AGU ADVANCING EARTH AND SPACE SCIENCES

Earth's Future

RESEARCH ARTICLE
10.1029/2024EF004972

What Will the European Climate Look Like in the Future? A Climate Analog Analysis Accounting for Dependencies Between Variables

B. Bulut[1,2], **M. Vrac**[1], **and N. de Noblet-Ducoudré**[1]

[1]Laboratoire des Sciences du Climat et de l'Environnement (LSCE-IPSL), CEA/CNRS/UVSQ, Université Paris-Saclay, Orsay, France, [2]Now at: UK Centre for Ecology & Hydrology, Wallingford, UK

Key Points:

- Wasserstein distance is used for the first time in climate analog calculations and compared with Euclidean and Mahalanobis distances
- Europe's future analogs are mostly located today south of Europe, except for the Balkans which need to look east to find their analogs
- As the climate warms, it will become more difficult to find a proper analog, leading to more challenges in thinking about how to adapt

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

B. Bulut,
burbul@ceh.ac.uk

Citation:

Bulut, B., Vrac, M., & de Noblet-Ducoudré, N. (2025). What will the European climate look like in the future? A climate analog analysis accounting for dependencies between variables. *Earth's Future*, *13*, e2024EF004972. https://doi.org/10.1029/2024EF004972

Received 13 JUL 2023
Accepted 9 DEC 2024

Author Contributions:

Conceptualization: B. Bulut, M. Vrac, N. de Noblet-Ducoudré
Data curation: B. Bulut
Formal analysis: B. Bulut
Funding acquisition: M. Vrac, N. de Noblet-Ducoudré
Methodology: B. Bulut, M. Vrac, N. de Noblet-Ducoudré
Project administration: M. Vrac, N. de Noblet-Ducoudré
Software: B. Bulut
Supervision: N. de Noblet-Ducoudré
Visualization: B. Bulut
Writing – original draft: B. Bulut, M. Vrac, N. de Noblet-Ducoudré
Writing – review & editing: B. Bulut, M. Vrac, N. de Noblet-Ducoudré

Abstract

Increasing the awareness of society about climate change by using a simplified way for the explanation of its impacts might be one of the key elements to adaptation and mitigation of its possible effects. This study investigates climate analogs, which allow the possibility to find, today, a place on land where climatic conditions are similar to those that a specific area will face in the future. The grid-based calculation of analogs over the selected European domain was carried out using a newly proposed distance between multivariate distributions, the Wasserstein distance, that has never been used so far for climate analog calculations. By working on the whole multivariate distributions, the Wasserstein distance allows us to account for dependencies between the variables of interest and for the shape of their distribution. Its features are compared with the Euclidean and the Mahalanobis distances, which are the most used methods up to now. Multi-model climate analogs analysis is achieved between the reference period 1981–2010 and three future periods 2011–2040, 2041–2070, and 2071–2100, for seasonal temperatures (mean, min, and max) and precipitation, from five different climate models and three different socio-economic scenarios. The agreement between climate models in the location and degree of similarity of the best analogs decreases as warming intensifies and/or as time approaches the end of the century. As the climate warms, the similarity between future and current climatic conditions gradually decreases and the spatial (geographical) distance between a location and its best analog increases.

Plain Language Summary

This study explores the concept of climate analogs, which can help us understand and prepare for future climate conditions. Climate analogs are places on Earth today that have similar climate conditions to what a specific area will experience in the future. The study focuses on Europe and uses a new method called the Wasserstein distance to calculate these analogs. This method takes into account the relationships between different climate variables. We analyze multiple climate models and emission scenarios for different time periods. The findings indicate that as we approach the end of the century and as scenarios become more severe, the agreement between climate models on best analogs decreases, although they point to similar geographical areas. Toward the end of the century, the similarity between future and current climate conditions will decline, and the distance between a location and its best analog will increase. This means that finding suitable climate analogs becomes more challenging. Overall, this study highlights the importance of understanding climate change impacts and finding ways to adapt and mitigate its effects through simplified explanations and climate analogs.

1. Introduction

Changing behaviors starts with awareness as the first step even if it is not sufficient by itself (Arlt et al., 2011; Halady & Rao, 2010). Therefore, awareness of climate change is required to reach the goals of the Paris Agreement and has been recently highlighted in the last Intergovernmental Panel on Climate Change (IPCC) reports (IPCC, 2022a, 2022b). The gap between recognizing the risks of climate change and taking social action is influenced by various factors, but shifting public perceptions and understanding of climate change is seen as essential for increasing public involvement (Fitzpatrick & Dunn, 2019; Khatibi et al., 2021; Owen, 2020). Explaining in a way that society can better understand is one of the most important factors in raising awareness against the threats that climate change will cause (Lee et al., 2015; O'Neill & Nicholson-Cole, 2009; Rohat et al., 2018). However, it is not easy to translate a very complex and uncertain phenomenon into a popular language that people can relate to their daily lives (Lorenzoni & Pidgeon, 2006). Therefore, climate analogs (CA)

are an effective method that can be used to make the studies of scientists understandable to the public (Fitzpatrick & Dunn, 2019).

The idea of CA is based on matching, that is, pairing, climate conditions at one location in a given time period (i.e., present), with climate conditions at another location and another time period (i.e., future). Matching the future and current climatic conditions of different locations provides a simplified representation of the changes due to climate change. As a simple example, while it makes sense for a climate scientist that, for a given region, the summer precipitation will decrease by one standard deviation from the mean and that temperature will increase by 1.5°C, it is a more understandable definition for the public that the city (i.e., Paris) they live in today will resemble the city in a region located further south/equatorward (i.e., Bordeaux) in the future. In this way, a farmer or a municipality can act by understanding how its activities can be adapted in the future, building experience from current conditions at this other southern location. Therefore, the utilization of CA in the presentation of potential regional changes not only facilitates a basic understanding but also can support the development of dependable adaptation strategies.

CA, as a (dis)similarity or distance-based method, requires the use of an appropriate metric to accurately measure the (dis)similarity of matches made using climate variables like precipitation and temperature. This measurement is achieved through the use of similarity (or, inversely, dissimilarity) calculations, which is a mathematical distance. By comparing the climate conditions of different locations and determining their similarity, the climatic similarity of these two locations can be established. There are many studies in the literature in which various similarity metrics are used and developed for the calculation of CA at different scales. The choice of similarity metric depends on the research question, computational complexity and data scale. The most commonly used metric to calculate climate analogs is the Euclidean distance (ED) which measures the straight-line distance between two points in a multidimensional space. ED based CA were used to investigate different topics such as climate change from a global city analogs analysis (Bastin et al., 2019), European cities' climate (Rohat et al., 2018), novel and disappearing climates on a global scale (Williams et al., 2007), and production potential under climate change on existing agricultural areas around the world (Pugh et al., 2016). The methodology of CA has been widely employed in studies of cities, primarily to investigate which cities will be climatically similar to current cities in the future, or to identify the global climate types that will emerge or decline over time. Apart from studies that use recent historical data, CA have also been examined using past earth system paleoclimate variables (Burke et al., 2018). Furthermore, Grenier et al. (2013) examined six different techniques previously used in the CA literature and concluded that standardized ED is the best metric for selecting spatial analogs. However, it is important to note that the evaluation was conducted using the methods available at that time. Therefore, it did not include the two other methods besides ED, which are used in this study and explained in the following paragraphs.

In the literature, the CA are generally calculated from statistical parameters, such as the mean or standard deviation, derived from chosen multiple climate variables (generally based on temperature and precipitation) over the present and future time periods. However, the dependence (i.e., correlation) between these multiple variables of interest was generally left unaddressed, while the possible change of the dependence between variables within time (e.g., interannual variability or extremes) also affects the joint distribution of the variables. The importance of the dependency between climatic variables has been clearly illustrated for extreme weather events (Leonard et al., 2014; Salvadori et al., 2016). In recent years, the Mahalanobis distance (MD) presented by Mahony et al. (2017) started to become a more dominant method in CA studies (Fitzpatrick & Dunn, 2019; King, 2023; Lotterhos et al., 2021). MD method offers more benefits compared to ED as it uses of variable dependencies at the location of interest while searching for the best analogs in candidate locations. Even though the MD method considers the dependencies between the variables at the focal location, due to the use of climatological means (long-term means) as variables for analog locations calculation, it still leaves out some points (such as; dependencies between variables at candidate locations and shapes of the variable distributions) which are later discussed in the following sections. When considering the dependencies between variables as a criterion for selecting the best climate analog, it's important to note that while the ED method lacks this feature, the MD method incorporates the interannual variability and the correlation between climate variables but only for the focal/reference location, thereby losing an important part of the climate signal, the interannual variability and the correlation at the analog position. On the other hand, the Wasserstein distance (WD) method uses dependencies as a criterion by calculating them for both focal and candidate locations (see Section 2.2 about the calculation of the different distances). What we mean by "criterion" in this context is that similarly to sharing climatology being a criterion for identifying the best analog, having reference and analog locations sharing similar dependencies

between variables is a significant criterion for the WD method. In other words, only the WD method searches for analogs by considering the dependencies between variables. Therefore, considering the dependence between variables in the calculation of multivariate climate analogs should allow a more complete understanding of the relationships between the climate variables, which can improve the accuracy and reliability of the analogs and ultimately provide more robust information for decision-making related to climate change.

In this paper, we propose a new approach, based on the WD Rüschendorf (1985) for climate analog calculations, that considers the multivariate (i.e., joint) distribution of the climate variables of interest and, therefore, is able to account for both univariate statistics and intervariable dependencies. The method is applied to grid-based climate model outputs in order to identify analogs for every region (grid point) on Earth, as opposed to being limited to specific cities or sites. The grid-based analysis aims to enhance the comprehensiveness of results in CA analysis through the implementation of unconstrained location matching, apart from looking at analogs on land only. Thus, the best analog of each grid point within Europe is found from globally available grid points over lands. In addition, this study enables the identification of robust analogs through a multi-model analysis and also evaluates the agreement across climate models from a CA perspective.

The paper is organized as follows. In Section 2, information about the datasets used is given and the definitions of Euclidean, Mahalanobis and Wasserstein distance methods are provided. In Section 3, a comparison between the three methods by using a synthetic dataset is performed and discussed. In Section 4, the CA results obtained by using WD for the example city of Paris from various GCMs are provided, as well as the investigation of multi-model best CA results and their consistencies. In addition, the detailed overall CA results for the selected European domain are presented and discussed. In Section 5, conclusions and some future research directions are given.

