Stochastic Environmental Research and Risk Assessment (2024) 38:2261–2281
https://doi.org/10.1007/s00477-024-02678-3

ORIGINAL PAPER

A new diagram for performance evaluation of complex models

Aras Izzaddin[1] · Andreas Langousis[2] · Vincenzo Totaro[1] · Marwah Yaseen[1] · Vito Iacobellis[1]

Accepted: 29 January 2024 / Published online: 2 March 2024

Abstract

In the latest years the capacity and complexity of climate and environmental modeling has increased considerably. Therefore, tools and criteria for model performance evaluation are needed to ensure that different users can benefit from model selection. Among graphical tools, Taylor's diagram is widely used to provide evaluation and comparison of model performances, with particular emphasis on climate models. Taylor's diagram accounts for different statistical features of model outputs and observations, including correlation, variability and centered root mean square error. Not included is model bias, which is an essential feature for climate model evaluations, and it is usually calculated separately to complement the information embedded in Taylor's diagram. In this paper a new diagram is proposed, referred to as Aras' diagram, which allows for visual assessments of the correspondence between model outputs and reference data in terms of total error, correlation, as well as bias and variability ratios through an easy-to-interpret two-dimensional (2D) plot, allowing for proper weighting of different model features. The strengths of the new diagram are exemplified in a case study of performance evaluation of EURO-CORDEX historical experiment over Southern Italy using E-OBS as reference dataset, for three hydrological variables (i.e. daily precipitation, daily surface minimum temperature, and daily maximum surface temperature), and four popular climate indices (i.e. total annual precipitation, annual maxima of daily precipitation, annual minima of daily minimum temperatures, and annual maxima of daily maximum temperatures). The proposed diagram shows interesting properties, in addition to those already included in Taylor's diagram, which may help promoting climate model evaluations based on their accuracy in reproducing the climatological patterns observed in time and space.

Keywords Bias ratio · Variability ratio · Correlation · Climate models · Kling Gupta efficiency · Taylor's diagram · Performance evaluation · EURO-CORDEX experiment · E-OBS · Aras' diagram

1 Introduction

Gaining understanding of natural phenomena has been crucial for advancing knowledge and supporting societal development. Models are vital tools towards this aim, as they allow for simulation of the dynamics of natural processes in both space and time (see e.g. Chaulya and Prasad 2016). Under this setting, it is important to evaluate model performances by comparing model outputs with observations (see e.g. Legates and McCabe 1999; Flato et al. 2013; Paul et al. 2023). Typically, this step can be carried out using qualitative and quantitative measures supported by the application of criteria and benchmarks for model selection (see e.g. Moriasi et al. 2012). The evaluation of models is often an iterative process, in which the model is calibrated and tested against observed data, and then refined and tested again. Sensitivity and uncertainty analyses are also important components of performance evaluation, as they provide insights into the robustness and reliability of model predictions. However, this should be considered only a module of model evaluation, as the latter should be complemented by a detailed investigation of the accuracy of the model structure in reproducing the different processes involved (see e.g. Gupta et al. 2008; Knutti 2010; Biondi et al. 2012; Kaleris & Langousis 2017). Qualitative analysis can involve the use of graphical tools (see e.g. Kundzewicz & Robson 2004), as graphical measures may allow for direct or adjusted comparisons, depending on the transformations imposed to data

✉ Aras Izzaddin
arasbotan.izzaddin@poliba.it

[1] Department of Civil, Environmental, Land, Building Engineering and Chemistry, Polytechnic University of Bari, 70125 Bari, Italy

[2] Department of Civil Engineering, University of Patras, 26504 Patras, Greece

(see e.g. Moriasi et al. 2015). Quantitative assessment of model performances, instead, relies on computing statistical measures, aimed at providing an evaluation of the goodness-of-fit of model outputs to observations (see e.g. Ritter & Muñoz-Carpena 2013). To this end, a number of criteria has been developed, used, and critically reviewed in the scientific literature, highlighting the importance of this topic (see e.g. Moriasi et al. 2015; Krause et al. 2005) as well as the distance from reaching a rigorous procedure for model performance assessment (see e.g. Ritter and Muñoz-Carpena 2013; Baker and Taylor 2016).

