as the ensemble permits.

Our application of the alpha-pooling method focuses on the 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 determine the weights during a calibration phase, utilizing a reference, and subsequently employ these weights to generate projections from the bias-corrected ensemble. This study compares the alpha-pooling method with prevalent BC techniques, including Multi-Model Mean (MMM), linear pooling, log-linear pooling, and CDF transformation (CDFt). Our analysis spans both short-term and extended projections of temperatures (T) and precipitation (PR), encompassed in two distinct experiments. In the first experiment, ERA5 serves as the reference, enabling performance evaluation against this benchmark. Subsequently, a perfect model experiment is employed, wherein each model is iteratively used as the reference. This iterative approach offers insights into the stability of the alpha-pooling projections compared to other BC techniques, extending to the end of the century. Our findings demonstrate the superiority of the proposed method, indicating that alpha-pooling presents a robust and efficient way to combine GCMs.

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, Section 3 presents the new α-pooling. Section 4 describes the experiments carried out in this work and Section 5 describes the obtained results. In Section 6 we provide some conclusions and perspectives. Two appendices provide an approximate, faster, solution to the α-pooling 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^o horizontal spatial resolution. The Western Europe domain, defined as $[10^oW, 30^oE] \times [30^oN, 70^oN]$, is considered.

Moreover, the same variables (T and PR) are also extracted for the period 1981–2100 from 12 Global Climate Models (GCMs) contributing to the 6^{th} exercise of the "Coupled Models Intercomparison Project" (CMIP6, Eyring et al., 2016). This selection was dictated by the availability of T and PR fields on daily time scales at the time of the analyses: we have only selected models whose data were fully available for the whole period 1981–2100. The list of the GCMs 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 $1^\circ \times 1^\circ$. Moreover, for 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 developed CDF 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 non-parametric CDF modelling methods can be used if needed and appropriate.

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

Table 1. List of CMIP6 simulations used in this study, their run, approximate horizontal atmospheric resolution and references. The models preceded by a "*" correspond to the 5 models used in the "ERA5 experiment" (sections 4.1 and 5.1) and the "Perfect Model Experiment" (sections 4.2 and 5.2). All 12 models are used in the "Sensitivity" experiment (sections 4.3 and 5.3). See text for details.

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 amounts thus essentially to combine, or aggregate, 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) offers 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. Those playing an important role in this work are briefly recalled. Interested readers are referred to Allard et al. (2012) for a more detailed exposition. Let us denote $p_1, \ldots, p_N$ the probabilities to be pooled together and p_G 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. Finally, let us suppose that we are in the following case: there exists at least one index i such that $p_i = 0$ (resp. $p_i = 1$) with $0 < p_j < 1$ for $j \neq i$. A pooling method which returns $p_G = 0$ (resp. $p_G = 1$) 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.

3.1 Pre-processing: Standardising data

CDFs from climate model simulations can be much different from each other or from ERA5 CDFs and it is then necessary to perform a preliminary standardization (i.e., basic adjustment) before pooling them. Note that it is basically what is performed in many IPCC figures (WGI, 2021) when working on anomalies (instead of raw simulated or reference data). This allows to more easily compare (and then combine) the different datasets. In the present study, temperature and precipitation are standardised differently. For temperature, the simulated data are rescaled such that the mean and standard deviation correspond to those of the reference data:

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

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:

$$PR_{rescaled} = PR \times Q90_{ref}/Q90_{mod} \tag{2}$$

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 compromise between a robust enough quantile and a large enough range of precipitation values (Vrac et al., 2016). In the rest of this paper, all tested pooling methods are then applied to standardized data.

Before presenting our new pooling approach, named α-pooling, we must first present briefly the linear and log-linear pooling with their main properties.

3.2 Linear pooling

The linear pooling, whose the resulting pooled CDF is denoted F_L, is simply a weighted average of all CDFs:

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

F_L is a proper CDF if and only if all w_is 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 to averaging the quantiles for a given probability, as done in a usual weighted MMM. Indeed, in our linear-pooling (3), the weighted average is performed on the CDFs (i.e., probabilities $F_i(x)$) and not on quantiles (values) of the variable.

3.3 Log-linear pooling

The log-linear pooled CDF, denoted 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:

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

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

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

which is a proper CDF for all non-negative weights w_i. The condition $S = \sum_{i=1}^{N} w_i = 1$ entails unanimity. On simulations, Allard et al. (2012) showed that log-linear pooling leads consistently 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 there exists a pair i, j with $i \neq j$ such that $F_i(x) = 0$ and $F_j(x) = 1$.

3.4 α-Pooling

In order to mitigate the problem faced with the log-linear pooling, we propose a new pooling method. Our approach builds on the $A_{\alpha-IT}$ 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, $A_{\alpha-IT}$ transforms a compositions 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 2-part composition. In this case, the $A_{\alpha-IT}$ transformation of $\boldsymbol{F}(x)$ results in a scalar:

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

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

$$z_G(x) = \sum_{i=1}^{N} w_i z_i(x) = \frac{\sqrt{2}}{\alpha} \sum_{i=1}^{N} w_i \big(F_i(x)^{\alpha} - (1 - F_i(x))^{\alpha}\big), \tag{6}$$

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

$$F_G(x)^{\alpha} - (1 - F_G(x))^{\alpha} = z_G(x) = \sum_{i=1}^{N} w_i \big(F_i(x)^{\alpha} - (1 - F_i(x))^{\alpha}\big). \tag{7}$$

Let us define the function $G(y) = y^{\alpha} - (1 - y)^{\alpha}$ with $0 \leq y \leq 1$. Then, one can write that $F_G(x) = G^{-1}(z_G(x))$, where G^{-1} is the inverse function of G. There is unfortunately no general closed form solution to (7) for all values of α. It is however straightforward to check that when $\alpha = 1$, the solution to (7) is the linear pooling provided that $\sum_{i=1}^{N} w_i = 1$. Likewise, using that $\lim_{\alpha \to 0} F_i(x)^{\alpha} = 1 + \alpha \ln F_i(x)$, it is easy to check that the α-pooling tends to the log-linear pooling as $\alpha \to 0$. In practice, the solution to (7) is found by minimizing $Q = \big(G(y) - z_G(x)\big)^2$ constrained to $y \in [0, 1]$ and by setting $F_G(x) = y_{min}$, where y_{min} is the location of the minimum of Q. No restrictions to the sum $\sum_{i=1}^{N} w_i = 1$ is necessary. We can show the following:

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

Proof: The derivative of $z_G(x)$ with respect to x is $z_G(x)' = \alpha \sum_{i=1}^{N} w_i f_i(x)\big(F_i(x)^{\alpha-1} + (1-F_i(x))^{\alpha-1}\big) \geq 0$. 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 non negative, the function $F_G(x) = G^{-1}(z_G(x))$ is non-decreasing because it is the composition of two non-decreasing functions. In addition, since $F_G(x)$ is constrained to belong to the interval $[0,1]$, it is a proper CDF. □

The α-pooling presented in (7) mitigates the principal inconvenient 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 accommodates seamlessly the case $F_i(x) = 0$ and $F_j(x) = 1$ with $i \neq j$.

In Appendix A, we present a closed-form expression which is a very good approximate solution to (7) in most cases, i.e. except when $S = \sum_{i=1}^{N} w_i > 1$. Then in appendix B, we present some optimal properties of the α-pooling presented above deriving from the fact that α-pooling belongs to 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. 1(a) for $N = 2$ Gaussian distributions F_1 and F_2 with means 2 and 4 respectively and with standard deviations 1 and 1.3 respectively. A Gaussian reference CDF is arbitrarily fixed with mean 2.5 and standard deviation 1.5. For this example, the estimated α-pooling parameters are $w_1 = 0.489$, $w_2 = 0.212$ and $\alpha = 0.108$. The higher value for w_1 than for w_2 indicates that the reference CDF is closer to F_1 than to the other, which was expected when looking at the 2 CDFs. Overall, the α-pooling is clearly able to approximate correctly the reference CDF (blue line in Fig. 1(a)), as the latter is almost perfectly recovered by the resulting CDF (red dashed line).

3.5 Estimating and interpreting the parameters

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

$$Q = \sum_{k=1}^{K} (x_k - x_{k-1})\big(F_0(x_k) - F_G(x_k)\big)^2, \tag{8}$$

where $F_G(x)$ is obtained by solving (7) and where $x_0, \ldots, x_K$ is an increasing sequence discretizing the real line. A usual optimisation procedure is launched to find the weights and the α parameter. The weights must be positive and they can be constrained to sum to 1 or not. When unconstrained, it was found that in most cases the sum S was close to one.

The weights are easily interpretable, since as rule, the higher the weight w_i, the closer F_i is to the reference F_0. The parameter α is also interpretable. In the sum $\sum_i w_i F_i(x)^\alpha$, the higher is the exponent α, the stronger is the influence of the highest value among all values $F_i(x)$. An exponent larger than one increases the importance of the highest value among the CDFs relatively to the other ones. Conversely, an exponent smaller than one tends to reduce the differences between the CDFs. For example, let us consider that for a certain value x, three CDFs provide the following probabilities: $(0.1, 0.2, 0.4)$. Then if $\alpha = 2$, the transformed probabilities become $(0.01, 0.04, 0.16)$. Higher values of α are thus found when the reference is close to the higher values, i.e. to the left of the models. The converse is not always true however. As $\alpha \to 0$, the alpha-pooling

tends to the log-linear pooling. Generally speaking, the differences between the log-probabilities are less pronounced, as can be seen in the example above where the log-probabilities are equal to $(-2.3, -1.6, -0.9)$. Small values of α are thus able to accommodate different situations of the reference with respect to the models.

3.6 Benchmarking α-pooling: CDF-Multi-Model Mean (MMM) and linear pooling

As a benchmark for evaluating the α-pooling approach, two CDFs pooling methods are also applied. The first one is the simplest and consists in defining a "mean" CDF based on the N CDFs to be combined. Let's take an example with $N = 2$ GCMs with respectively CDFs F_1 and F_2, say of temperature, for a given grid-cell. For any value x of temperature, the mean CDF value $F_{MMM}(x)$ corresponds to the average of $F_1(x)$ and $F_2(x)$. An example is given in Fig. 1(b) for the same two Gaussian distributions as used to illustrate the α-pooling method. 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 reanalyis. Hence, it is not surprising that α-pooling approximates better the reference CDF over the calibration period.