2. Materials and Methods

In this study, the CA methodology aims to look for the location with today's (or very recent past) climate conditions similar to the simulated future climatic conditions of a selected reference location (location of interest). This approach is basically based on the calculation of distances between the simulated future climatic conditions at the location of interest and today's conditions at any other terrestrial location. The outcome of such an analysis provides comprehensive information regarding the temporal (as specified by the designated time periods) and spatial variations of climatic conditions, contingent upon Shared Socioeconomic Pathways (SSP).

The smallest distance corresponds to the highest (i.e., best) similarity, so the location where the current climatic conditions are the most similar to the possible future conditions at the reference location can be called the best analog. Unlike previous CA studies, this study aims to identify the best analog locations not based on similarity comparisons between selected cities or specific areas, but on a globally gridded scale. Specifically, similarity calculations are performed between each grid point within the study region and all other grid points available over land across the globe.

In this study two different analyses are performed; first, Euclidean, Mahalanobis and Wasserstein distance methods are applied to synthetic datasets in order to compare the properties of the methods. Then, the calculation of CA using climate variables is carried out with the most novel approach, the Wasserstein distance method. All analyses regarding this study are performed using the *R* environment (R Core Team, 2021).

2.1. ISIMIP3b Climate Data

The input climate variables used in this study are obtained from the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP) which aims to assess the impacts of climate change on different sectors at various time horizons (Warszawski et al., 2014). The project's third protocol (ISIMIP3b) offers five bias-adjusted CMIP6 (Coupled Model Inter-comparison Project Phase 6) climate models (Table 1) for three different socio-economic scenarios including low SSP126 (SSP1-RCP2.6), mid SSP370 (SSP3-RCP7.0), and high SSP585 (SSP5-RCP8.5), interpolated at 0.5° spatial resolution (Lange & Büchner, 2021).

In total, the 15 different available global climate datasets (five Global Climate Models (GCMs) and three SSPs) are used in the study. Climate conditions for a specific time period are defined using 30 consecutive years: current/reference climate (referred as Historical—HS) uses years between 1981 and 2010, Early Future (EF) conditions are for 2011–2040, Mid Future (MF) for 2041–2070, and Far Future (FF) for 2071–2100.

Table 1
List of Used GCMs in ISIMIP3b

Model	Institution	Reference
GFDL-ESM4	Geophysical Fluid Dynamics Laboratory, USA	(Dunne et al., 2020)
IPSL-CM6A-LR	Institut Pierre-Simon Laplace, France	(Boucher et al., 2020)
MPI-ESM1-2-HR	Max Planck Institute for Meteorology, Germany	(Mauritsen et al., 2019)
MRI-ESM2-0	Meteorological Research Institute, Japan	(Yukimoto et al., 2019)
UKESM1-0-LL	Met Office Hadley Center, UK	(Sellar et al., 2019)

A general circulation model (GCM) is a type of climate model

Shared Socioeconomic Pathways (SSPs) are climate change scenarios of projected socioeconomic global changes up to 2100 as defined in the IPCC Sixth Assessment Report on climate change in 2021.

In our analysis, four climate variables are used to diagnose the CA: total daily precipitation, as well as mean, minimum and maximum daily temperature. All four seasons (e.g., the summer season in the northern hemisphere is JJA—June-July-August—while it is DJF—December-January-February—in the southern hemisphere) are considered therefore, in total, 16 variables are used to calculate the mathematical distances. Seasonal values are computed from daily datasets by taking the sum of daily precipitation for total seasonal precipitation and taking the mean, minimum and maximum temperature values within each season respectively for the mean, minimum and maximum seasonal temperature values. The seasonal minimum and maximum temperatures are thus respectively the smallest and highest daily temperature within the related season. Consequently, datasets of $16 \times 30 \times 4$ (variables × number of years × defined time periods) dimensions are generated for each grid point on a global scale for each GCM and SSP. Therefore, for a given triplet (GCM, SSP and future time period) the method will measure the distances between two sets of gridded climate data: today's climate (i.e., the HS period, hereafter referred to as the $\boldsymbol{A}$ matrix) and the time horizon targeted (i.e., the selected future period which can be EF, MF or FF, hereafter referred to as the $\boldsymbol{B}$ matrix).

The gridded climate data, $\boldsymbol{A}$ and $\boldsymbol{B}$, are $(n \times \mathrm{K}) \times \mathrm{T}$ matrices, n being the number of land points examined, K the number of climate variables per grid point (16 herein), and T the number of years per climate data (30 herein). In our analysis, n equals 92,889 land points for matrix $\boldsymbol{A}$ (all land points on the globe), while for matrix $\boldsymbol{B}$ it either equals 6,797 when we concentrate over Europe or one if we are interested in only one grid point/city (hereafter Paris).

In our study, searching for the best analog for a focal location j, in matrix $\mathbf{B}$, means finding the candidate analog location, i, from matrix $\boldsymbol{A}$ that minimizes the mathematical distance between their climatic conditions.

2.2. Methods of Calculating Distances

We intend to compare three methods, two of them being commonly used in the field of climate analogs (the standardized Euclidean Distance, hereafter referred to as ED, and the Mahalanobis Distance, hereafter referred to as MD). The third method is the one we propose, the Wasserstein Distance, hereafter referred to as WD.

We define some notations below that will be further used for the distance calculations:

a_{ikt} is the value of the climate variable k, at location i and for year t, within the matrix $\boldsymbol{A}$, which means for the HS period 1981–2010 (i.e., summer precipitation at Paris for year 1981);

b_{jkt} is the value of the climate variable k, at location j and for year t, within the matrix $\boldsymbol{B}$, which means for the targeted future period (i.e., summer precipitation at Barcelona for year 2071).

Therefore, $\boldsymbol{a_i}$ and $\boldsymbol{b_j}$ refer to (K × T) sub-matrices that store 30 years of data (at the seasonal time scale) for K variables. The other quantities we need for the distance calculations are:

$\overline{a}_{ik}$: the climatological (temporal) mean of a_{ikt};

$\overline{b}_{jk}$: the climatological (temporal) mean of b_{jkt};

s_{jk}: the standard deviation of variable k (i.e., measuring the interannual variability) at location j in matrix $\boldsymbol{B}$. It is the standard deviation of the projected future climate data at our focal location j ($(b_{jkt})_{t = 1,..,T}$ row vector).

2.2.1. Euclidean Distance

The Euclidean distance (ED_{ji}) between the projected climate data $\overline{b}_{jk}$ at a focal location j (over future period) and the historical climate data $\overline{a}_{ik}$ at any location i, is formulated by Williams et al. (2007) as follows:

$$ED_{ji}^{\ 2} = \sum_{k=1}^{K} \frac{\left(\overline{b}_{jk} - \overline{a}_{ik}\right)^2}{s_{jk}} \tag{1}$$

Here, location i is the analog candidate. The ED is computed separately for each climate variable and the squared distances are summed up. All climate variables are treated equally, no weighting of variables is applied, no correlation between variables is accounted for. The standardization of each variable (to make them comparable and summable) is accounted for via the standard deviation of the projected climate s_{jk}. The final best analog is the location i whose $\overline{a}_{ik}$ minimizes the ED with respect to $\overline{b}_{jk}$. More details can be found in Williams et al. (2007).

2.2.2. Mahalanobis Distance

The Mahalanobis distance, MD_{ji}, between the focal location j (over the future period) and a location i, its analog candidate (over the historical period) has been formulated by Mahony et al. (2017) as follows:

$$MD_{ji}^{\ 2} = \left[\overline{b'}_j - \overline{a'}_i\right]^T [\mathbf{R}_j]^{-1} \left[\overline{b'}_j - \overline{a'}_i\right] \tag{2}$$

where $\boldsymbol{R_j}$ is the correlation matrix of the $\boldsymbol{b_j}$ (K × K) and $\overline{a'}_i$ and $\overline{b'}_j$ are the row vectors (1 × K) of standardized 30 year mean climatological values at locations i and j respectively, defined as

$$\overline{a'}_i = \left(\overline{a'}_{ik}\right)_{k=1,\ldots,k} \text{ and } \overline{b'}_j = \left(\overline{b'}_{jk}\right)_{k=1,\ldots,k} \tag{3}$$

$$\text{with } \overline{a'}_{ik} = \frac{\overline{a}_{ik} - \overline{c}_{lk}}{\sigma(c_{lk})} \text{ and } \overline{b'}_{jk} = \frac{\overline{b}_{jk} - \overline{c}_{lk}}{\sigma(c_{lk})}$$

where $\overline{c}_{lk}$ and $\sigma(c_{lk})$ are the mean and standard deviation of variable k at the focal location l. It is important to note that in Mahony et al. (2017), both vectors of past $(\overline{a'}_i)$ and future $(\overline{b'}_j)$ mean climatological values were standardized with respect to $\overline{c}_{lk}$ and $\sigma(c_{lk})$, see Equation 3, that is, which in the original study was based on observed dataset. In this study, the standardization is done with respect to the data of the focal location j, over the selected future period, that is, with mean $\overline{b}_{jk}$ and standard deviation s_{jk}. Therefore, the $\overline{a'}_i$ and $\overline{b'}_j$ terms in Equation 2 become

$$\overline{a'}_{ik} = \frac{\overline{a}_{ik} - \overline{b}_{jk}}{s_{jk}} \text{ and } \overline{b'}_{jk} = \frac{\overline{b}_{jk} - \overline{b}_{jk}}{s_{jk}} = 0 \tag{4}$$

and the final MD_{ji} formulation to use becomes

$$MD_{ji}^{\ 2} = \left[\overline{a'}_i\right]^T [\mathbf{R}_j]^{-1} \left[\overline{a'}_i\right] \tag{5}$$

For clarification, the MD method is applied as in Mahony et al. (2017). However, the inclusion of an additional dataset (observational data, c_{lk}) for standardization introduced an extra term in the equation. In this study, we used only the GCM dataset, applying the standardization to a single dataset. As a result, only the standardized past values of the possible analog locations remain, while the standardized future focal location value canceled out.

The final best analog is the location i whose $\overline{a'}_i$ minimizes MD_{ji}. More details can be found in Mahony et al. (2017). Compared to the ED, the correlations between the selected climate variables in the targeted future are accounted for (via the $\boldsymbol{R}$ matrix). However, these correlations are not compared to the correlations at present time and at other locations, they allow the distance metric to weigh correlated variables more heavily. In other words,

when variables are highly correlated, the MD further adjusts by 'shrinking' the distance in the direction of high correlation, making it more sensitive to the true geometry of the data.

In the study by Mahony et al. (2017), MD was converted into Sigma Dissimilarity (SD) measures by transforming distances into percentiles of the chi distribution, with degrees of freedom equal to the number of climate variables, thus accounting for dimensionality. In our study, we also calculated SD values as the final results of MD method. However, for simplicity, we refer to this method simply as MD.