Several quantitative statistical metrics can be used for evaluating the performance of models, including: (1) Bias, which measures the average difference between the model's output and observed data; (2) correlation, (3) root mean square error (RMSE), which is the square root of the average of the squared differences between the model's output and the observed data; (4) the Nash–Sutcliffe efficiency (NSE, Nash and Sutcliffe 1970), which is a measure of how well the model predicts the observed data relative to the mean of the observed data; (5) the Kling–Gupta Efficiency (KGE), proposed by Gupta et al. (2009), a measure of how well a model simulates the observed data, including explicit evaluation of correlation, variability, and bias. The latter study also showed that both NSE and RMSE provide an evaluation of model performance that is conditionally biased by correlation. Evidently, the combined use of different metrics should be retained as the most appropriate way to approach model selection (see e.g. Flato et al. 2013; Moriasi et al. 2015).

On these grounds, a practical and easy way to compare multiple models is by graphically representing different and preferentially complementary performance measures in a single diagram (see e.g. Jolliff et al. 2009). An example is Taylor's diagram (Taylor 2001), which exploits the law of cosines to display the analytical relationship between standard deviations, correlation, and centered root-mean-squared error (CRMSE) of different model outputs and reference data. Plotted in a polar coordinate system, Taylor's diagram eases comparison of multiple model outputs (represented by distinct points) by utilizing the conceptual equivalence of their distance from the origin to CRMSE. The radial and azimuthal coordinates of each point, which corresponds to a different model, allow for visual assessment of the potential reasons of the discrepancies observed between model outputs and observations. Due to its compactness in yielding an immediate assessment of multiple model performances, Taylor's diagram has been widely used in environmental sciences. Particularly popular is its use in the evaluation and comparison of climate models. The latter are fundamental tools for simulating and predicting the Earth's climate system, including its atmospheric, oceanic, and land surface components. The development and evaluation of these models is crucial for improving our understanding of the climate system, predicting future climate change, and informing policy decisions (see e.g. Deidda et al. 2013; Langousis and Kaleris 2014; Langousis et al. 2016; Emmanouil et al. 2022, 2023).

However, Taylor's diagram has well known limitations that should be considered when interpreting results. In particular, (1) it does not explicitly consider the overall bias of model outputs (see e.g. Gleckler et al. 2008; Hu et al. 2019) and (2) it exploits the root-mean-squared error (RMSE), which is a measure of fit conditional on correlation. More precisely, as shown in Appendix, selection of best performing models via minimization of the RMSE implicitly benefits those that underestimate the variability of observations, unless the correlation coefficient between model results and observations is equal to 1. With regard to point (1), please note that a model characterized by significant bias could still be assessed to provide a good fit to observations based on standard deviation and correlation. According to several performance evaluation studies, climate models heavily suffer from large biases and cannot be used directly for impact studies, such as hydrological model applications, unless they are treated using bias correction methods (see e.g. Mamalakis et al. 2017; Perra et al. 2020; Emmanouil et al. 2021, 2023).To remedy the aforementioned shortcoming, Taylor (2001) proposed a methodology to complement the original diagram with information on bias. In addition, several alternative metrics and tools were progressively developed (see e.g. Xu et al. 2016; Hu et al. 2019; Sáenz et al. 2020; Paul et al. 2023).

Evidently, the development of alternative performance evaluation approaches indicates that the choice of the performance measure to be used should be consistent with the requirements of the application (see e.g. Zhou et al. 2021). In this study, we propose Aras' diagram, a new tool which exploits in a geometrical setting the structure of the Kling–Gupta efficiency (KGE) index. The choice of KGE is motivated by several advantages over other measures used for model performance assessments. First, KGE accounts for three metrics (i.e. correlation, variability ratio and normalized bias ratio, see Sect. 2). Second, KGE is dimensionless, making it insensitive to the variables considered and their range of values. Third, KGE is a relative measure, allowing for easy comparison of model performances across different sites and time periods. Lately, KGE has become a widely used measure for model performance assessments due to its comprehensiveness, simplicity, and ease of interpretation, being used in a variety of studies, including climate change impact assessments (see e.g. Liu 2020; Agyekum et al. 2022; Ahmed et al. 2019; Castaneda-Gonzalez et al. 2018; Ta et al. 2018), water resources management, and flood forecasting (see e.g. Pechlivanidis and Arheimer 2015; Mwangi et al.