In addition, a second CDFs pooling method is applied for comparison, 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-poolingmethods, we can assess the potential added-value brought by the alpha parameter.

The same illustration as previously is now given in Fig. 1(c) for linear pooling. Based on this toy 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. However, one major objective of this study is also to evaluate how MMM, linear pooling and α-poolingbehave in a projection period where climate changes occurs. When driven only by model CDFs over a projection (future) period, are the three polling methods able to capture the changes in reference (temperature or precipitation) CDFs?

3.7 Bias corrections from CDF-pooling results

Based on the obtained CDFs $\hat{F}$ (from one the three pooling methods), like from any statistical distribution, it is possible to generate values. This can be done in an "independent and identically distributed" (iid) mode, where the n^{th} simulated value does not depend on the $(n-1)^{\text{th}}$ one. However, it can also be performed in a more constrained way such that the generated values are bias corrections of the climate model simulations, preserving the rank dynamics (i.e., the temporal dependence structure) of the raw GCM time series. Indeed, once $\hat{F}$ is estimated over a projection period, one can apply a quantile-mapping technique 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 in finding the value y such that $\hat{F}(y) = F_m(x)$ which is equivalent to:

$$y = \hat{F}^{-1}(F_m(x)) \tag{9}$$

where $\hat{F}^{-1}$ is the inverse CDF function, allowing to compute the quantile associated to a given probability. Therefore, by applying Eq. (9) successively to all simulations from model m, we obtain time series with the same rank chronology as that of model m but whose values are corrected to follow distribution $\hat{F}$. By applying this bias correction technique to the

different models employed within the MMM, linear 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 combining 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). Note that, here, no preliminary basic adjustment (i.e., standardisation) is made on the data, as the goal of CDF-t is precisely to perform bias corrections. It consists in a quantile-mapping technique (e.g., Panofsky and Brier, 1968; Haddad and Rosenfeld, 1997; Déqué, 2007; Gudmundsson et al., 2012) allowing to account 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}, respectively the reference CDF over the calibration period, the model CDF over the calibration period and the projection period:

$$\hat{F}_{Rp}(x) = F_{Rc}(F_{Mc}^{-1}(F_{Mp}(x))) \tag{10}$$

where F_{Mc}^{-1} is the inverse CDF function 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 $\hat{F}_{Rp}$ 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 different 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 as supplementary materials.

4.1 ERA5 experiment

The first experiment considers ERA5 reanalysis as reference. For linear and α-pooling methods, for each grid-point and variable, it consists in calibrating the approaches with N climate models with respect to ERA5 data over the calibration period 1981-2000 and, then, in using the parameters to combine the models CDFs over the projection period 2001-2020. For CDF-t, in the same manner the calibration period is 1981-2000. The corrections are made for each model independently for 2001-2020. For MMM, the CDFs of the climate models are directly averaged over 2001-2020. Results are next compared to data from ERA5 over 2001-2020.

In this experiment, only 5 GCMs are used. This is partly constrained by the α-pooling method that can have stability issues to infer the parameters for a too high number of models to combine. Although it has been tested with more than 10 models, 5 GCMs appeared as a good compromise between a reasonable computation time, stable parameter estimates and a sufficient

number of simulations to get robust results. These 5 GCMs (indicated with "*" in table 1) were selected on the basis of a preliminary analysis showing that they approximately represent the spread of future evolutions of all 12 GCMs (not shown). Note that 4 models (IPSL-CM6A-LR, MRI-ESM2-0, UKESM1-0-LL, GFDL-CM4) out of the 5 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 results with respect to ERA5. For each grid-point, dataset, variable, and season (winter, DJF or summer, JJA), some statistics S are calculated from the daily values time series. For temperature, statistics include the mean, standard deviation and 99% quantile (Q99), while they include the conditional mean given wet (Cm), probability of dry day (P_1) and the 99% quantile for precipitation. Here, a day with PR value lower than 1mm is considered as dry (and thus >1mm as wet).

Then, absolute biases are calculated as

$$B(m,S) = S(m) - S(ERA5) \tag{11}$$

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

$$B(m,S) = \frac{S(m) - S(ERA5)}{S(ERA5)} \tag{12}$$

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

4.2 Perfect Model Experiment (PME)

As the ERA5 experiment evaluates the methods on a projection period (2001-2020) very close to the calibration one (1981-2000), it does not allow to understand 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; Krinner and Flanner, 2018; Robin and Vrac, 2021; Thao et al., 2022; Vrac 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., weights and α for α-pooling or weights 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; 2081-2100.
- The same approach is followed for CDF-t: one model serves as reference over 1981-2000 to calibrate CDF-t – here, separately for each of the $N-1$ remaining models – that is next used to bias-correct each model simulation over the five future periods.
- As previously for the ERA5 experiment, MMM does not require any calibration. The 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 of some statistics of the results can then be evaluated with respect to the reference model. For temperature, it includes: absolute biases (Eq. (11)) of mean temperature, 1% quantile, 99% quantile, minimum temperature and maximum temperature, as well as relative biases (Eq. (12)) of standard deviation. For precipitation, relative biases are computed for conditional mean precipitation given wet, probability of dry ($<$ 1mm) day, standard deviation, conditional 99% quantile given wet, unconditional 99% quantile, and maximum precipitation.

Hence, no observational or reanalysis data are used as 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 5 models as in the ERA5 experiment (Section 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 multimodel 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 much 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 5 models (as in ERA5 experiment). Instead, 100 different sets of N=5 models among the 12 presented in table 1 are randomly drawn. The obtained 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 5 models to combine, while CDF-t is applied to the 12 models separately. The 2081-2100 results obtained from each method and model or set of models do not allow us any evaluation per se, as there is no reference over the future period. However, the use of multiple models or sets of models permits to quantify and compare the statistical uncertainty of each method brought by the choice of models. For sake of clarity, in this experiment, for both temperature and precipitation, only 6 grid-points are considered, corresponding to major capitals of the geographical domain: Paris (France), London (UK), Rome (Italy), Madrid (Spain), Berlin (Germany), 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, Figure 2 displays the maps of the winter (2.a and c) and summer (2.b and d) α parameters, for temperature (2.a and b) and precipitation (2.c and d). First, note that the range of α is not the same for T and PR. While most of the values are lower than 1 (no unit) for temperature, the range goes up to 2.5 for precipitation. Moreover, for the two seasons, we can see more pronounced spatial structures from T than from PR, the latter α maps appearing more "pixelated". This was somehow expected from the well-known spatially heterogeneous nature of rain properties. However, globally, even for PR, large regions share similar α values, indicating some spatial consistency of the parameters.

Regarding the weights parameters of α-pooling, their winter maps are provided in Figures 3 and 4, for temperature and precipitation respectively. The spatial structures of the weights are clearly visible (for both T and PR) and even more pronounced than from the α maps. This strongly indicates that α-pooling identifies large zones where some models have a bigger influence than others on the combination and, thus, whose CDFs are closer to the ERA5 ones. Note however that, for the 2 variables, none of the models has the highest weights for all grid-points of the domain. In other words, globally, over this European region, each of the 5 models brings valuable contributions, depending on the sub-region. For example in temperature, UKESM (panel 3.e) shows the strongest contributions over the Mediterranean sea, while MRI-ESM2 (panel 3.d) displays the largest weights over the Noth-East 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 contributions in terms of T and those in terms of PR, confirming that results from one variable cannot be generalised to another. In addition, a concentration index is also displayed in panels (f) of Figures 3 and 4. It is equal to the sum of the squares of the 5 normalized weights. This index allows to see if one single GCM takes all the weight – in this case, the concentration index has a value 1 – or if the 5 normalized weights are equally distributed – and in this case, the its value reaches the minimum value euqal to $1/N = 0.2$. The concentration index can only be applied to weights summing to one. In our implementation of α-pooling, the sum of weights is let free and, thus, not constrained to one. Although this sum remains quite close to 1 (mostly between 0.95 and 1.05 for temperature and between 0.92 and 1.1 for precipitation, not shown), it is therefore needed to normalize them by dividing each one by $S = \sum_{i=1}^{N} w_i$. For temperature, panel 3.f) 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 panel 3.e). For precipitation, more zones show the concentration index close to one: for examples, Northwestern part of the domain and North of France (MRI-ESM2, panel 4.f), South of Norway and Northeastern part of the domain (CNRM-CM6, 4.a), or Eastern Adriatic coast (UKESM1, 3.e). 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 modify structurally the interpretation of the weights but brings additional flexibility.

The biases of the different methods with respect to 2001-2020 ERA5 data in terms of mean, standard deviation and Q99 are now given for winter temperature in Fig. 5 and in terms of conditional mean given wet, probability of dry day (P_1) and Q99 for winter precipitation in Fig. 6. 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 fourth row for linear-pooling. Note that, because CDF-t is applied separately for each GCM, for sake of clarity, 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 boxplots of biases.

For temperature (Fig. 5), at first sight, the differences between the maps of biases from the four methods are not very pronounced. The biases have similar patterns and intensities from one method to another. This is specially true for the biases in mean temperature and standard deviation (sd). Some more differences appear for Q99. For example MMM (panel 5.c) shows relatively high positive bias ($\sim 4^oC$) over the Northeastern part of the domain (Sweden and Finland), while α-pooling Q99 biases (panel 5.f) CDF-t (median) ones (panel 5.i) and linear pooling ones (panel 5.l) do not present this structure. Also, CDF-t median Q99 (panel 5.i) have a positive ($\sim 1-2^oC$) bias pattern over the central domain (Germany, Italy, Poland, Hungary, Romania, etc.) while the three other methods show more nuanced and mixed structures. When looking at the more integrated boxplots view (bottom row in Fig. 5), the similar behaviour of α-pooling, linear-pooling and MMM is visible for the three biases: the boxplots are relatively equivalent from one method to another. However, if 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 a 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 (Figure 6), conclusions are about the same, although some more differences between the methods are present. For example, on the Norwegian sea, P_1 relative biases of MMM (Fig. 6.b) has 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 (6.f) 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 boxplots view for winter precipitation is similar to that for temperature: roughly equivalent boxplots for the four methods, with more variability from the individual CDF-t results.

Note, however, that the ERA5 experiment results for summer (Figures SM-7 and SM-8 of the supplementary material) show more differences between the four methods – specially in the boxplots –, slightly in favour of the linear an α-pooling methods, which show boxplots more centered around 0 for all biases and variables.