2.2.3. Wasserstein Distance

The Wasserstein distance (WD_{ji}) between the focal location j (over the future period) and i, its analog candidate (over the historical period) is calculated between the standardized yearly values of the (K × T) matrices $\boldsymbol{b}_j'$ and $\boldsymbol{a}_i'$. The values of the $\boldsymbol{b}_j'$ and $\boldsymbol{a}_i'$ matrices are calculated using:

$$a'_{ikt} = \frac{a_{ikt} - \overline{b}_{jk}}{s_{jk}} \; and \; b'_{jkt} = \frac{b_{jkt} - \overline{b}_{jk}}{s_{jk}} \tag{6}$$

One major difference with the other two methods is that the WD is using all the 30 year annual values of each variable, while the other methods are only or mostly based on climatological mean values of 30 years.

The WD is a metric based on optimal transport theory. It measures the optimum total transport cost required to move a set of n points from one distribution to another distribution in an m-dimensional phase space (Villani, 2009). Let's take a simple example with one-dimensional data (i.e., a single variable for each location). In this case, the location i over the historical period is only described by a'_{it} values (no more k index anymore) and the location j over the future period is only described by b'_{jt} values. Now, let μ and ν be two discrete probability measures, characterizing non-parametric distributions respectively of a'_{it} and b'_{jt}, defined as:

$$\mu = \sum_{t=1}^{T} \mu_t \delta_{a'_{it}} \; and \; \nu = \sum_{t=1}^{T} \nu_t \delta_{b'_{jt}} \tag{7}$$

where $\delta_{a'_{it}}$ and $\delta_{b'_{jt}}$ are Dirac masses at points a'_{it} and b'_{jt} respectively (i.e., functions such that $\delta_{a'_{it}}(x) = 1$ if $x = a'_{it}$ and 0 otherwise; and $\delta_{b'_{jt}}(x) = 1$ if $x = b'_{jt}$ and 0 otherwise) and whose fractional masses are μ_t and ν_t, respectively, with $\sum_{t=1}^{T} \mu_t = \sum_{t=1}^{T} \nu_t = 1$ and all terms in the summation are positive.

The quadratic Wasserstein distance, WD^2, between these two discrete distributions μ and ν can be written as (Vissio et al., 2020):

$$WD^2_{(\mu,\nu)} = \inf_{\gamma'_{t,t}} \sum_{t,t'} \gamma_{t,t'} \left[d(a'_{it}, b'_{jt'}) \right] \tag{8}$$

where, $\gamma_{t,t'}$ is the set of coefficients called the transport plan which describes how the fraction of mass transports from a'_{it} to b'_{jt} and $d(a'_{it}, b'_{jt'})$ is the usual ED between a'_{it} and $b'_{jt'}$. When all possible transport plans ($\gamma_{t,t'}$) are considered, Equation 8 is an optimization problem based on minimizing the transport cost. Therefore, the optimum transportation plan, in other words the WD value, is the result of this problem. The optimization is done by using the network simplex algorithm due to its availability and previous usage in a climate study (Robin et al., 2017, 2019). Since the WD is a mathematical distance, a value of zero means exact match while values greater than zero indicate distances between the distributions. Therefore, in practice in our case, the final best analog is the location i whose $\boldsymbol{a}_i'$ minimizes $WD(\mu,\nu)$.

In simple terms, WD refers to the overall transportation cost needed to move the distribution of $\boldsymbol{a}_i'$ values to match that of $\boldsymbol{b}_j'$ values. The values $\boldsymbol{a}_i'$ and $\boldsymbol{b}_i'$ can be visualized as the coordinates of datapoints (30 yearly climate variables data in 16 dimensions). The cost calculation involves utilizing the optimal transportation plan between

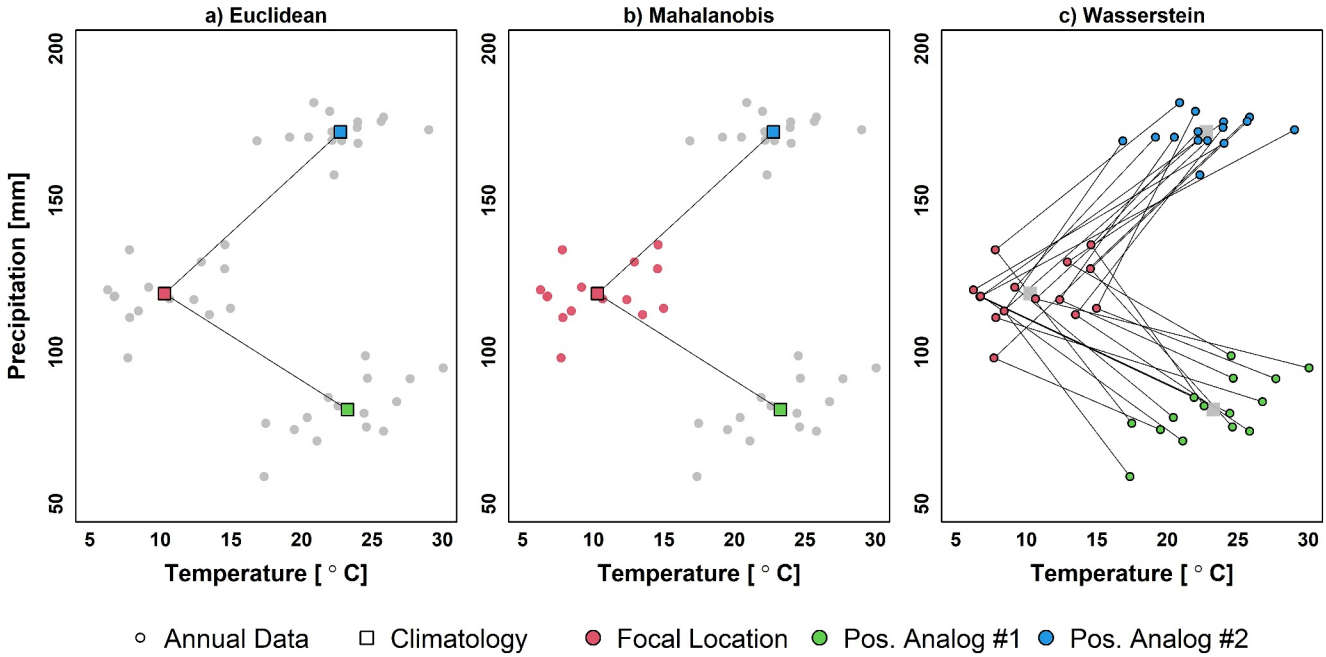

Figure 1. Scatter plots of the annual temperature, precipitation values and their climatological (temporal) means for three illustrative locations (red focal location, green and blue for possible analog locations) used in each distance calculation methods: (a) Euclidean, (b) Mahalanobis and (c) Wasserstein. The colored points indicate the used data in calculation, circular points are the annual values where square points show climatological means and lines between focal location and possible candidate locations represent calculated distances. When the points are colored it means they are used for the distance calculations (e.g., at panel b and c interannual values are used to calculate correlation).

these coordinates of $\boldsymbol{a}_i'$ and $\boldsymbol{b}_i'$ and the location i where the minimum cost is found is the final best analog. More information about the WD and its mathematical explanation can be reached from the studies of Robin et al. (2017, 2019) and references therein. In this study, WD calculations are done by using the "transport" package (Schuhmacher et al., 2020) in the *R* environment.

In summary, the key differences between the methodologies of the three distance calculation methods first lies in their use of datasets. ED relies on mean values for both focal and other locations (Figure 1a); MD utilizes the complete 30 year dataset for the focal point but only average values for other locations (Figure 1b), while WD uses all 30 year datasets in all calculations and locations (Figure 1c). Another major distinction is how they handle dependencies between variables. WD is the only method that accounts for dependencies between variables because of its optimum transport plan calculation applied to the multivariate distributions (lines in Figure 1c); MD uses intervariable correlations only at the focal point and employs it only as weighting factors applied to ED values in Equation 2; and ED does not consider any dependency.

3. Comparison of the Methods

3.1. Synthetic Dataset

In order to evaluate the performance of the different distance calculation methods we have constructed synthetic bivariate datasets that have been divided into four categories based on the correlation between variables and the probability distribution types. These four categories are labeled as (a) "dependent and gaussian", (b) "independent and gaussian", (c) "dependent and skewed", and (d) "independent and skewed". The terms "dependent" and "independent" refer to the correlation between variables within the bivariate dataset, while "Gaussian" (the data points cluster around the mean and the distribution is symmetric) and "Skewed" (the data points are not distributed symmetrically around the mean, with more observations on one side of the distribution) refer to the shape of the probability distribution of the data. The distinction between categories is solely based on dependency and marginal probability distribution types, while other statistical properties such as the mean and standard deviation of each variable are kept similar. Each category contains 100 samples of bivariate data, and each sample

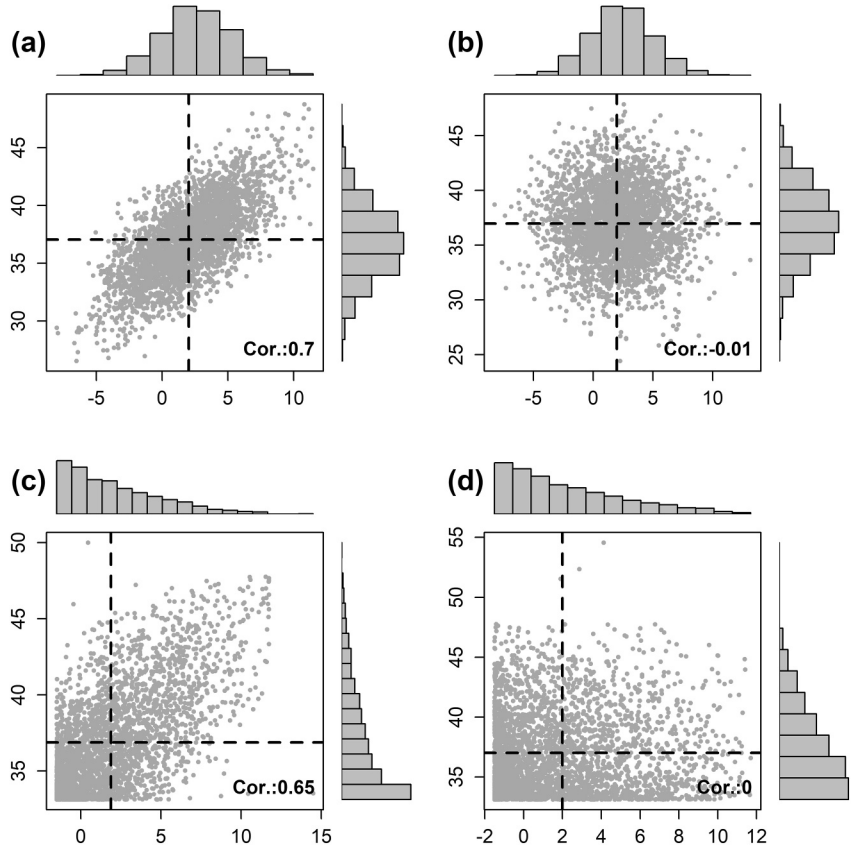

Figure 2. Scatter plots and distributions of the 100 samples for each synthetic bivariate dataset category. Categories are; (a) dependent-gaussian, (b) independent-gaussian, (c) dependent-skewed, and (d) independent-skewed. The mean values of the variable are shown in dashed lines and the Pearson correlation values between variables are given at the bottom right.

has 30 data points for each of the two variables. This number of data points is chosen to be consistent with the main analysis of the study, which uses climate variables obtained from 30 year periods.