2016; Pool et al. 2018; Lamontagne et al. 2020; Brunner et al. 2021).

In order to demonstrate the attributes and skills of Aras' diagram, we use it for performance evaluation of EURO-CORDEX historical experiment over Southern Italy using E-OBS as reference dataset, for three hydrological variables (i.e. daily precipitation, daily minimum surface temperature, and daily maximum surface temperature), and 4 climate indices (i.e. total annual precipitation, annual maxima of daily precipitation, annual minima of daily minimum surface temperatures, and annual maxima of daily maximum surface temperatures).

The paper is structured as follows. In Sect. 2, the underlying theoretical background of Aras' diagram is discussed. Section 3 describes the study area and provides a brief summary of the climate model and reference data used. In Sect. 4 we describe the most important findings of the present study, and in Sect. 5 we discuss the competitive advantages of the newly developed diagram. Concluding remarks are presented in Sect. 6.

2 Theoretical foundation and construction of Aras' diagram

Assume n couplets of values $\left\{x_{o,i}, x_{m,i}\right\}_{i=1,\ldots,n}$, where subscripts o and m refer to observations and model predictions respectively, and n is the length of each time series. Keeping the same subscripts, μ_o, σ_o and μ_m, σ_m stand for the estimated mean and standard deviation of observed and simulated series, respectively, while ρ denotes their correlation coefficient.

Taylor's diagram (Taylor 2001) is based on a graphical representation of the relationship between the centered root-mean-squared error (CRMSE), standard deviation and correlation between observations and predictions:

$$CRMSE^2 = \sigma_m^2 + \sigma_o^2 - 2\sigma_m\sigma_o\rho \tag{1}$$

noting that CRMSE is linked to RMSE and BIAS as:

$$RMSE^2 = BIAS^2 + CRMSE^2 \tag{2}$$

with CRMSE, RMSE and BIAS defined as:

$$CRMSE = \left\{ \frac{1}{n}\sum_{i=1}^{n}\left[\left(x_{m,i} - \mu_m\right) - \left(x_{o,i} - \mu_o\right)\right]^2 \right\}^{\frac{1}{2}} \tag{3}$$

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}\left(x_{m,i} - x_{o,i}\right)^2\right]^{\frac{1}{2}} \tag{4}$$

and.

$$BIAS = \mu_m - \mu_o \tag{5}$$

Comparing the structure of Eq. (1) with the geometrical law of cosines:

$$d^2 = x^2 + y^2 - 2xycos\phi \tag{6}$$

a graphical representation of CRMSE, σ_m, σ_o and ρ can be obtained by setting:

$$d = \text{CRMSE},\ x = \sigma_m,\ y = \sigma_o \text{ and } \phi = \arccos(\rho) \tag{7}$$

The Kling–Gupta efficiency (KGE, Gupta et al. 2009) can be regarded as a statistical measure of the accuracy of different model simulations. It is a dimensionless index that ranges between $-\infty$ and 1, where the value of 1 indicates perfect agreement between model simulations and observed data (see e.g. Lamontagne et al. 2020).

Applying the same notation as in Gupta et al. (2009), we define the variability ratio:

$$\alpha = \frac{\sigma_m}{\sigma_o} \tag{8}$$

and the normalized bias ratio:

$$\beta = \frac{\mu_m}{\mu_o} \tag{9}$$

The Kling–Gupta efficiency is defined as:

$$KGE = 1 - \sqrt{[(\alpha - 1)]^2 + [(\beta - 1)]^2 + [(\rho - 1)]^2} \tag{10}$$

In a geometrical interpretation, Eq. (10) involves the concept of Euclidian distance from two points in a three-dimensional space. Equation (10) can also be formulated by introducing the total error, E:

$$E = (1 - KGE)^2 = [(\alpha - 1)]^2 + [(\beta - 1)]^2 + [(\rho - 1)]^2 \tag{11}$$

Equation (11) can also be interpreted as the three-dimensional formulation of the Pythagorean theorem,

$$d^2 = y^2 + x^2 + z^2 \tag{12}$$

where d=$1 - KGE$, $y = \alpha - 1,\ x = \beta - 1$ and $z = -1$.

In such a 3D representation, d is the Euclidean distance between the point representing model performance and the origin of the diagram (the ideal point) and is equal to the error measure E=1—KGE.