5.2 PME results

In the ERA5 experiment, evaluation (2001-2020) and calibration (1981-2000) periods are quite close to each other. This leads to similar results for the four methods. Hence, the "Perfect Model Experiment" (PME) described in section 4.2 is applied.

PME is first applied to winter temperature. For each period and method, the boxplots of the different biases are provided in Figure 7 (PME summer temperature results in Fig. SM-9). As expected, for all biases, the more distant the period, the larger the boxplots, indicating an increase in possible statistical errors for periods further in the future. For brevity, We now focus on the

last (i.e. p6) 2081-2100 period, which contains the most pronounced distinctions between the methods. For mean T bias (7.a), the four BC methods have a similar quality, even though CDF-t has a larger boxplot. The bias in minimum temperature (7.e) 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 temperature 1% quantile (Q01, 7.c), with CDF-t having a median bias (i.e. boxplot center) equivalent to α-pooling but with a larger variability. For maximum temperature (7.f), CDF-t shows a strongly positive bias, while its biases look reasonable – at least more comparable to the other methods – for standard deviation (7.b) and 99% quantile (Q99, 7.d). Globally, for temperature standard deviation (7.b), Q99 (7.d) and maximum value (7.f), 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 in Fig. SM-10). As for temperature, the more distant the period, the larger the boxplots precipitation biases, although this feature is less pronounced here. Over the 2081-2100 period, CDF-t results are often the most biased, except for the probability of dry day (P_1, 8.b) where it is as good as the two methods. As in Fig. 7.f, the maximum values of precipitation from CDF-t (green boxplot in Figure 8.f) show strong biases with a high variability. Regarding MMM, linear and α-pooling methods, they give about similar biases in terms of conditional mean precipitation given wet (Cm, 8.a) and $P-1$ (8.b) but more differences are visible for all other types of bias in favour of α-pooling. Indeed, for precipitation standard deviation(8.c), condition 99% quantile (CQ99, 8.d), unconditional 99% quantile (Q99, 8.e) and maximum value (8.f), the α-pooling biases (blue boxplots) are always more centred around 0 and with a smaller variability than the linear pooling and MMM biases.

The results from this PME experiment allows 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 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 5 climate models, somehow 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 had been selected. Hence, the sensitivity experiment detailed in section 4.3 is performed.

For each of the 6 selected cities over 2081-2100, Figure 9 shows the 75% confidence envelop 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% envelop from the 12 CDF-t results (green lines). Figure 10 show the 75% confidence envelops for winter precipitation CDFs.

All temperature corrections show a shift of the CDFs towards higher values, for all 6 cities. All combination approaches (i.e., MMM, linear- and α-pooling) have very similar 75% envelops for Paris (9.a) and relatively close for Berlin (9.e) and

Stockholm (9.f). 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- and α-pooling techniques (depending on the quantiles); Rome and Madrid have an MMM envelop shifted towards smaller temperature with respect to the other methods. CDF-t 75% envelops are generally larger and, thus comprise most of the envelops for any of the 6 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 envelops showing more rain in Paris, London, Berlin and, to some extent, Stockholm (10.a, b, e, f), while they give less rain in Rome (10.c). Madrid (10.d) appears as the most uncertain for linear and α-pooling – whose the CDF envelop contains the ERA5 precipitation CDF –, while MMM shows more frequent low to medium rain but less frequent heavy rain. For most cities, CDF-t envelops 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 envelops, an important information is their size. Hence, the lengths of the 75% CDF confidence envelops for the 6 cities over 2081-2100 in winter are given in Figure 11 for temperature and Figure 12 for precipitation. For temperature, it is clear that CDF-t has, by far, the largest envelops lengths, while MMM has generally the smallest ones. It was somehow expected that the linear- and α-pooling have a larger uncertainty than MMM. Indeed, since they associate weights to the models to combine, models with higher weights will have a stronger influence on the resulting CDFs and bias corrections. Thus, the calibrated combined projections will also be more influenced by these models and can, hence, be deviated 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 proposed a new approach to perform bias correction of climate simulations, taking advantage of combinations of climate models. Combinations are realised via mathematical pooling of cumulative distribution functions (CDF) – characterising 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. 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. The comparison of these four methods applied to temperature and precipitation has been performed according to 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), section 5.1), the four methods generally behave similarly, with most biases relatively well centered around 0, both in temperature and precipitation. However, the application

of the "pure" bias correction method CDF-t on separate GCMs can generate more biases, with more variability. This is due to the fact that 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 the CDFs of the different GCMs, the pooling techniques are also combining the changes over time, resulting in bias corrected projections more in agreement with the reanalyses.

The results of the PME showed 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 by CDF-t here. Overall, a quasi-systematic ranking of the four methods is observed in this PME: while CDF-t can present some recurrent and pronounced biases – getting larger for time periods that are further in time –, the MMM correction approach improves the results, the linear approach even more and the best results are obtained by the alpha-pooling technique for both variables. This confirms the interest 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 to get more robust estimates of future dependencies.

However, the CDFs resulting of our linear or α-pooling approaches might depend on the selected ensemble of model CDFs to combine. Hence, the choice of the models to combine 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, etc.). The sensitivity analysis of the future (2081-2100) CDFs to the choice of the ensemble of models showed that the uncertainty in long-term projections was found globally comparable for the three pooling-based methods, although slightly higher for α-pooling and slightly lower for MMM pooling. Indeed, as the α- and linear-pooling associate non-uniform weights to the different CDFs, they pull the results towards those of the models with the highest weights, hence generating more variability depending on the selected ensemble of models to combine. On the other side, the MMM pooling corresponds to a linear pooling with weights forced to be uniform. Therefore, it provides smoother CDF results, 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.

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

This work can be extended in various ways. First, event though only temperature and precipitation were considered in this study, many other climate variables – such as wind, humidity, etc. – 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 about climate changes.

In addition, some more technical and statistical developments could be explored to improve the CDF-pooling approach. For example, the present linear- and α-pooling methods are based on the $L2$ norm to estimate the parameters associated with the pooled CDF the closest to the reference one over the calibration period. Other distances could be used, and more specifically distances between distribution functions, e.g., the Kullback-Leibler divergence, the Hellinger distance or the Wasserstein distance. 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 weights and in parameter α, there is a remaingin variability between neighbour grid cells (Fig. 3 and 4) that complicates the interpretation of the parameters. Such a variability could be reduced by constraining the approach to provide more continuous and smoother spatial structures, possibly at the cost of longer computations.

Note also 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 days frequencies and distributions of wet intensity could be made by having two separate pooling. Although the α-pooling results for precipitation in this article were quite satisfying, such a rainfall-specific design could provide additional improvements and would deserve to 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 values theory – such as the Generalized Pareto Distribution (GPD) or the Generalized Extreme Value distribution (GEV) – would then have to be used. Behind the practical results that such an application could bring, the statistical properties of the resulting pooled (extreme) CDFs would also be worth studying from the theoretical point-of-view.

Another interesting perspective, both from the 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 to combine joint (i.e., multivariate) CDFs would allow to improve the modelling of dependencies between the variables and, thus, to provide 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 also allow us to investigate compound events (e.g., Zscheischler et al., 2018, 2020) and their potential future changes more robustly.

Beyond possible practical and theoretical improvements of our proposed α-pooling strategy, many uses of this approach can be made. For example, the weights obtained from the linear- and α-pooling methods inform us about the relative closeness of the models to the reference data, in terms of CDF. Hence, they could be used as basic tools for some model selection-related questions. Of course, this tool would only consider the distributional aspects and not the climate trajectory itself. However, by considering CDF-pooling on two different historical periods – potentially with a bivariate pooling – it would be possible to include an additional information about historical changes within such a selection.

Finally, more generally, it is worth noticing 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, max, etc., as usually done), which is, in itself, an original contribution to the combination framework. This CDF-pooling strategy and this hybrid combination-correction method would deserve to be further explored, as well as its potential applications, behind combination and bias correction.

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, 10, 91–105, https://doi.org/10.5194/esd-10-91-2019, publisher: Copernicus GmbH, 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, 23, 4803–4824, https://doi.org/10.5194/hess-23-4803-2019, publisher: Copernicus GmbH, 2019.

Allard, D., Comunian, A., and Renard, P.: Probability aggregation methods in geoscience, Mathematical Geosciences, 44, 545–581, 2012.

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

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. et al.: Verification of forecasts expressed in terms of probability, Monthly weather review, 78, 1–3, 1950.

Clarotto, L., Allard, D., and Menafoglio, A.: A new class of α-transformations for the spatial analysis of Compositional Data, Spatial Statistics, 47, 100 570, https://doi.org/https://doi.org/10.1016/j.spasta.2021.100570, 2022.

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), Journal of Advances in Modeling Earth Systems, 12, e2019MS001 916, https://doi.org/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, Monthly Weather Review, 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 re-analysis evaporation datasets for hydrological modelling under various model calibration strategies, 143, 103 667, 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.

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, Geoscientific Model Development, 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

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

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

Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, Journal of the American statistical Association, 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, Hydrology and Earth System Sciences, 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. R. Meteorol. 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, Journal of Advances in Modeling Earth Systems, 11, 3691–3727, https://doi.org/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, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/https://doi.org/10.1002/qj.3803, 2020.

Kleiber, W., Raftery, A. E., and Gneiting, T.: Geostatistical Model Averaging for Locally Calibrated Probabilistic Quantitative Precipitation Forecasting, 106, 1291–1303, https://doi.org/10.1198/jasa.2011.ap10433, publisher: Taylor & Francis _eprint: 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, https://doi.org/10.1002/2016GL072012, 2017.

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

Krinner, G. and Flanner, M. G.: Striking stationarity of large-scale climate model bias patterns under strong climate change, Proceedings of the National Academy of Sciences, 115, 9462–9466, https://doi.org/10.1073/pnas.1807912115, 2018.

Lange, S.: ISIMIP3b bias adjustment fact sheet, Technical report, ISIMIP, https://www.isimip.org/documents/413/ISIMIP3b_bias_adjustment_fact_sheet_Gnsz7CO.pdf, 2021.

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

Michelangeli, P., Vrac, M., and Loukos, H.: Probabilistic downscaling approaches: application to wind cumulative distribution functions, Geophys. Res. Lett., 36, L11 708, doi:10.1029/2009GL038 401, 2009.