The rmvnorm (Genz et al., 2021) and unonr (Qu & Zhang, 2020) *R* functions are used to generate synthetic data from a multivariate normal and multivariate non-normal distributions, respectively. In both functions three arguments: n, which specifies the number of data to be generated, the mean vector, which specifies the average value of each variable in the distribution, and the sigma which is the covariance matrix specifies the relationships between the variables, are required. We used the rmvnorm function with $n = 30$ and the mean vector (2, 37) for all four categories while only changing the covariance matrix (variances are 9 and 11) to generate dependent (Pearson's $r = 0.7$) and independent ($r = 0$) normal bivariate datasets. For the non-normal datasets, we used the same set of arguments and just added the skewness argument as one in the unonr function, allowing to generate bivariate non-Normal data using the Vale and Maurelli's method (Vale & Maurelli, 1983).

The distribution and scatter plot of the synthetic datasets (100 samples × 4 categories × 30 data points × 2 variables) are presented in Figure 2. The mean values of the first and second variables are indicated by dashed lines, and the correlation between the variables is provided in the bottom right corner of each scatter plot.

For instance, the first sample from the categories of gaussian distribution (a) and (b) include bivariate data sets ($a_{1,i}$, $a_{2,i}$) and ($b_{1,i}$, $b_{2,i}$) (where $i = 1, \ldots, 30$), respectively. The means and standard deviations of the ($a_{1,i}$, $b_{1,i}$) and ($a_{2,i}$, $b_{2,i}$) pairs are identical in both categories. Therefore, the only distinction between categories (a) and (b) is the degree of correlation between their bivariate data sets; there is a non-zero correlation between ($a_{1,i}$, $a_{2,i}$), while the correlation between ($b_{1,i}$, $b_{2,i}$) is null. Similarly, categories (c) and (d) are created by adhering to the same logic but

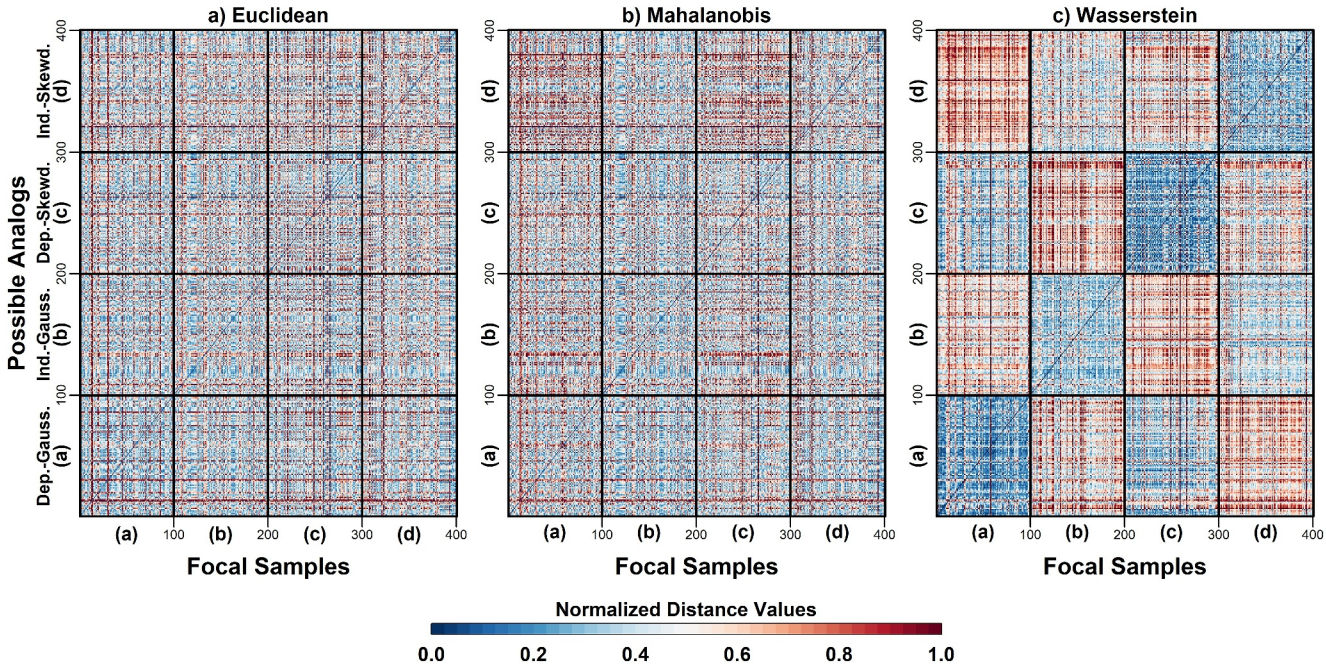

Figure 3. Matrices of the normalized distance values between each pair of samples, as calculated by the Euclidean Distance (ED, panel (a), Mahalanobis Distance (MD, panel (b) and the Wasserstein Distance (WD, panel (c). The focal samples are located in the *x*-axis where distance values calculated with respect to those are given in the *y*-axis. The diagonal values are all zeros, and the category labels are the synthetic data categories; (a) dependent-gaussian, (b) independent-gaussian, (c) dependent-skewed, and (d) independent-skewed. Blue indicates the smallest distance and red the highest one.

with the use of skewed distribution (positive skewness, with most values concentrated around the left tail) instead of gaussian distribution, in order to examine the effect of the distribution type on the distance calculations.

3.2. Comparison of Distance Calculation Methods

We calculated the distance between the different synthetic datasets by using all three distance methods. As the calculated distances from each method have different value ranges, we normalized the distances in order to allow a visual comparison of the results. The normalization process involves subtracting the mean and dividing by the standard deviation. Subsequently, for visualization purposes, values less than −2 and greater than +2 are censored to these thresholds. Finally, all values are rescaled to fit within the 0 to 1 range. In Figure 3, blue indicates the closest distance and red the furthest. The diagonal of the matrix, representing the distance between a sample and itself, is always zero.

The results obtained with the ED method (Figure 3a) show that all samples are nearly equidistant, regardless of the category, meaning that the ED method cannot distinguish between correlated and uncorrelated datasets. This is because ED only compares climatological means, variable per variable. The potential correlation between variables is not accounted for when calculating the distances.

The MD method results are displayed in Figure 3b and are expected to outperform the results obtained using the ED method in identifying the best analog from samples according to their dependencies between variables. This expectation stems from its use of a correlation matrix derived from the focal location (R correlation matrix in Equation 2). Figure 3b shows that the MD method differs from the ED one only when the focal point belongs to categories with dependent variables (columns a and c). In those columns indeed, lower distances are calculated when analogs are looked for within dependent samples (rows a and c), while relatively higher distances are obtained when analogs are searched for in the independent samples (rows *b* and d). When the focal point belongs to some independent categories (columns *b* and d) ED and MD distances are similar. Thus, the MD method is not more capable than the ED method to distinguish correlated samples from uncorrelated ones. This is because MD uses the correlation information only from the focal location, and not from both the focal and the candidate locations.

Table 2
Distribution of the Best Analog (Only 1) Category (Rows a to d) for Each Dataset

		Number of selected best analogs from each category											
		ED				MD				WD			
Synthetic data categories		(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
Focal Sample	**(a)** dependent-gaussian	36	22	29	13	37	23	25	15	94	0	6	0
	(**b**) independent-gaussian	20	31	17	32	20	27	18	35	5	83	2	10
	(**c**) dependent-skewed	26	14	36	24	24	15	37	24	4	1	93	2
	(**d**) independent-skewed	16	29	16	39	17	29	17	37	0	2	10	88

Note. Each category has 100 samples and their best analog can be found in any of the 4 categories (columns a to d), for each method to compute the distances. The sum of each row for each method thus equals the size of the sample (100 data).

In contrast, the results obtained using the WD method (Figure 3c) demonstrate a clear clustering of the categories. The majority of the minimum distances are observed between samples from the same category. This implies that, if a selected sample is characterized by correlated bivariate data with a Gaussian distribution, it is most likely that the sample with the smallest distance to the selected one also possesses similar characteristics. Furthermore, it is evident that the correlation between variables has a significant impact on the calculation of distances when comparing the results obtained in the category of correlated variables (the darker blue grids on the diagonal of columns/rows a and c), and in the category of uncorrelated variables (the lighter blue grids on the diagonal of columns/rows *b* and d). The overall findings indicate that the WD method demonstrates enhanced utility for determining distances between distributions, considering both the dependence between variables and the types of distributions.

Table 2 shows the category of the nearest sample for each method used (ED, MD, or WD) across all samples in each category. For instance, for the first row, which corresponds to the dependent-gaussian samples, ED and MD both split the best analogs over all four categories, favoring the dependent samples (a and (c) but finding quite a significant number of analogs in the samples with skewed distribution (b and (d). Moreover, within the dependent samples (i.e., a and (b), ED and MD are only slightly favoring the Gaussian compared to the skewed data (37 vs. 25 for MD, 36 vs. 29 for ED), showing the relatively poor distinction between the different categories. On the other hand, the WD method (a) finds most of the analogs (94 out of 100) in the original distribution, (b) finds a very small number of analogs (6 out of 100) in the other dependent distribution, and (c) finds no analog at all in both independent distributions (b and (d). The similar results can be seen for the other categories (i.e., in other rows a, c, and (d). This shows that only the WD method (more than 83 out of 100 for each category) is able to find the best analog from the same exact category that the focal sample belongs to, in a robust way.