Gupta et al. (2009) also introduced a general "scaled version" of KGE, by incorporating into Eq. (11) the scaling factors s_α,s_β and s_r, aimed to rescale the criteria space before computing the Euclidian distance from the ideal point, as a means of weighting differently the components of the total error:

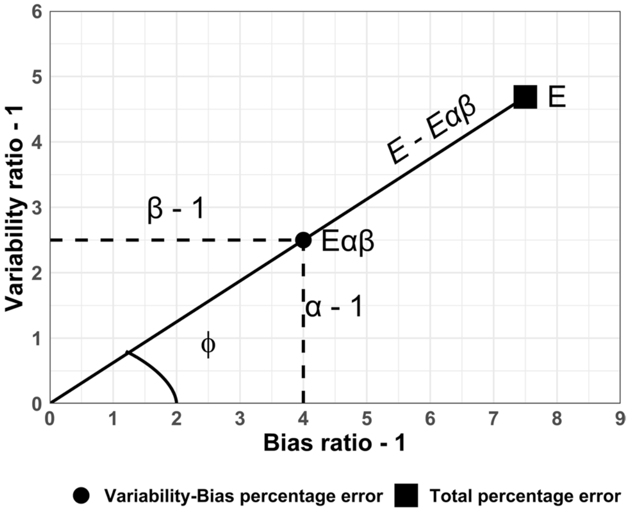

Fig. 1 Model error representation used in Aras' diagram

$$KGE_s = 1 - \sqrt{\left[s_\alpha(\alpha - 1)\right]^2 + \left[s_\beta(\beta - 1)\right]^2 + \left[s_r(\rho - 1)\right]^2} \tag{13}$$

Practical implications of the aforementioned concepts can be represented in a Cartesian coordinate system using a two-dimensional diagram, where β -1 and α -1 take values on the x and y axes, respectively; see Fig. 1 and discussion below.

To do so, one first exploits the scaled KGE_s formulation by assuming $s_\alpha = s_\beta = 1$ and $s_r = 0$. Under this setting, Eq. (11) reduces to:

$$E_{\alpha\beta} = \left(1 - KGE_{\alpha\beta}\right)^2 = (\alpha - 1)^2 + (\beta - 1)^2 \tag{14}$$

and by setting $y = \alpha - 1$, $x = \beta - 1$ and using the Pythagorean theorem, one obtains a graphical representation on a 2D Cartesian coordinate plane of the error component $E_{\alpha\beta}$ induced by the bias and variability ratios, as:

$$E_{\alpha\beta} = d^2 = y^2 + x^2 \tag{15}$$

where d denotes the Euclidean distance of point $E_{\alpha\beta}$ in Fig. 1 from the origin.

At a second step, one calculates the total percentage error E given by Eq. (11), and draws a line segment that starts from point $E_{\alpha\beta}$, has length equal to the difference $E - E_{\alpha\beta}$ (computed from Eqs. (11) and (14)), and slope:

$$\tan\phi = \frac{a - 1}{\beta - 1} \tag{16}$$

Figure 1, which forms the foundation of Aras' diagram (named by the first name of the first Author), illustrates the construction of the aforementioned error representation, where the length of the segment that links point E (i.e. total percentage error) to point $E_{\alpha\beta}$ (i.e. percentage error induced by the bias and variability ratios) provides an indication of the error component induced by discrepancies in correlation (i.e. the longer the segment the larger the disagreement). The slope of the segment indicates the relative contribution of variance and bias ratios (i.e. larger slopes correspond to smaller bias-induced errors). The endpoint of the segment, which represents the error $E_{\alpha\beta}$, is marked using a circle which is filled (empty) in case of positive (negative) correlation.

Figure 2 presents a guide to Aras' diagram. The origin of the diagram (i.e. point (0, 0)) indicates perfect model performance relative to observations. The closer the model mark (i.e. point E in Figs. 1, 2) is to the origin the better is model performance. In the diagram, circles representing % errors are drawn (e.g. 10%, 25% and 50%). Any model mark located inside the inner circle has error below 10%, any model mark located inside the second circle has error below 25% and any model mark located inside the third circle has error below 50%.