Neyman, E. and Roughgarden, T.: From Proper Scoring Rules to Max-Min Optimal Forecast Aggregation, Operations Research, 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, 43, 7661–7669, https://doi.org/10.1002/2016GL069704, _eprint: https://onlinelibrary.wiley.com/doi/pdf/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, Climate Dynamics, 48, 367–386, 2017.

Robin, Y. and Vrac, M.: Is time a variable like the others in multivariate statistical downscaling and bias correction?, Earth System Dynamics Discussions, 2021, 1–32, https://doi.org/10.5194/esd-2021-12, 2021.

Rougier, J., Goldstein, M., and House, L.: Second-Order Exchangeability Analysis for Multimodel Ensembles, 108, 852–863, https://doi.org/10.1080/01621459.2013.802963, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2013.802963, 2013.

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

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

Strobach, E. and Bel, G.: Learning algorithms allow for improved reliability and accuracy of global mean surface temperature projections, 11, 451, https://doi.org/10.1038/s41467-020-14342-9, number: 1 Publisher: Nature Publishing Group, 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 M.Vrac: Combining Global Climate Models Using Graph Cuts, Clim. Dyn., 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, 173, 371–388, https://doi.org/10.1111/j.1467-985X.2009.00616.x, _eprint: https://onlinelibrary.wiley.com/doi/pdf/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., Stein, M. L., Hayhoe, K., and Liang, X.-Z.: A general method for validating statistical downscaling methods under future climate change, Geophysical Research Letters, 34, https://doi.org/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, doi:10.5194/nhess–12–2769–2012, 2012.

Vrac, M., Noël, T., and Vautard, R.: Bias correction of precipitation through Singularity Stochastic Removal: Because occurrencesmatter, Journal of Geophysical Research: Atmospheres, 121, 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?, Journal of Geophysical Research: Atmospheres, 127, e2022JD036 562, https://doi.org/https://doi.org/10.1029/2022JD036562, e2022JD036562 2022JD036562, 2022.

Wanders, N. and Wood, E. F.: Improved sub-seasonal meteorological forecast skill using weighted multi-model ensemble simulations, 11, 094 007, https://doi.org/10.1088/1748-9326/11/9/094007, publisher: IOP Publishing, 2016.

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

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

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.

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, Nature Clim Change, 8, 469–477, https://doi.org/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.

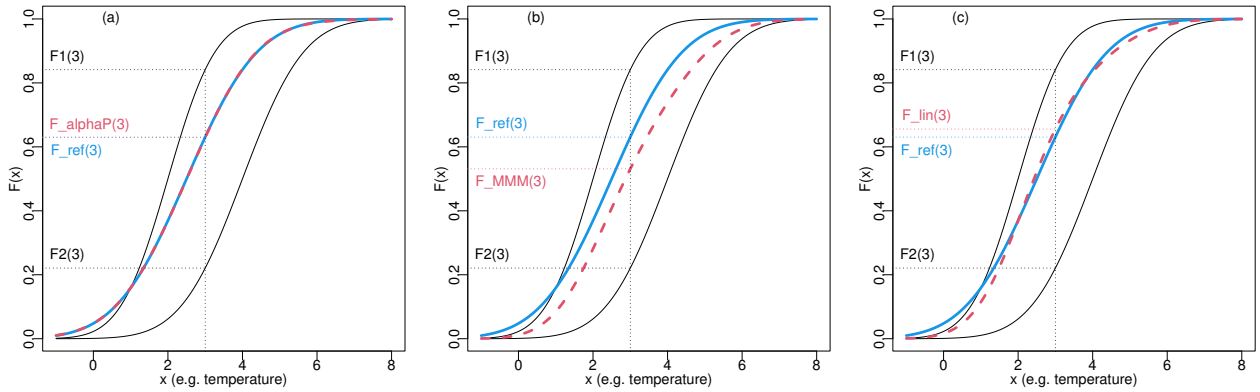

Figure 1. Examples for $N = 2$ GCMs with CDFs $F_1(x)$ and $F_2(x)$ (say of temperature) of: (a) the α-pooling approach, (b) the MMM approach and (c) the linear pooling approach. See text for details. In each panel, the black lines are the 2 CDFs. CDFs, the blue line is the reference CDF and the red dashed line is the resulting CDF from α-pooling (panel a), MMM (panel b), or linear pooling (panel c). Note that the reference is not used to perform MMM. An illustration is displayed for $x = 3$.

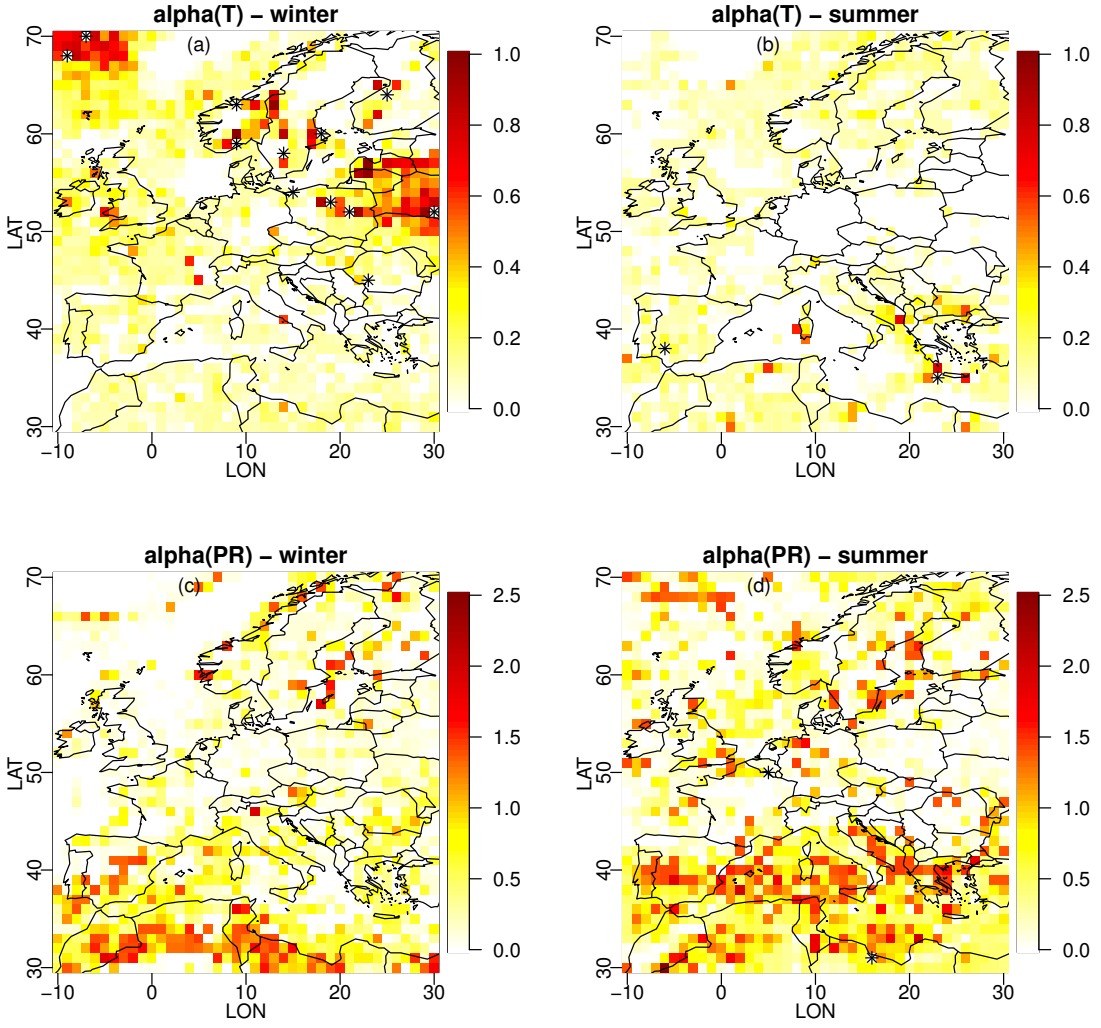

Figure 2. From α-pooling, maps of the parameters α obtained within the ERA5 experiment for temperature (a, b) and precipitation (c, d), over winter (a, c) and summer (b, d) seasons.

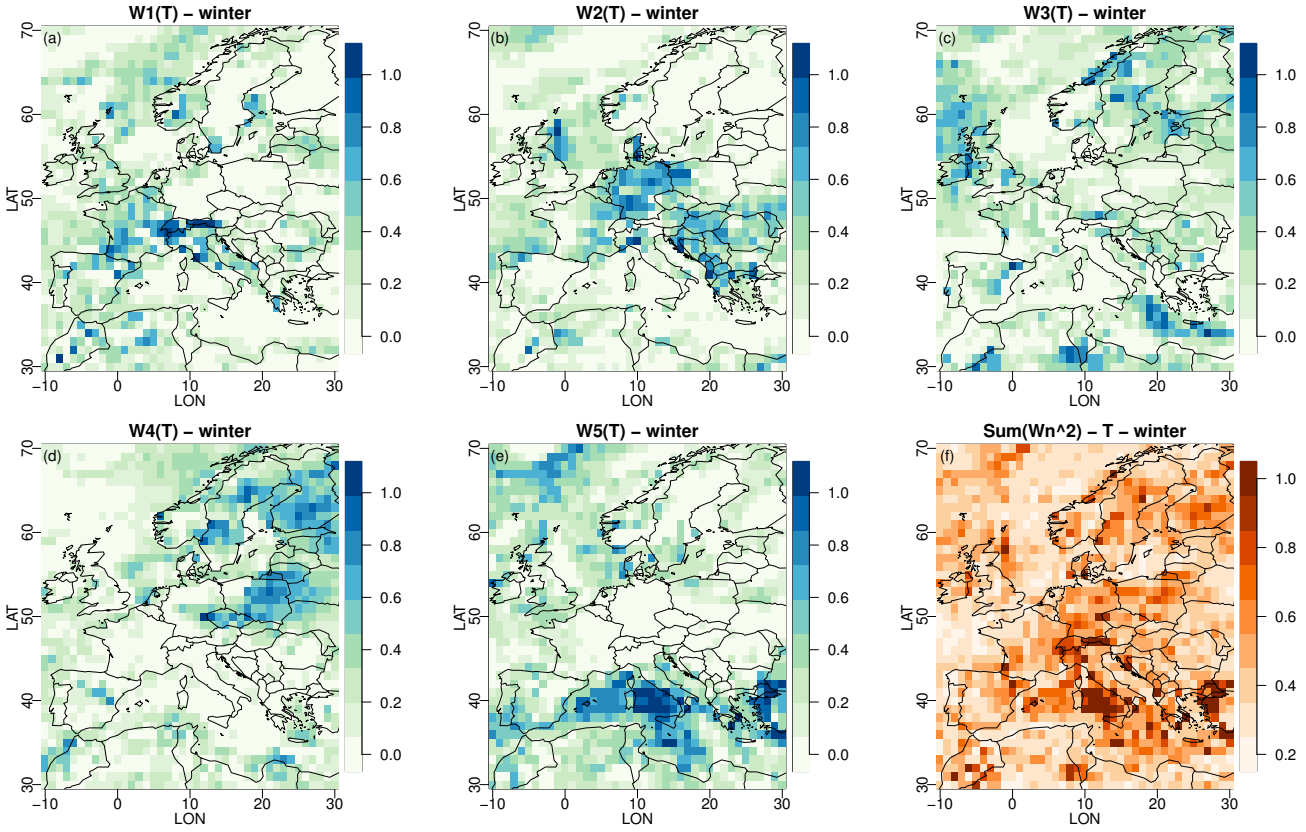

Figure 3. Maps of the weights parameters from α-pooling for winter obtained with the ERA5 experiment for temperature, over winter. Models 1 to 5 correspond respectively to CNRM-CM6-1-HR, GFDL-CM4, IPSL-CM6A-LR, MRI-ESM2-0 and UKESM1-0-LL. Panel (f) displays the concentration index, equal to sum of the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