In addition to this comparison, we provide, in the supplementary document (Figure S1 in Supporting Information S1 and Table S1 in Supporting Information S2), a similar comparison using 16 variables (instead of 2 and by using the same methodology given in Section 3.1. The covariance matrix, mean and standard deviations of 16 variable synthetic data are given in Tables S2 and S3 in Supporting Information S2), that is, a number of variables similar to our real-case application (Section 4). This 16-variables comparison provides results and conclusions similar to the 2-variables comparison discussed above.

Based on these results, in the following, we only focus on the WD method, to search for climate analogs. However, to illustrate the differences in climate analog (CA) results among the three methods, we have included maps of Paris's CA calculated by each method using Sigma Dissimilarity (Figure S2 in Supporting Information S1).

4. Results From the Climate Simulations

4.1. Where Are the Analogs for the City of Paris Located?

We start our analysis focusing on Paris and looking at the grid points "today", throughout the globe, where the climate today looks like Paris's climate "tomorrow", with "tomorrow" being calculated from 5 GCMs in the mid-future (MF, 2041–2070) for the SSP370 scenario, and "today" being the reference period HS, 1981–2010. The maps showing the calculated WD are displayed in Figure 4. The maps also show, in a small plot zooming over

10.1029/2024EF004972

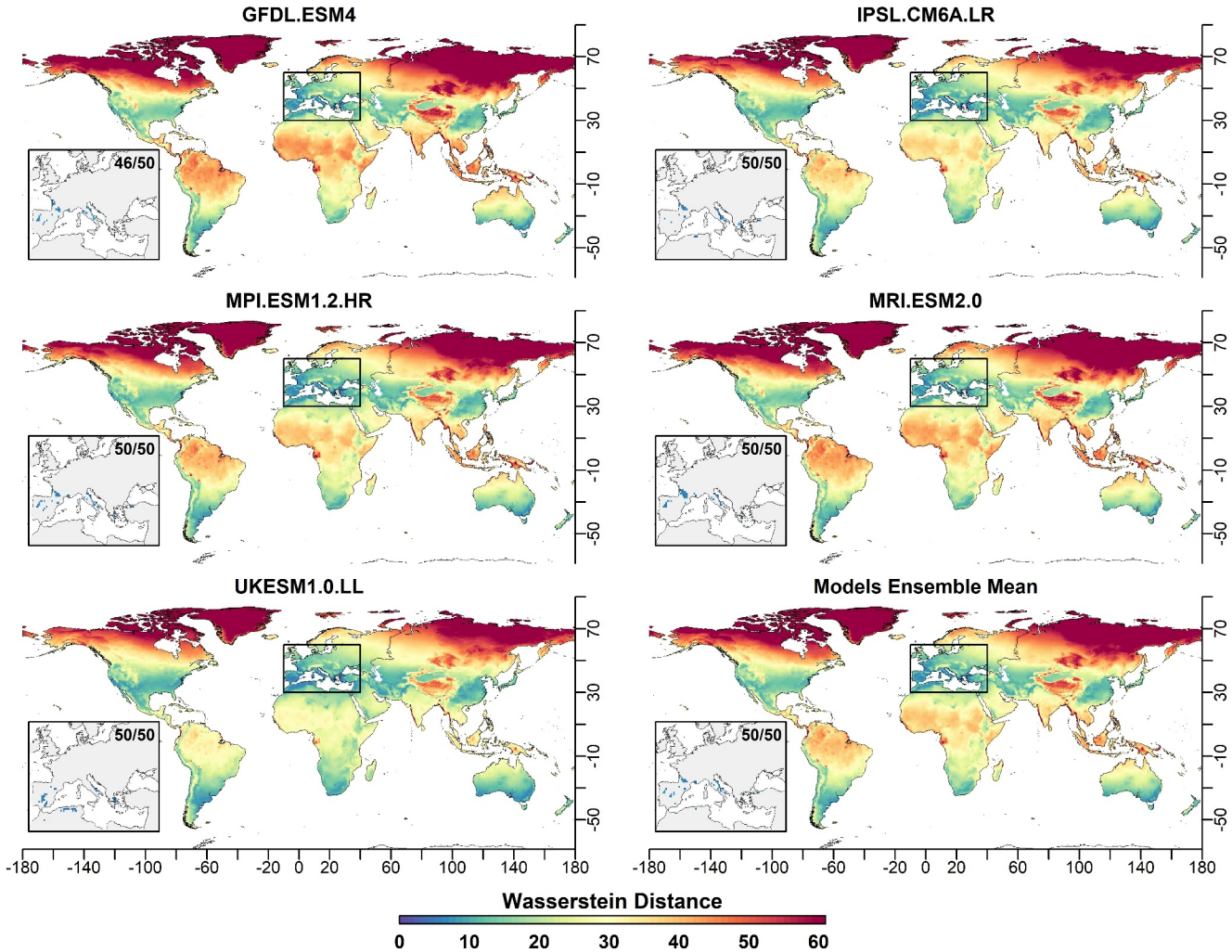

Figure 4. Wasserstein Distances (WD, dimensionless) calculated between the climate of Paris in the mid-future (2041–2070) and today's climate (1981–2010) everywhere on the globe, for the SSP370 scenario, and the 5 GCMs. The bottom right figure is the ensemble mean of all calculated WDs. In the inserts that show the map of Europe, blue dots are the analog locations, among the 50 best, that are located in Europe. The values (NN/50) give the number of analogs, among the 50 best, that are located in this zoom. We see that all 'NN' are 50, confirming that the 50 best analogs are all located within the zoom except the GFDL model (46/50) for this time horizon and level of warming.

Europe, the analogs, among the 50 best ones (i.e., grid points with the 50 lowest WD values), that are located in Europe. The lower the WD value (bluest colors, WD < 10), the more similar the climate is to the one that Paris will experience in the mid future. The maps show that the best analogs remain located within the temperate climate zone (between the 30° N and 50° N latitudes) for all models, and the 50 best ones can be found in southern Europe and the northern part of north Africa. For the UKESM model, relatively low WD values (<20) are also found in more southern regions, implying that future climate conditions in Paris are more likely to be similar to hotter regions than in the other four models. For UKESM the 50 best analogs are located at more southerly positions (Spain and north-eastern Algeria) than for the other GCMs. However, for the ensemble mean, the 50 best analogs are located in southern France, Italy, the Pyrénées and Spain.

In the supplementary material (Figure S2 in Supporting Information S1) we show that, although both other methods (ED and MD) also show analog locations south and south west of Paris, there are two major differences with WD. First, ED and MD find that today's climate in Paris and Bordeaux (in Nouvelle Aquitaine) are very similar, with a sigma dissimilarity value smaller than 1. In reality, both climates today are quite different, as captured by the WD method. Second, when the climate warms, the sigma dissimilarity values for ED and MD notably differ from those obtained using WD. Particularly, only the WD method suggests that no analog locations

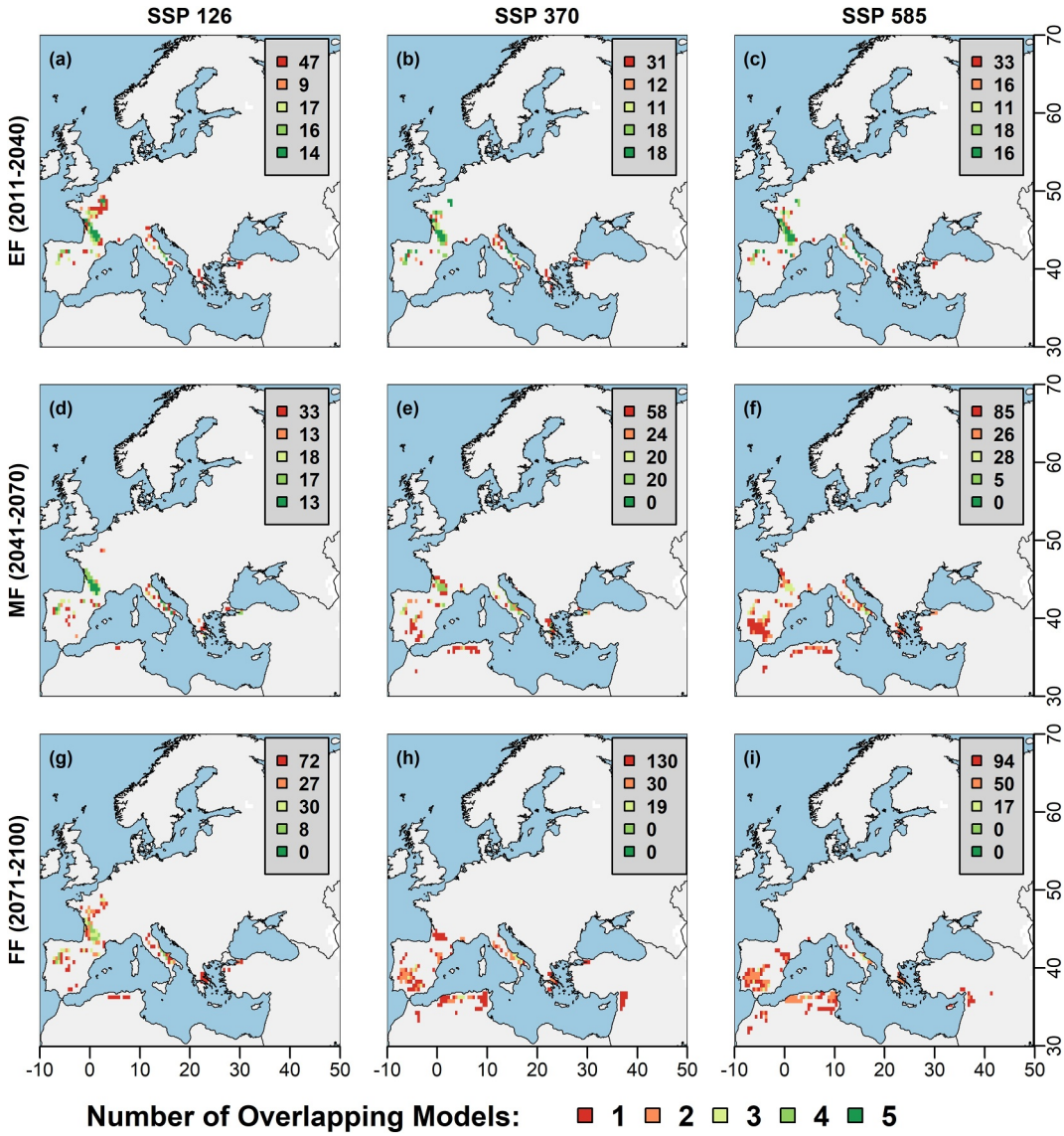

Figure 5. Agreement between the 5 GCMs in projecting the locations of the best 50 climate analogs for Paris with the WD method, for the 3 scenarios (in column) and 3 future time-periods (in rows; EF: Early Future, MF: Mid Future, FF: Far Future). Colors refer to the number of models that agree on a specific location: red when only 1 model finds a CA at this location, orange when 2 models agree, pale green when 3 models agree, light green when 4 models agree and dark green when all 5 models agree. Numbers on each map for each color refer to the number of locations showing similar model agreement.

with less than 1 sigma dissimilarity will exist in the future. This means that, even for the closest analog, the distribution of climate variables will be substantially different and this is not caught by ED nor by MD. Such dissimilar climates, even for the closest analog, may have significant implications for adaptation.