The two axes split the entire diagram into four quadrants. With respect to the y-axis, positive and negative y-values correspond to variability ratios $\alpha > 1$ and $\alpha < 1$, respectively. Therefore, the quadrants above and below the x-axis are indicative of over- and under-estimation of the variability of observed data, respectively. Similarly, the quadrants on the right and left of the y-axis correspond to bias ratios $\beta > 1$, (i.e. positive bias) and $\beta < 1$ (i.e. negative bias), respectively, indicating over- and under-estimation of the mean value of the observed data.

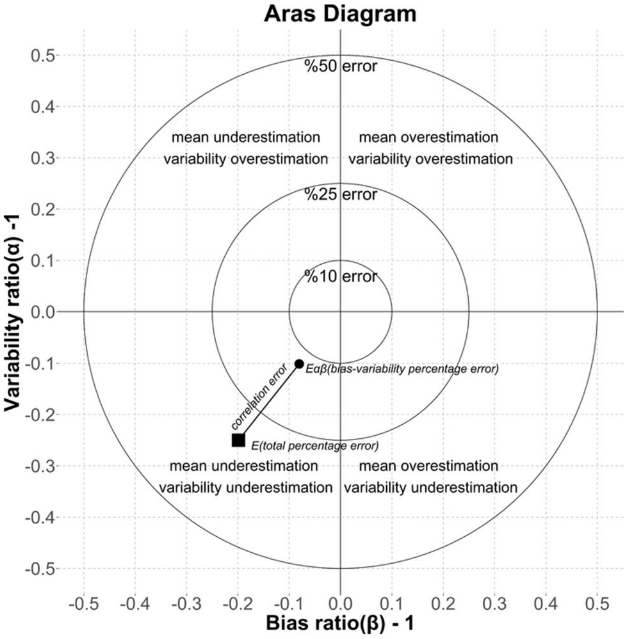

Fig. 2 Aras' diagram guide

3 Case study and data

In what follows, we use Aras' diagram to assess the accuracy of climate model outputs from EURO-CORDEX (Jacob et al. 2014) historical experiments over southern Italy, using E-OBS (Haylock et al. 2008a, b) gridded data as reference dataset.

3.1 Description of the study area

The study area is the entire southern part of the Italian peninsula (latitude 37.5–42.5 N; longitude 11–19E) (see Fig. 3) with approximate area of 90,000 km^2 crossed by the mountainous ridge of Southern Italy, and a long coastal-line on the order of 2500 km. It is characterized by a wide range of elevations, with peaks reaching up to 2912 m and it includes the "Tavoliere delle Puglie", the second larger plain in Italy. Precipitation in southern Italy varies depending on elevation and location. The regional average of mean annual rainfall ranges from 1000 to 1500 mm. Precipitation generally increases with elevation, with some of the highest elevations receiving more than 2000 mm of rain annually. Seasonality also affects precipitation which is mainly distributed in the autumn and winter months, with dry summers. The average temperature in the entire area varies between 7–9 °C in winter and 21–23 °C during summer. Temperature is influenced by elevation and proximity to the sea. The coastal areas of the region have a Mediterranean climate, with mild winters and hot summers. At higher elevations, temperatures generally decrease, and the climate becomes more continental, with colder winters and milder summers. The region is an important area for agriculture, with crops such as olives, grapes, and citrus.

3.2 Climate data

The only way to assess the magnitude and probable causes of climate change is by understanding the natural climate variability (Tsonis et al. 2017). Climate models are computer-based algorithms used to simulate and predict climate conditions under different deterministic scenarios. They are fundamental tools for simulating and predicting the Earth's climate system, including its atmospheric, oceanic, and land