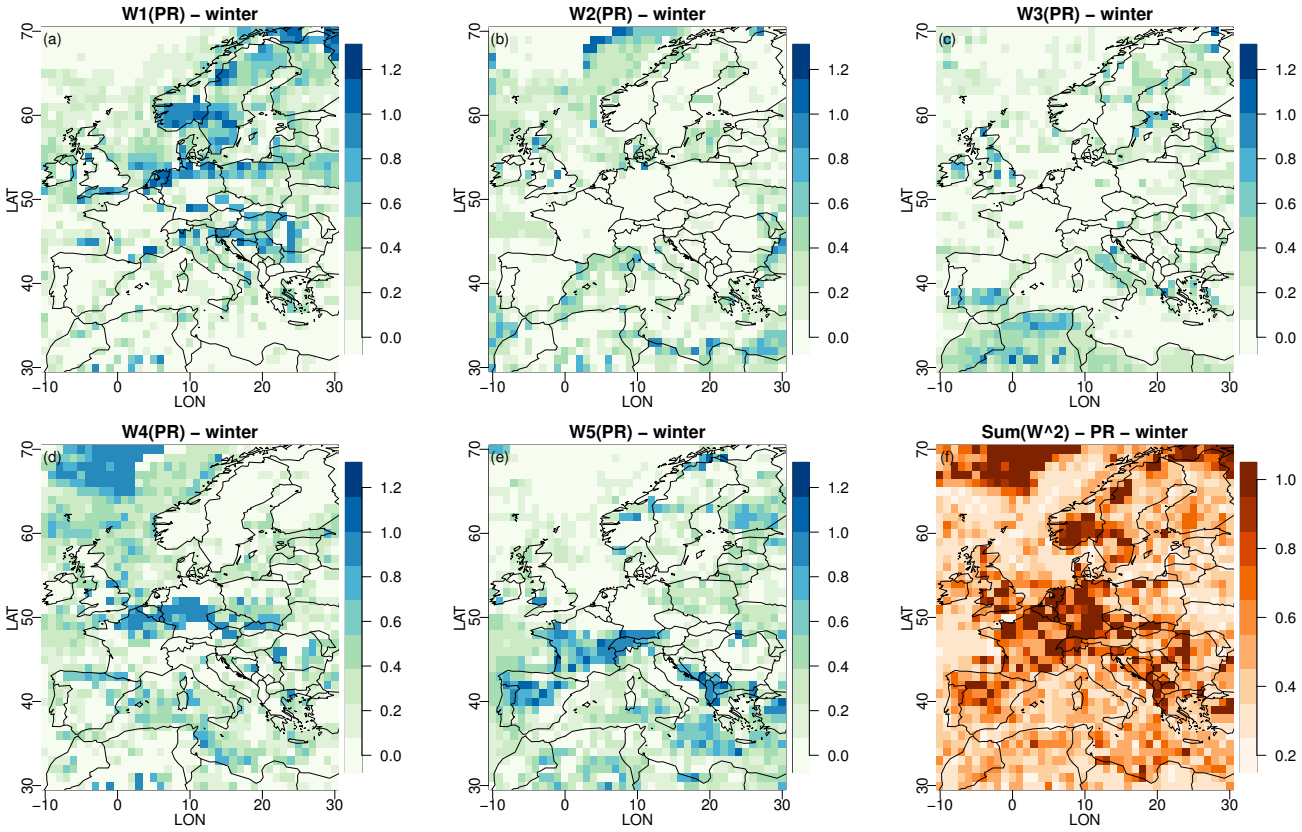

Figure 4. Same as Fig. 3 but for precipitation.

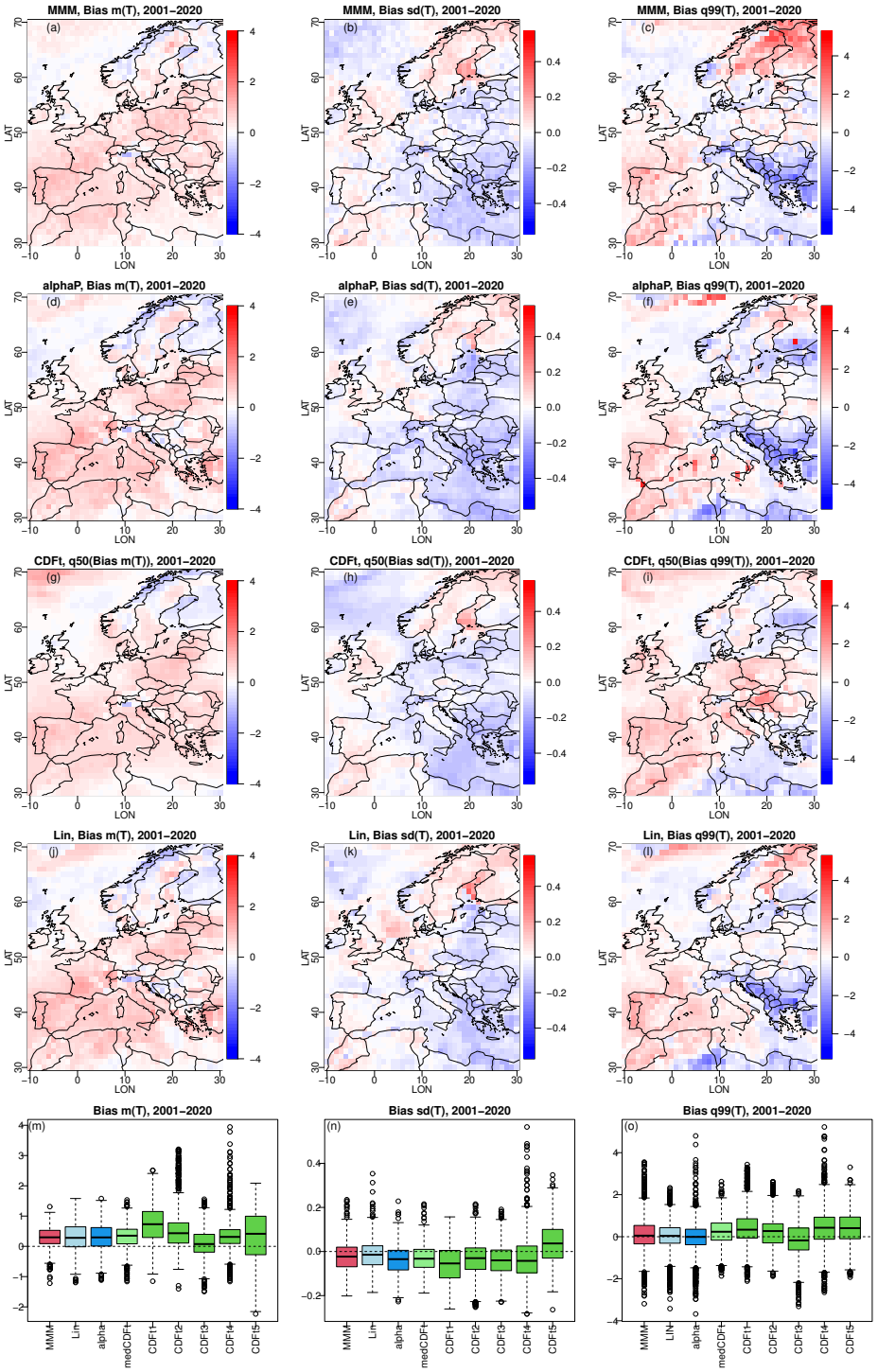

Figure 5. Biases in mean (left column), standard deviation (middle column) and 99% quantile (right column) for winter temperature from MMM (top row), α-pooling (row 2), CDF-t (third row) and linear pooling (fourth row) under the 2001-2020 (projection) time period of the ERA5 experiment. Third row corresponds to the grid-point median of the CDF-t biases. Fifth (bottom) row: boxplots of biases for MMM, linear pooling, α-pooling the median CDF-t biases, as well as for each of the 5 CDF-t results.