4.2. Do GCMs and Scenarios Agree on the Locations of Analogs for Paris?

We now investigate the level of agreement among the five climate models, the three socio-economic scenarios and the three chosen time horizons in the exact location of the best 50 analogs for Paris with the WD method (Figure 5). If all models agree on these locations (perfect agreement), then each European map should only include 50 dark green grid points. If all models disagree (complete disagreement), then the maps should include 5 × 50 (i.e., 250) red grid points. In other words, Figure 5e represents the overlapping of the inserted panels

(spatial distribution of the best 50 analogs located within the study area) of all individual model results from Figure 4.

The best agreement between all 5 models is found for the EF period, whatever the scenario Figures 5a–5c. For all scenarios, agreement between GCMs decreases with time, that is with increased global warming (going from greenish to reddish colors on the maps). Up to 2040 (early future time horizon), scenarios can hardly be distinguished: most analogs are located south west of Paris in France, in the Center Val de Loire and eastern Pays de la Loire, as well as in southern Nouvelle Aquitaine and Occitanie. Some marginal analogs can also be found in Spain and Italy. Agreement is the largest between various models for locations in the southern part of Nouvelle Aquitaine and in Occitanie. There is consensus that by 2040, the climate in Paris is likely to resemble the current climate of those areas.

Starting from the mid-future (after 2040), agreement between GCMs worsen with the socio-economic scenario (from SSP126 to SSP370 and then to SSP585). Most analogs are found outside France for both SSP370 and SSP585 scenarios: in Spain, Italy, northern Africa (Morocco, Algeria, Tunisia), Greece, with some marginal locations in Türkiye and Syria. Although the spatial distribution of the best 50 analogs gets less and less consistent between models with increasing warming, they remain clustered in very specific areas: along the Pyrénées, south-western Spain, Portugal and coastal north Africa. These high concentrations of individual analogs suggest that these particular regions are likely to be the analog regions of future Paris' climate at more distant time horizons.

In summary, Paris climate in the future will resemble today's south western French climatic conditions within the next 20 years, while it will progressively move to more Mediterranean conditions after 2040 with increasing global warming.

4.3. Climate Analogs for Europe

The CA analysis, previously performed for the sole city of Paris (Section 4.1), is repeated for all grid points throughout the European domain in order to identify their best analogs with the WD method. As for Paris, analogs for each grid point in the studied area are looked for over the entire globe, but on terrestrial areas only. For each grid point we have performed ensemble averages of the calculated WDs using the 5 climate models (hereafter referred to as the ensemble WD). Figure 6 shows, for one socio-economic scenario (SSP370) and three time horizons in the future, three information for each European grid point: the ensemble WD value obtained for its best analog (dimensionless, Figures 6a, 6d and 6g), the geographical distance to its best analog (in km, Figures 6b, 6e and 6h) and the cardinal direction where its (Figures 6c, 6f and 6i) best analog is located.

The warmer the climate (from EF to FF), the less similar are the analogs, as illustrated in Figures 6a, 6d and 6g where the colors go from dominant blue and dark green in EF to light green, orange and even red in FF, which indicates increasing WD. This means that, the further we move toward the future (or the warmer the climate gets), the more difficult it will be to find historical locations with similar climate conditions. The increase in WD values is larger south of 35°N in northern Africa and the eastern side of the Mediterranean in Iran and Iraq. In those regions, future climate conditions are found to be less similar to historical global conditions in comparison to the rest of the study area.

In the EF period (Figure 6a), WD values greater than five (greenish colors) are mostly distributed at relatively high altitude in the Alps, in Scandinavia, Central Taurus, Caucasus and High Atlas. This indicates that changes in mountainous regions will more quickly move away from known current climate conditions than any other location. In addition to the mountainous areas, the regions over the eastern shore of the Mediterranean Sea where hot-summer Mediterranean climate is observed (Peel et al., 2007) and the Nile Delta will experience the same higher WD values. In other words, the availability of a location with analog climate conditions will be lower for these regions even in the near future.

When climate warms, that is, when we move towards the end of the century, not only analogs are less and less similar (increase in WD) but also the geographical distance (in km) between the grid point of interest and its closest analog increases (Figures 6b, 6e and 6h). In the EF, nearly all analogs can be found within a 500 km radius (three darkest blue colors), with some exceptions for which analogs are located more than 500 km away, in the north-east of Germany, in Poland and parts of north eastern Russia, and in some mountainous regions. In the MF, distances from their CA increase to more than 1,000 km (reddish colors) for the south of Spain, the western, northern and eastern coasts of the Black Sea, central Anatolia, and some countries in northern Africa and on the

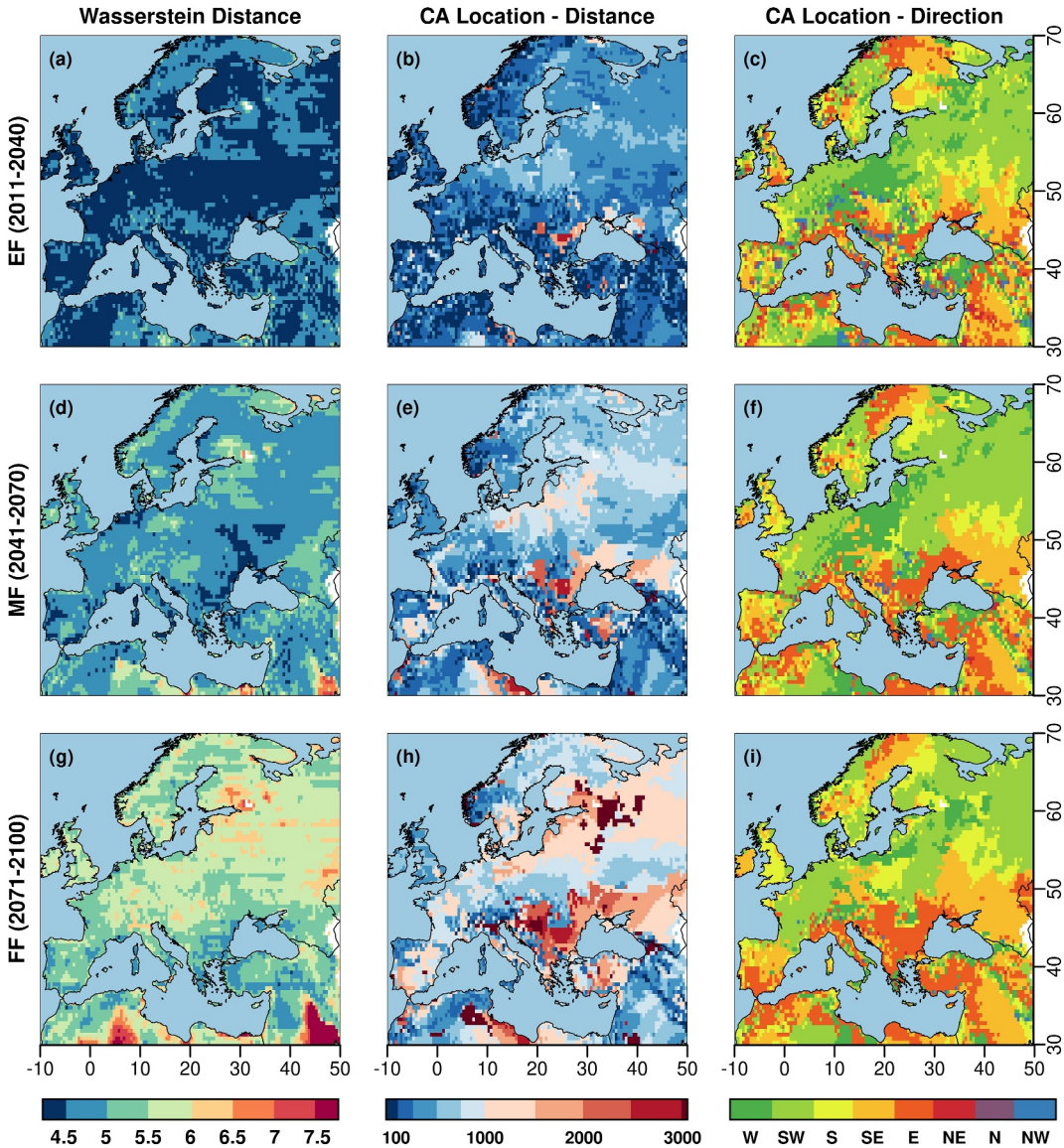

Figure 6. (a-d-g) Ensemble mean WD (dimensionless) calculated from the best analog for each grid point (b-e-h) geographical distance (km) between a grid point and its best analog (c-f-i) ordinal or cardinal direction indicating where to find the closest analog of a grid point. All results are for the SSP370 scenario and three future time horizons, the early future (EF, 2011–2040; (a-b-c), the mid future (MF, 2041–2070; (d-e-f) and the far future (FF, 2071–2100; (g-h-i).

southern edge of the Baltic Sea. In the FF, distances from their CA exceed 1,000 km in nearly all eastern Europe locations, the coastal zone of Norway, and most Mediterranean regions. In western Europe, CA locations remain mostly within a 500–750 km radius. The Rioni basin between the Greater and Lesser Caucasus is an exception with analog distances greater than 3,000 km for all periods, due to its unique climate regime (considered as humid subtropical climate using the Köppen-Geiger climate classification, Peel et al. (2007)), distinct from nearby regions even in current conditions.

The right column of Figures 6c–6f–6i shows that, whatever the targeted time period or whatever the level of warming, almost all climate analogs of our selected domain are located southward, from south-west to south-east (light green, yellow, light orange). Some climate analogs can be found east (dark orange) or west (dark green) along the same latitude, especially in the EF. Northward locations are marginal and found essentially in the earliest future (Figure 6g). This is consistent with the warming of all regions in the future. For many European regions, EF analogs are located south-west or west while when moving forward in time, their locations become

south and south-east, that is not only hotter but also drier and more continental. The climate analogs found on most oceanic or sea facades tend to be located towards the east. The direction of analogs remains eastward only over the Balkans during all study periods. The best climate analog can also be outside of the study area, such as in some grid points in Spain and Russia in the far future. In these cases, the calculated analog locations indicate distances greater than 3,000 km to the west, beyond the Atlantic Ocean (Figures 6h and 6i). The overall results are in accordance with the previous CA studies which generally concluded the CA locations are moving towards southerly directions for the cities located in the northern hemisphere.