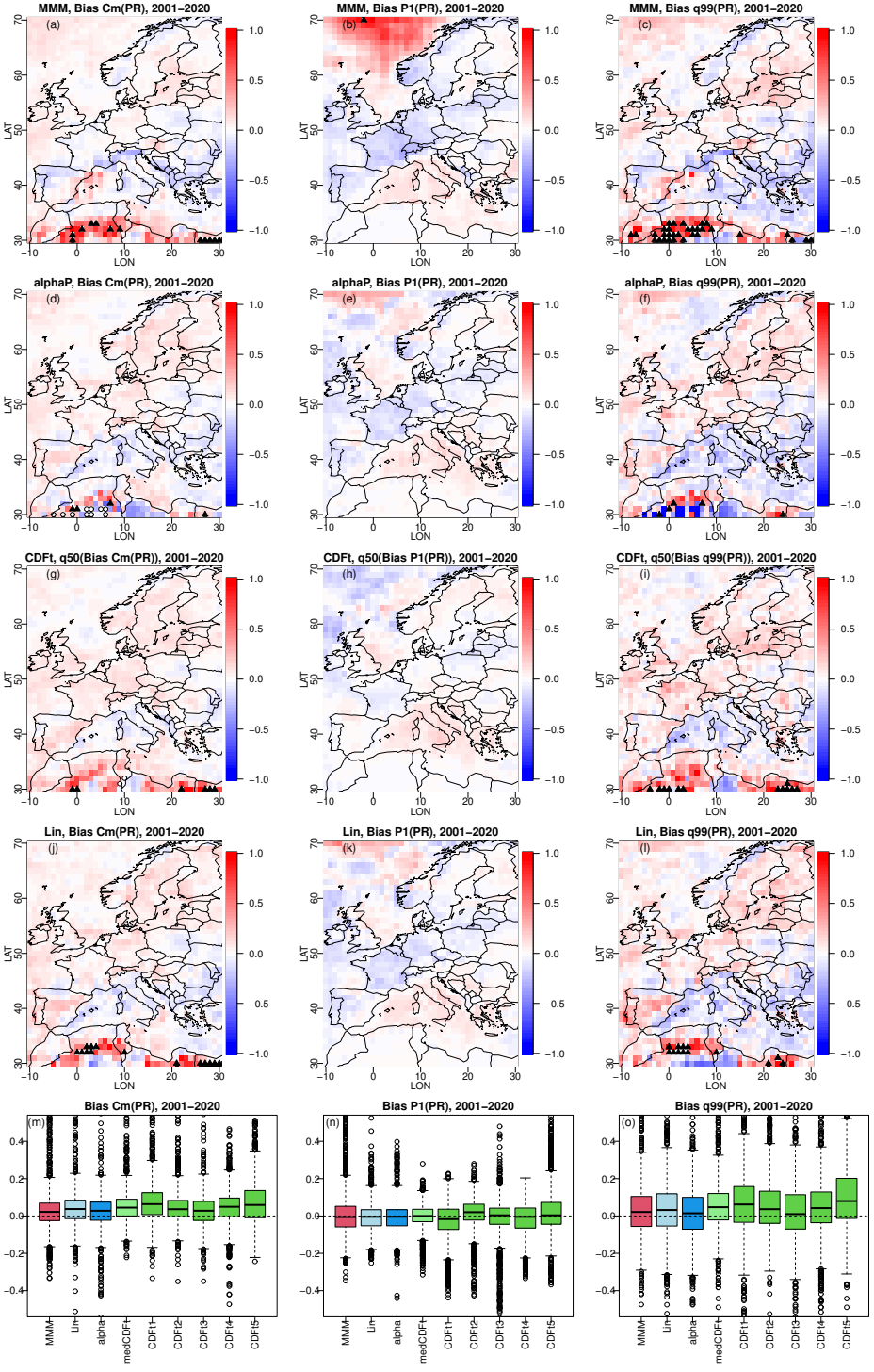

Figure 6. Same as Fig. 5 but for winter precipitation with biases in conditional mean given wet (left column), probability of dry day (P_1, middle column) and 99% quantile (right column).

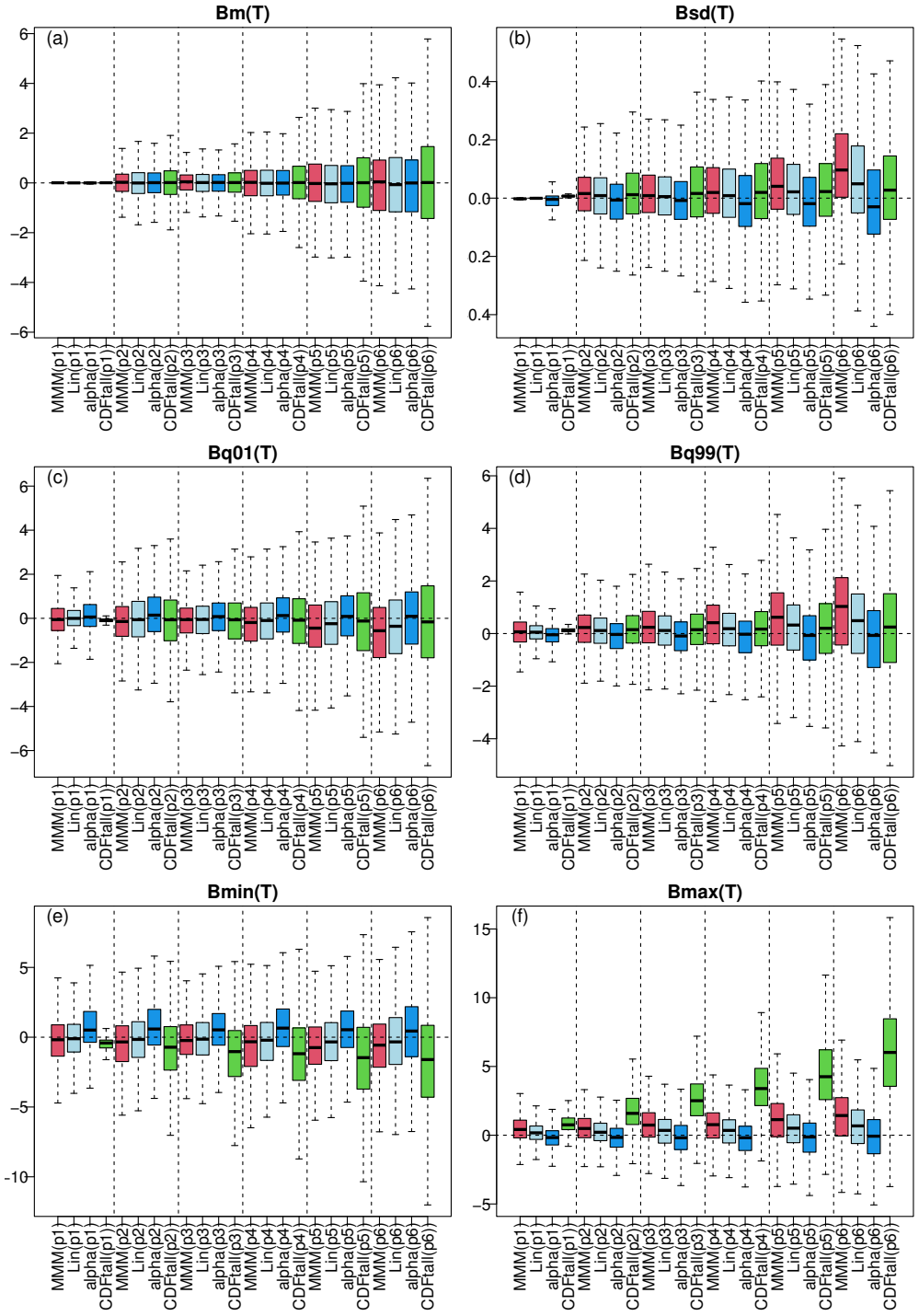

Figure 7. Results of the perfect model experiment for winter temperature: Boxplots of biases from the three methods (red=MMM, light blue=linear pooling, blue=α-pooling, green=CDFt) for the six 20-year time periods (from p1=1981-2000=calibration to p6=2081-2100). The different panels display biases in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

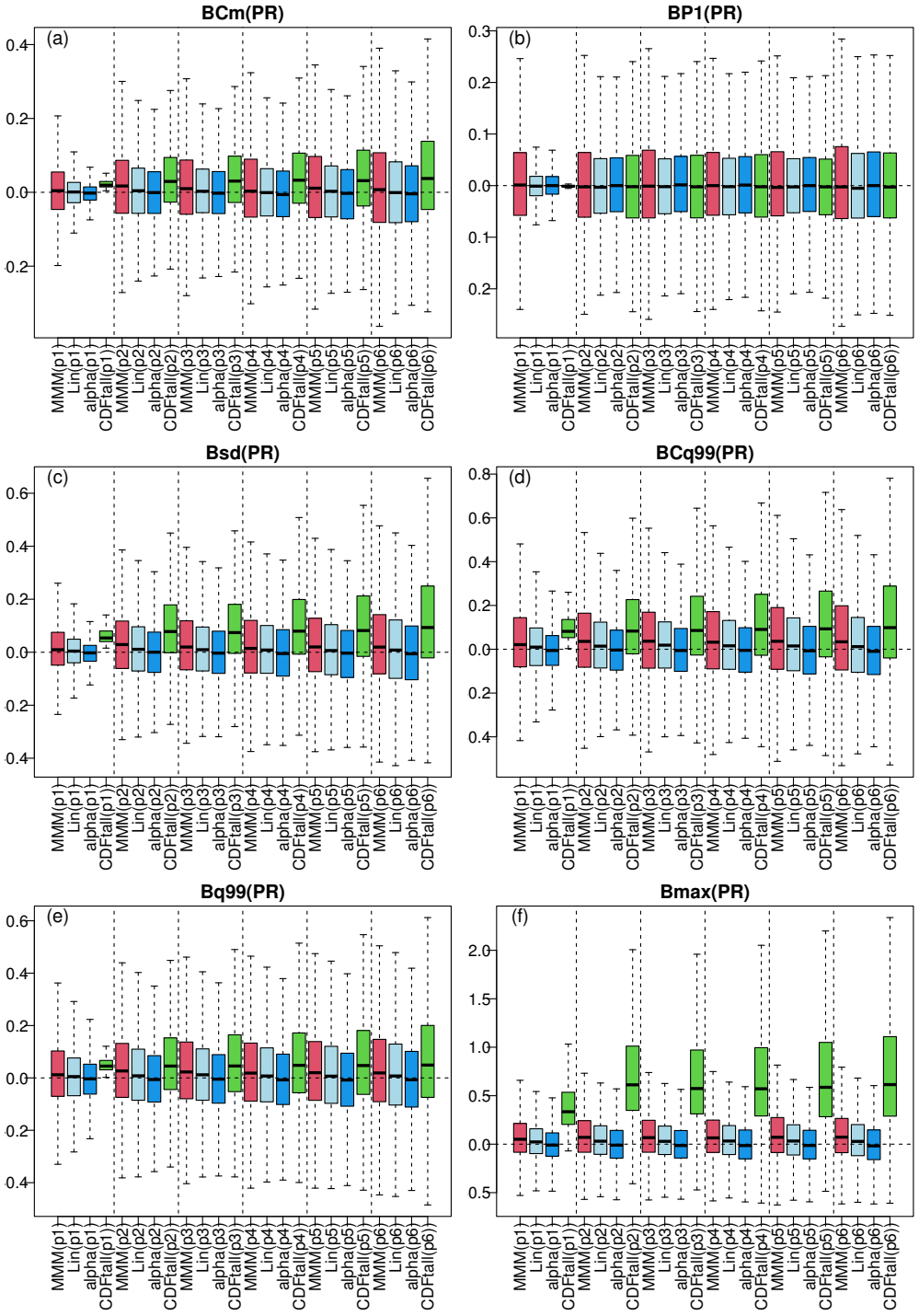

Figure 8. Results of the perfect model experiment for winter precipitation: Same as Fig. 7 but for precipitation. The different panels display biases in (a) conditional mean precipitation given wet, (b) probability of dry ($<$ 1mm) day, (c) standard deviation, (d) conditional 99% quantile given wet,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

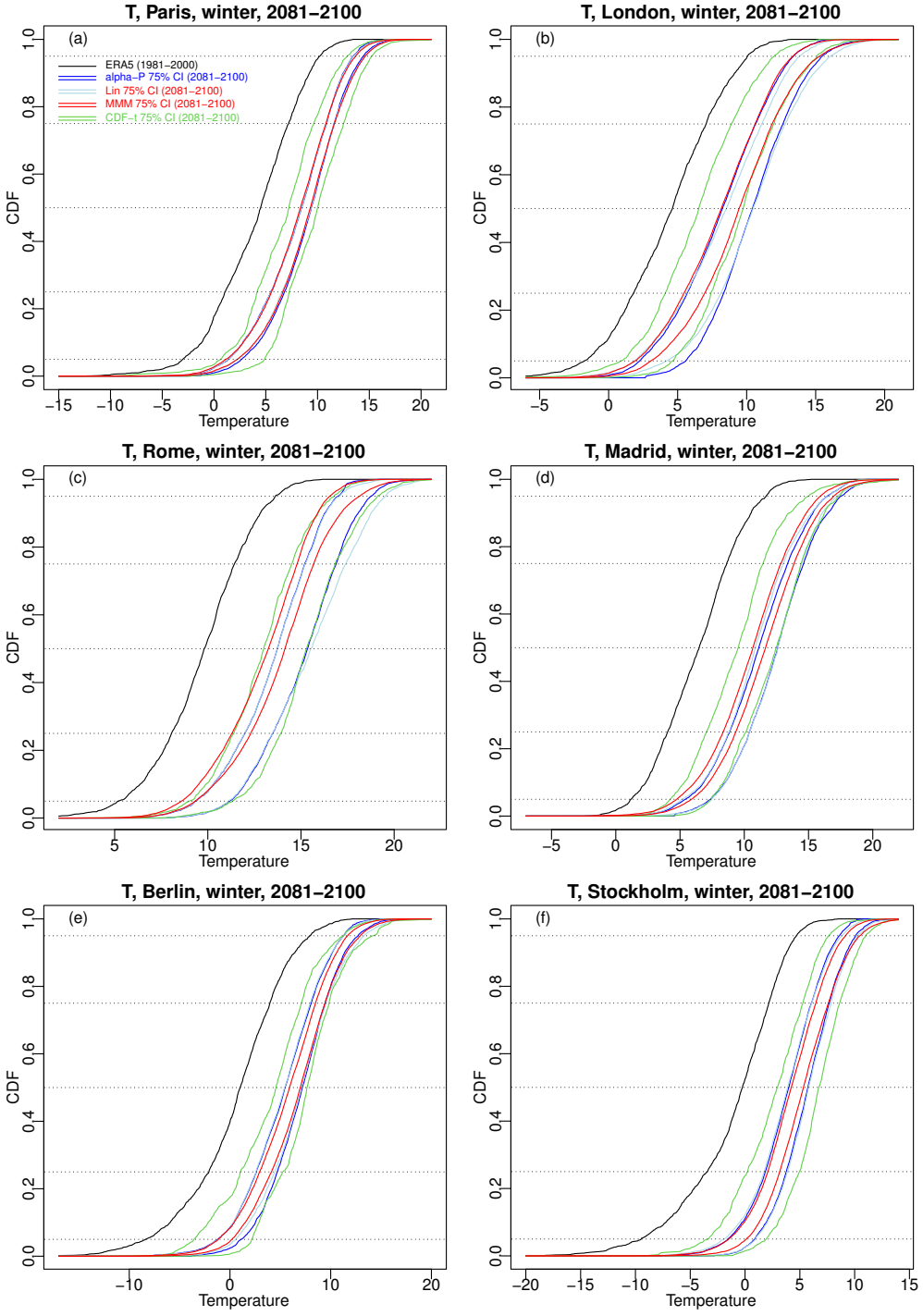

Figure 9. Results of the sensitivity experiment: For winter temperature over 2081-2100 and 6 major cities in western Europe, 75% confidence intervals for α-pooling (blue lines), linear pooling (light blue lines), MMM (red lines) and CDFt (green lines). The temperature ERA5 CDF (black line) over 1981-2000 is also displayed for visual evaluation of changes.

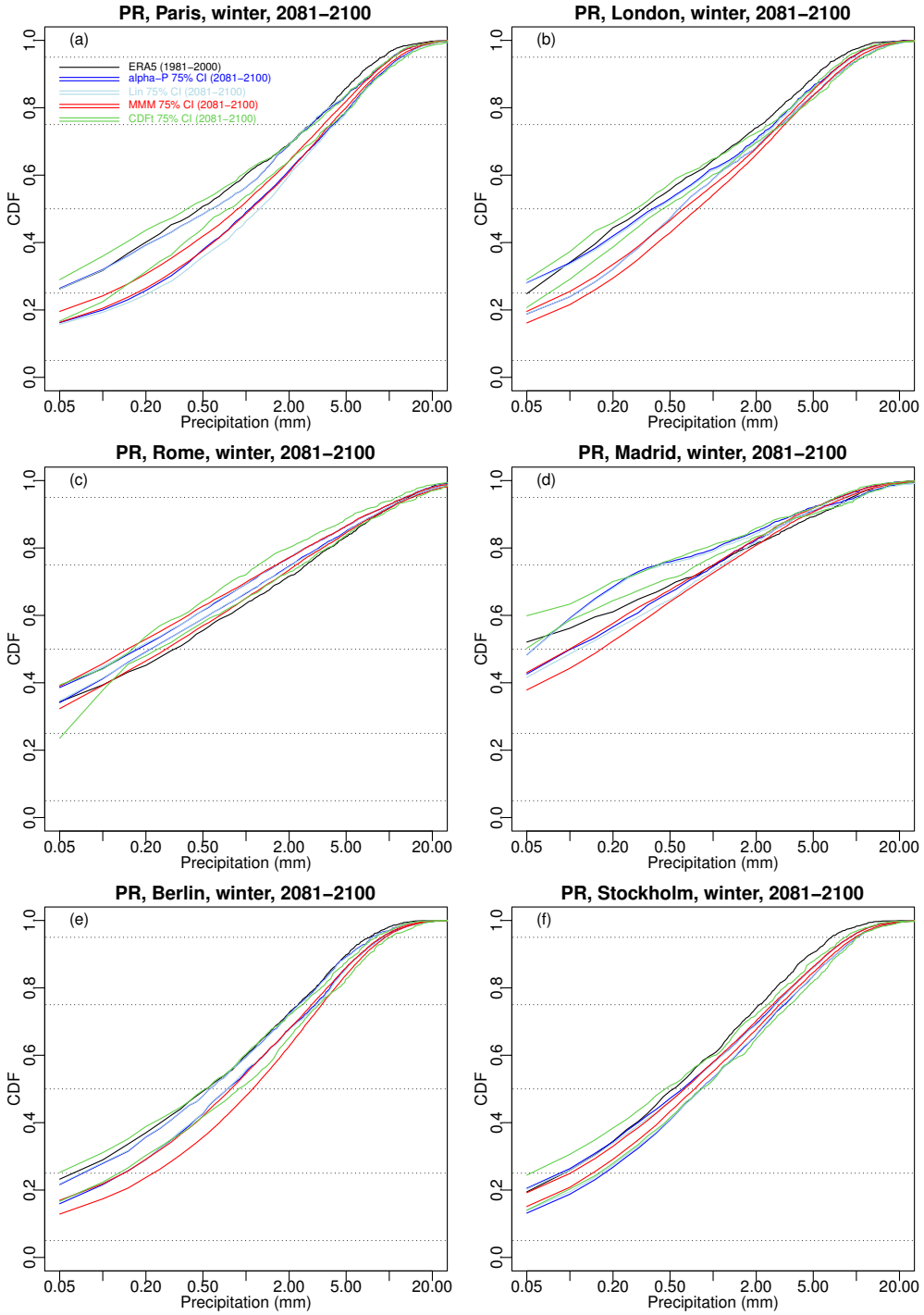

Figure 10. Results of the sensitivity experiment: Same as Fig. 9 but for precipitation. Note that the x-axis is displayed in log-scale to ease evaluation.