In addition, the same figures prepared for the other two socio-economic scenarios (SSP126 to SSP585) are given in supplementary documents (Figures S3 and S4 in Supporting Information S1). Messages are relatively similar to the ones discussed above: when climate warms, WD values and spatial distances increase, and analogs move globally to the south, with a dominant eastward direction south of 50°N. In SSP126 however, WD increase from EF to FF with no further change in geographical distance and direction, implying less similar analogs. In SSP585, WD and geographical distances increase significantly from EF to FF suggesting that the term "analog" may not be appropriate any more.

4.4. Where Can We Find the Analogs of Big Capital Cities?

In Table 3, the best climate analogs for the selected capitals from the study domain are given at three future horizons. The results are obtained from the best CA grid point based on the ensemble mean of all GCMs and under the SSP370 scenario (Figures 6a, 6d and 6g). It is important to note that the method used to identify the analog cities is based on finding the city closest to the best analog grid point, rather than relying solely on analog calculations between cities. Paris is projected to experience future climate conditions similar to those currently experienced in Toulouse in the south west of France, Ancona at the shore of the Adriatic Sea in the northern part of Italy, and Potenza in the southern part of Italy, inland, at respectively the early, mid, and late future periods of the century. Paris is the analog of Berlin's climate in the mid future, and that of Brussels in the near and mid future.

Although making comparisons with previous studies in the literature is desired, it is important to note that variations in study design can hinder such comparisons. For example, the majority of studies generally focused on evaluating the conditions of selected cities to identify the most appropriate analog, instead of considering possibilities across all locations. Additionally, the use of different General Circulation Models (GCMs) or Regional Climate Models (RCMs) introduces further complexity for the comparison. Nonetheless, when examined in a general manner, a common observation across multiple studies is that the future climate in Europe is expected to resemble climates found in more southerly regions.

For more detailed information on CA of the selected capital cities and also CA of cities with a population greater than 250 k, please refer to the Table S4 in Supporting Information S2 (as a separate file), where readers can find the precise locations of the CA and the calculated WD values for all SSP scenarios.

5. Conclusion and Potential Future Use of Those Climate Analogs

This study aimed to show how climate will change within a selected European domain, by comparing the projected climate in Europe to historical (recent past) climatic conditions everywhere on the globe, on land areas only. This is known as the climate analog method that allows one to look away and find a place that already experiences the climate its "home" will experience in the future.

To do so, we have used the Wasserstein distance (WD) method that offers a unique contribution to the field by using the complete multivariate distributions, instead of some statistical parameters only, and thus accounting for dependencies between variables in the analog calculation process. We first compared the WD method with the Euclidean distance (ED) and Mahalanobis distance (MD) methods, that are more traditionally employed for climate analogs, using synthetic data we have generated on purpose. While such pronounced differences may not always be observed in real-world scenarios compared to those obtained with synthetic data, our results demonstrate the added value of the WD method in selecting analogs more accurately than both the ED and WD methods, based on a comparison of the calculated CA from all methods in the Paris example.

Regarding the terminology of dependencies between climate variables, the MD method addresses variance inflation caused by correlations at the focal location. In contrast, our study emphasizes the WD method's ability to account for dependencies and distributions of climate variables. For instance, if temperature and precipitation are

Table 3
The Nearest Cities to the Model Ensemble's CA of the Selected Capitals Located in the Study Domain (for SSP370)

	The nearest city to the best CA grid point location		
Reference location	EF (2011–2040)	MF (2041–2070)	FF (2071–2100)
Algiers/Algeria	Oujda-Angad/Morocco	Mahdia/Tunisia	Tripoli/Libya
Amman/Jordan	Beersheba/Israel	At Tafilah/Jordan	Tabuk/Saudi Arabia
Amsterdam/Netherlands	Rouen/France	Rennes/France	Nantes/France
Ankara/Türkiye	Ankara/Türkiye	Gorgan/Iran	Qazvin/Iran
Athens/Greece	Paphos/Cyprus	Paphos/Cyprus	Mahdia/Tunisia
Baghdad/Iraq	Al Kut/Iraq	Kuwait City/Kuwait	Al Farwaniyah/Kuwait
Baku/Azerbaijan	Baku/Azerbaijan	Gorgan/Iran	Semnan/Iran
Beirut/Lebanon	Al Qunaytirah/Syria	Dar'a/Syria	Madaba/Jordan
Belgrade/Serbia	Goeycay/Azerbaijan	Gorgan/Iran	Bojnurd/Iran
Berlin/Germany	Mainz/Germany	Paris/France	Bologna/Italy
Brussels/Belgium	Paris/France	Paris/France	Toulouse/France
Bucharest/Romania	Gorgan/Iran	Gorgan/Iran	Bojnurd/Iran
Budapest/Hungary	Szeged/Hungary	Belgrade/Serbia	Novyy Karanlug/Azerbaijan
Cairo/Egypt	El-Tor/Egypt	Hurghada/Egypt	Luxor/Egypt
Copenhagen/Denmark	Schwerin/Germany	Lille/France	Potenza/Italy
Damascus/Syria	Baalbek/Lebanon	As-Suwayda/Syria	Homs/Syria
Dublin/Ireland	Dublin/Ireland	Saint Savior/Guernsey	Torteval/Guernsey
Kyiv/Ukraine	Ialoveni/Moldova	Slobozia/Romania	Gorgan/Iran
London/United Kingdom	Rouen/France	Rennes/France	Valladolid/Spain
Madrid/Spain	Tissemsilt/Algeria	Tebessa/Algeria	Ouled Djellal/Algeria
Minsk/Belarus	Rivne/Ukraine	Bratislava/Slovakia	Timisoara/Romania
Moscow/Russia	Homyel/Belarus	Chernihiv/Ukraine	Zaporizhzhya/Ukraine
Oslo/Norway	Sarpsborg/Norway	Trento/Italy	Bremen/Germany
Paris/France	Toulouse/France	Ancona/Italy	Potenza/Italy
Prague/Czechia	Tatabanya/Hungary	Szeged/Hungary	Novi Sad/Serbia
Rabat/Morocco	Casablanca/Morocco	Al Khums/Libya	Misratah/Libya
Rome/Italy	Vlore/Albania	Jijel/Algeria	Skikda/Algeria
Sofia/Bulgaria	Mitrovice/Kosovo	Yambol/Bulgaria	Goeycay/Azerbaijan
Stockholm/Sweden	Vejle/Denmark	Kiel/Germany	Chiesanuova/San Marino
Tbilisi/Georgia	Terter/Azerbaijan	Zangilan/Azerbaijan	Novyy Karanlug/Azerbaijan
Tripoli/Libya	Zuwarah/Libya	Laayoune/W.Sahara	Medina/Saudi Arabia
Vienna/Austria	Resita/Romania	Bologna/Italy	Thessaloniki/Greece
Warsaw/Poland	Potsdam/Germany	Mainz/Germany	Belgrade/Serbia
Yerevan/Armenia	Kapan/Armenia	Yeghegnadzor/Armenia	Orumiyeh/Iran

correlated at a focal location, the WD method ensures that the analog location exhibits similar dependencies. The MD and ED methods do not provide this capability, as they do not account for such dependencies between variables at both focal and analog locations.

We have then applied our WD method to a set of 4 climate variables (rainfall, mean, maximum and minimum ambient air temperature) on four seasons (winter, spring, summer and fall), which means a set of 16 climate variables. Climate analogs are computed with the WD method using the distribution of these 16 variables over 30 year periods. It is thus the interannual distribution and the dependence between those variables, during a

climatological time-period (historical, early future, mid future or far future), that are compared from one grid point over future times to other grid-cells during the historical time to find the best match. We have examined three socio-economic scenarios and five global climate models, all projections being available following the CMIP6 exercise, downscaled and bias-corrected within the ISIMIP project. We have also performed ensemble calculations, averaging the WDs computed by the 5 climate models from which we have derived the best analogs for major European cities. This is different from what most studies have done so far, creating ensemble climate conditions (averaging climate variables) before computing the distances. This approach provides a clearer picture of the potential range of climate futures and allows us to assess the level of agreement or disagreement among the models, offering more robust and reliable insights for understanding climate analogs.

Our results show that, as the climate gets warmer, climate analogs for Europe will generally be found south-westward as well as westward for most of the northern part of Europe, while south-eastward and eastward movements are found in southern and eastern Europe. The warmer the climate is, the bigger the eastward component, suggesting an increase in continentality. In some isolated regions in central and eastern Europe, there are pure southward movements. Similarities between future and historical climates decrease with the level of warming, and the geographical distance from the best analogs increases and can exceed more than 1,000 km. This means that the warmer the climate gets, the more difficult it will be to find useful analogs elsewhere, based on our selected four seasonal climate variables.

What conclusions can be drawn from these results? What precise information should the authors of this paper provide to the readers? Being able to anticipate what climate could happen in the future, by an immersive experience, can be useful to accelerate adaptation and transitions. The immersive experience means a possibility to "go there, see, and feel" what the climate is like, and "learn" from what is being done at this analog place, how people live, grow crops and which crops, how they manage water resources, etc. This means there is an additional step to this paper, that is a more thorough analysis, per region, to clearly identify where the analogs are, and how robust they are. We have shown for example, that the analogs for Paris (and its surroundings) remain clustered despite different locations projected by the five climate models. More in-depth analyses may help choose some preferential ones based on soils, altitude, exposure to dominant winds, etc., before starting to suggest adaptation solutions to sectorial activities.

In this study, we offered climate analogs in a general manner using temperature and precipitation variables from each season. It offers a first insight into the effect of climate change for various sectors as well as in our daily routines and practices. Furthermore, to conduct a comprehensive sector-specific analysis, various variables relevant to each sector should be employed. For instance, in agriculture, variables like soil moisture and evapotranspiration, while in the energy sector, factors such as global horizontal irradiation or wind speed can be considered. Hence, this study is a brick to go further in the understanding and appropriation of future climate changes and to build the adaptation strategies that must be implemented.

Data Availability Statement

The *R* code and sample seasonal climate data to calculate and compare climate analogs from all methods at user-defined locations within the study area available in Bulut (2024). The ISIMIP3b climate forcing input data are available in Lange and Büchner (2021).