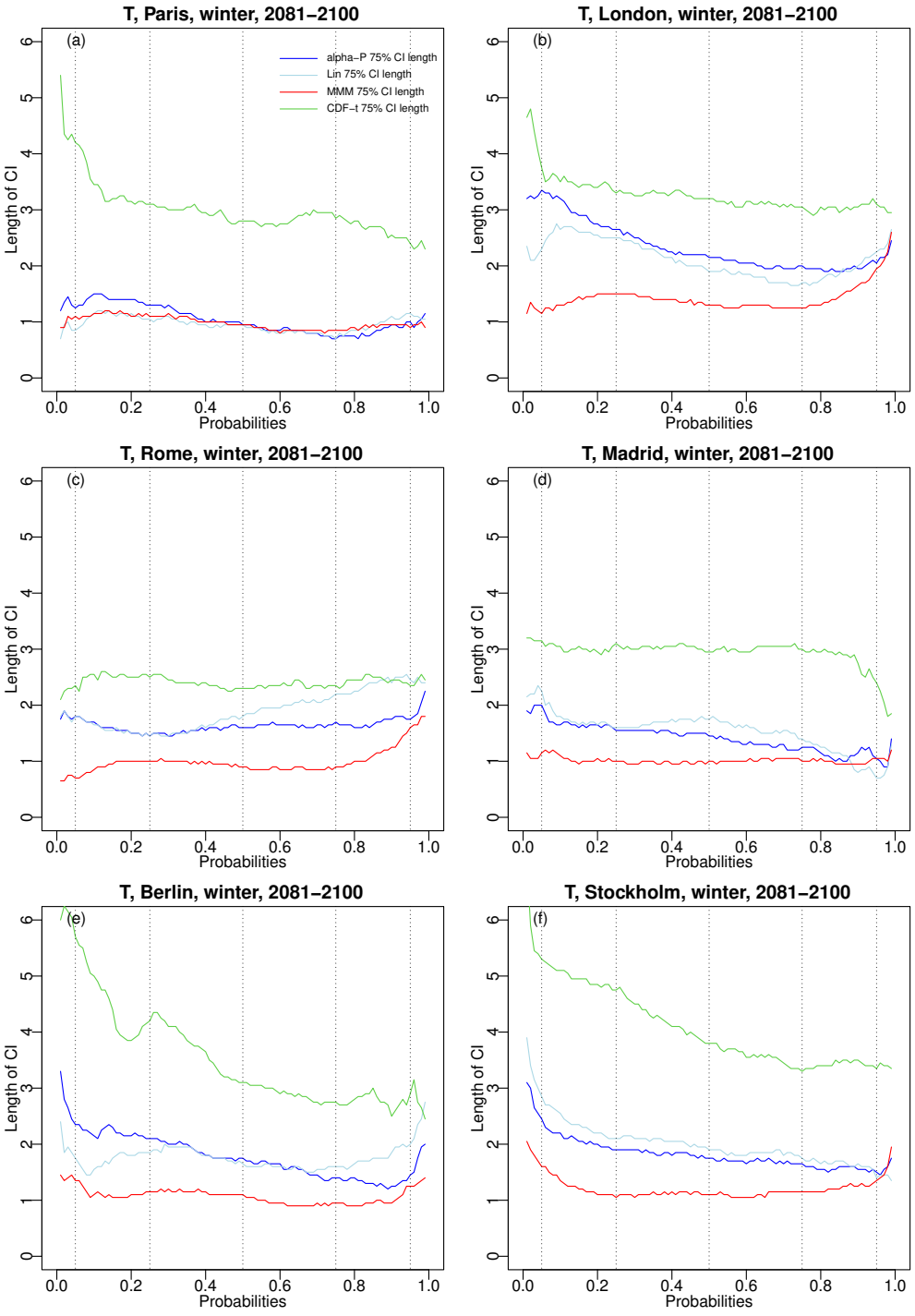

Figure 11. For winter temperature over 2081-2100 and 6 major cities in western Europe, length of the 75% CDF confidence intervals for MMM (red line), linear pooling (light blue line), α-pooling (blue line), and CDFt (green line).

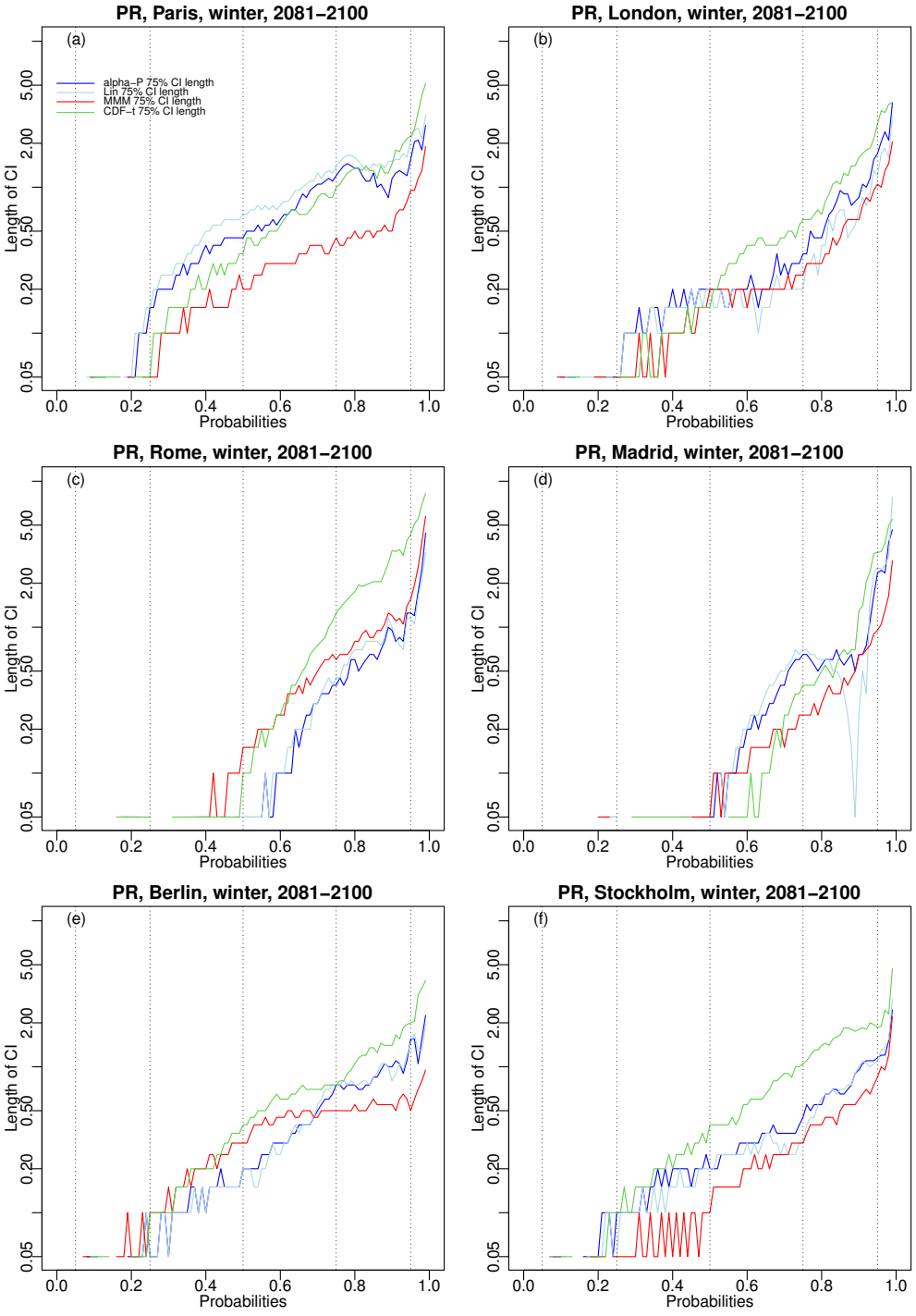

Figure 12. Same as Fig. 11 but for precipitation. Note that the y-axis is displayed in log-scale to ease evaluation.

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 to assume that the Box-Cox transformation of the pooled CDF is, up to a normalizing factor K, the weighted average of the Box-Cox transformation, ie:

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

After multiplying by α and rearranging, one gets

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

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

$$F_B(x) = \frac{\big[1 - S + \sum_{i=1}^{N} w_i F_i(x)^\alpha\big]^{1/\alpha}}{\big[1 - S + \sum_{i=1}^{N} w_i F_i(x)^\alpha\big]^{1/\alpha} + \big[1 - S + \sum_{i=1}^{N} w_i (1 - F_i(x))^\alpha\big]^{1/\alpha}}, \quad \forall x \in \mathbb{R} \tag{A1}$$

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

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

$$F_H(x) = \frac{\big[\sum_{i=1}^{N} w_i F_i(x)^\alpha + A\big]^{1/\alpha}}{\big[\sum_{i=1}^{N} w_i F_i(x)^\alpha + A\big]^{1/\alpha} + \big[\sum_{i=1}^{N} w_i (1 - F_i(x))^\alpha + A\big]^{1/\alpha}},$$

which is nothing but (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

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

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

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

When $\alpha = 1$, the pooling formula (A3) reduces to the linear pooling. As $\alpha \to 0$, it is straightforward to check that it boils down to the log-linear pooling (4). As was the case for the α-pooling presented in Section 3.4, this pooling formula generalizes thus

both the log-linear pooling and the linear pooling. It must be emphasized that replacing w_i by Kw_i with $K > 0$ in (A3) leads to the same value $F_{B,1}(x)$. Imposing or not $\sum_{i=1}^{N} w_i = 1$ in (A3) has thus no consequences on $F_{B,1}$.

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

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

- However, as discussed in Section 3.4, F_G is a proper CDF even if $S \neq 1$. 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 to pay is increased computation time.

- When using the direct approach in (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 (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 $\left((1-S)^{1/\alpha}+1\right)^{-1}$ was extremely small (say, less than 10^{-3}).

- In (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 report briefly some optimal properties of the α-pooling presented in Section 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 $\boldsymbol{p}$.

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

$$G(\boldsymbol{F}) = F^{1+\alpha} + (1-F)^{1+\alpha}, \tag{B2}$$

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

$$\boldsymbol{g}(\boldsymbol{F}) = (1+\alpha)(F^{\alpha},\ (1-F)^{\alpha})^t, \tag{B3}$$

with $\lim_{\alpha \to 0} \boldsymbol{g}(\boldsymbol{F}) = (1+\ln F,\ 1+\ln(1-F))^t$. Since in (B2) the function G is convex, the scoring rule given by (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 to $G(\boldsymbol{F})$ in (B2) varies thus 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 quasi-arithmetic pooling defined by

$$\boldsymbol{g}(\boldsymbol{F}_G) = \sum_{i=1}^{N} w_i \boldsymbol{g}(\boldsymbol{F}_i), \qquad w_i \geq 0,\ i = 1,\ldots,N, \qquad \sum_{i=1}^{N} w_i = 1, \tag{B4}$$

corresponds exactly to the α-pooling presented in Section 3.4. Neyman and Roughgarden (2023) showed that this pooling reflects a compromise between all probabilities $(\boldsymbol{F}_i)$, $i = 1,\ldots,n$, in the sense that it would correspond to the least wrong probabilities overall (ie accross all outcomes and all randomly chosen model according to $\boldsymbol{w}$) as measured by the scoring rule derived from (B2). They also showed (in Theorem 4.1) the following Max-Min property. Let us defined the following utility function

$$u(\boldsymbol{F}; j) := s(\boldsymbol{F}; j) - \sum_{i=1}^{N} w_i s(\boldsymbol{F}_i; j) \tag{B5}$$

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 $\min_j u(\boldsymbol{p}; j)$ is maximized by setting $\boldsymbol{F} = \boldsymbol{F}_G$ as given in (B4). In other words, the worst loss of scores (often interpreted as a reward) is maximized using quasi-arithmetic pooling.