Acknowledgments
This study is part of the "RechErche d'analogGues climAtiques pour sélectionneR Demain (REGARD)/ Searching for Climate Analogs to Select Tomorrow" project and funded by Le Fonds de Soutien à l'Obtention Végétal (FSOV)/The Support Fund for Plant Breeding (Project no: FSOV 2020 S-REGARD). MV acknowledges support from the "COESION" project funded by the French National program LEFE (Les Enveloppes Fluides et l'Environnement), as well as support from the Swiss national program FNS "Combine" project. This paper was supported by the Agence Nationale de la Recherche—France 2030, as part of the PEPR TRACCS programme under Grants ANR-22-EXTR-0002-DIALOG, ANR-22-EXTR-0004-DEMOCLIMA, and ANR-22-EXTR-0005-EXTENDING.

References

Arlt, D., Hoppe, I., & Wolling, J. (2011). Climate change and media usage: Effects on problem awareness and behavioural intentions. *International Communication Gazette*, *73*(1), 45–63. https://doi.org/10.1177/1748048510386741

Bastin, J. F., Clark, E., Elliott, T., Hart, S., van den Hoogen, J., Hordijki, I., et al. (2019). Understanding climate change from a global analysis of city analogues. *PLoS One*, *14*(7), e0217592. https://doi.org/10.1371/JOURNAL.PONE.0217592

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., et al. (2020). Presentation and evaluation of the IPSL-CM6A-LR climate model. *Journal of Advances in Modeling Earth Systems*, *12*(7), e2019MS002010. https://doi.org/10.1029/2019MS002010

Bulut, B. (2024). Bulut-etal_2024_EarthsFuture:What will the European climate look like in the future? A climate analog analysis accounting for dependencies between variables [collection]. *Zenodo*. https://doi.org/10.5281/zenodo.13763740

Burke, K. D., Williams, J. W., Chandler, M. A., Haywood, A. M., Lunt, D. J., & Otto-Bliesner, B. L. (2018). Pliocene and Eocene provide best analogs for near-future climates. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(52), 13288–13293. https://doi.org/10.1073/PNAS.1809600115

Dunne, J. P., Horowitz, L. W., Adcroft, A. J., Ginoux, P., Held, I. M., John, J. G., et al. (2020). The GFDL earth system model version 4.1 (GFDL-ESM 4.1): Overall coupled model description and simulation characteristics. *Journal of Advances in Modeling Earth Systems*, *12*(11), e2019MS002015. https://doi.org/10.1029/2019MS002015

Fitzpatrick, M. C., & Dunn, R. R. (2019). Contemporary climatic analogs for 540 North American urban areas in the late 21st century. *Nature Communications*, *10*(1), 1–7. https://doi.org/10.1038/s41467-019-08540-3

Genz, A., Bretz, F., Miwa, T., Xuefei, M., Leisch, F., Scheipl, F., & Hothorn, T. (2021). mvtnorm: Multivariate normal and t distributions. https://CRAN.R-project.org/package=mvtnorm

Grenier, P., Parent, A. C., Huard, D., Anctil, F., & Chaumont, D. (2013). An assessment of six dissimilarity metrics for climate analogs. *Journal of Applied Meteorology and Climatology*, *52*(4), 733–752. https://doi.org/10.1175/JAMC-D-12-0170.1

Halady, I. R., & Rao, P. H. (2010). Does awareness to climate change lead to behavioral change? *International Journal of Climate Change Strategies and Management*, *2*(1), 6–22. https://doi.org/10.1108/17568691011020229

IPCC. (2022a). Summary for policymakers. In P. R. Shukla, J. Skea, R. Slade, A. Al Khourdajie, R. van Diemen, et al. (Eds.), In: *Climate change 2022: Impacts, adaptation and vulnerability. Contribution of working group II to the sixth assessment report of the intergovernmental panel on climate change*. Cambridge University Press. https://doi.org/10.1017/9781009325844.001

IPCC. (2022b). Summary for policymakers. In R. S. A. A. K. R. van, D. D. M. M. P. S. S. P. V. R. F. M. B. A. H. G. L. S. L. J. M. P. R. Shukla, & J. Skea (Eds.), *Climate change 2022: Mitigation of climate change. Contribution of working group III to the sixth assessment report of the intergovernmental panel on climate change*. Cambridge University Press. https://doi.org/10.1017/9781009157926.001

Khatibi, F. S., Dedekorkut-Howes, A., Howes, M., & Torabi, E. (2021). Can public awareness, knowledge and engagement improve climate change adaptation policies? *Discover Sustainability*, *2*(1), 1–24. https://doi.org/10.1007/S43621-021-00024-Z

King, A. D. (2023). Identifying historical climate changes in Australia through spatial analogs. *Environmental Research Letters*, *18*(4), 044018. https://doi.org/10.1088/1748-9326/acc2d4

Lange, S., & Büchner, M. (2021). ISIMIP3b bias-adjusted atmospheric climate input data (v1.1). In *ISIMIP repository*. ISIMIP Repository. [Dataset]. https://doi.org/10.48364/ISIMIP.842396.1

Lee, T. M., Markowitz, E. M., Howe, P. D., Ko, C. Y., & Leiserowitz, A. A. (2015). Predictors of public climate change awareness and risk perception around the world. *Nature Climate Change*, *5*(11), 1014–1020. https://doi.org/10.1038/nclimate2728

Leonard, M., Westra, S., Phatak, A., Lambert, M., van den Hurk, B., Mcinnes, K., et al. (2014). A compound event framework for understanding extreme impacts. *Wiley Interdisciplinary Reviews: Climate Change*, *5*(1), 113–128. https://doi.org/10.1002/WCC.252

Lorenzoni, I., & Pidgeon, N. F. (2006). Public views on climate change: European and USA perspectives. *Climatic Change*, *77*(1), 73–95. https://doi.org/10.1007/S10584-006-9072-Z

Lotterhos, K. E., Láruson, Á. J., & Jiang, L. Q. (2021). Novel and disappearing climates in the global surface ocean from 1800 to 2100. *Scientific Reports*, *11*(1), 15535. https://doi.org/10.1038/s41598-021-94872-4

Mahony, C. R., Cannon, A. J., Wang, T., & Aitken, S. N. (2017). A closer look at novel climates: New methods and insights at continental to landscape scales. *Global Change Biology*, *23*(9), 3934–3955. https://doi.org/10.1111/GCB.13645

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., et al. (2019). Developments in the MPI-M earth system model version 1.2 (MPI-ESM1.2) and its response to increasing CO2. *Journal of Advances in Modeling Earth Systems*, *11*(4), 998–1038. https://doi.org/10.1029/2018MS001400

O'Neill, S., & Nicholson-Cole, S. (2009). Fear won't do it. *Science Communication*, *30*(3), 355–379. https://doi.org/10.1177/1075547008329201

Owen, G. (2020). What makes climate change adaptation effective? A systematic review of the literature. *Global Environmental Change*, *62*, 102071. https://doi.org/10.1016/J.GLOENVCHA.2020.102071

Peel, M. C., Finlayson, B. L., & McMahon, T. A. (2007). Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences*, *11*(5), 1633–1644. https://doi.org/10.5194/HESS-11-1633-2007

Pugh, T. A. M., Müller, C., Elliott, J., Deryng, D., Folberth, C., Olin, S., et al. (2016). Climate analogues suggest limited potential for intensification of production on current croplands under climate change. *Nature Communications*, *7*(1), 1–8. https://doi.org/10.1038/ncomms12608

Qu, W., & Zhang, Z. (2020). Mnonr: A generator of multivariate non-normal random numbers. https://CRAN.R-project.org/package=mnonr

R Core Team. (2021). R: A language and environment for statistical computing. Retrieved from https://www.R-project.org/

Robin, Y., Vrac, M., Naveau, P., & Yiou, P. (2019). Multivariate stochastic bias corrections with optimal transport. *Hydrology and Earth System Sciences*, *23*(2), 773–786. https://doi.org/10.5194/HESS-23-773-2019

Robin, Y., Yiou, P., & Naveau, P. (2017). Detecting changes in forced climate attractors with Wasserstein distance. *Nonlinear Processes in Geophysics*, *24*(3), 393–405. https://doi.org/10.5194/NPG-24-393-2017

Rohat, G., Goyette, S., & Flacke, J. (2018). Characterization of European cities' climate shift – An exploratory study based on climate analogues. *International Journal of Climate Change Strategies and Management*, *10*(3), 428–452. https://doi.org/10.1108/IJCCSM-05-2017-0108

Rüschendorf, L. (1985). The Wasserstein distance and approximation theorems. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *70*(1), 117–129. https://doi.org/10.1007/BF00532240

Salvadori, G., Durante, F., De Michele, C., Bernardi, M., & Petrella, L. (2016). A multivariate copula-based framework for dealing with hazard scenarios and failure probabilities. *Water Resources Research*, *52*(5), 3701–3721. https://doi.org/10.1002/2015WR017225

Schuhmacher, D., Bähre, B., Carsten, G., Hartmann, V., Heinemann, F., & Schmitzer, B. (2020). transport: Computation of optimal transport plans and Wasserstein distances. https://cran.r-project.org/package=transport

Sellar, A. A., Jones, C. G., Mulcahy, J. P., Tang, Y., Yool, A., Wiltshire, A., et al. (2019). UKESM1: Description and evaluation of the U.K. Earth system model. *Journal of Advances in Modeling Earth Systems*, *11*(12), 4513–4558. https://doi.org/10.1029/2019MS001739

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, *48*(3), 465–471. https://doi.org/10.1007/BF02293687

Villani, C. (2009). *Optimal transport: Old and new* (Vol. 338). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-71050-9

Vissio, G., Lembo, V., Lucarini, V., & Ghil, M. (2020). Evaluating the performance of climate models based on Wasserstein distance. *Geophysical Research Letters*, *47*(21). https://doi.org/10.1029/2020GL089385

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., & Schewe, J. (2014). The inter-sectoral impact model Intercomparison project (ISI–mip): Project framework. *Proceedings of the National Academy of Sciences*, *111*(9), 3228–3232. https://doi.org/10.1073/PNAS.1312330110

Williams, J. W., Jackson, S. T., & Kutzbach, J. E. (2007). Projected distributions of novel and disappearing climates by 2100 AD. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(14), 5738–5742. https://doi.org/10.1073/PNAS.0606292104

Yukimoto, S., Kawai, H., Koshiro, T., Oshima, N., Yoshida, K., Urakawa, S., et al. (2019). The meteorological research institute earth system model version 2.0, MRI-esm2.0: Description and basic evaluation of the physical component. *Journal of the Meteorological Society of Japan. Ser. II*, *97*(5), 931–965. https://doi.org/10.2151/JMSJ.2019-051