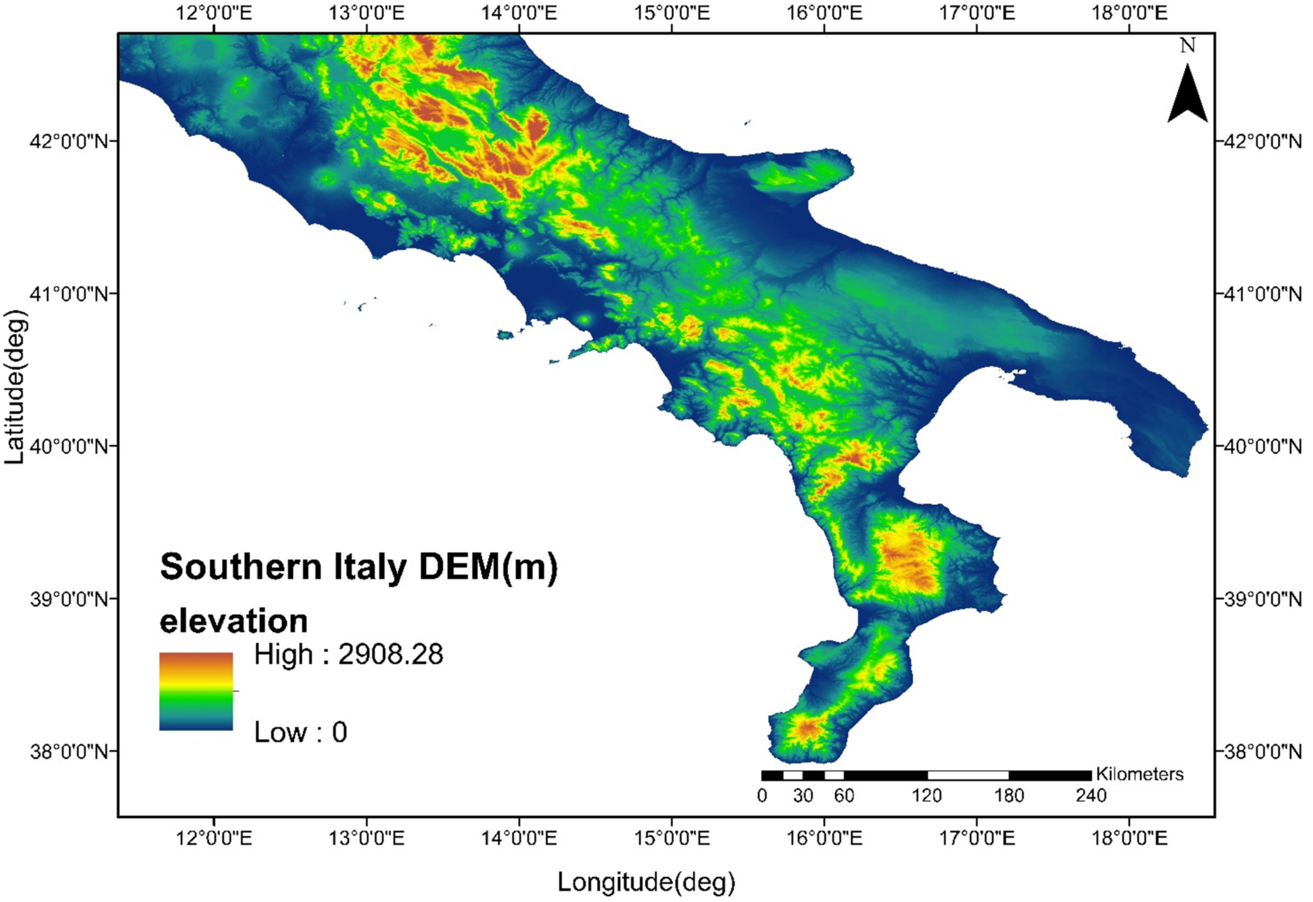

Fig. 3 Digital elevation model (DEM) of the study area

surface components. The development and evaluation of these models is crucial for improving our understanding of the climate evolution, analyze and predict climate change, and inform policy decisions (Wetterhall et al. 2009; Palatella et al. 2010).

The evaluation of climate models performance with respect to ground observations is a complex and challenging task, due to the highly nonlinear and dynamic nature of the climate system, the large spatial and temporal scales involved, and the complexity of the physical processes that govern the system. The evaluation process can be used also to diagnose and identify model strengths and weaknesses. Moreover, by evaluating the historical period and choosing the best performing climate model, one may assume that such a model should be the best option for future projections. The increasing complexity of climate models, incorporating more physical processes and components, also requires larger and powerful computational resources (Tsonis and Kirwan 2023).

A global climate model (GCM) is designed to represent the Earth's climate system on a global scale. It contains complex mathematical equations that simulate the interactions between the atmosphere, oceans, land surface, ice and other components of the climate system. GCMs are used to understand the long-term climate patterns, project future climate scenarios and assess the impact of greenhouse gas emissions on global climate (see e.g. Kirchmeier-Young and Zhang 2020; Moustakis et al. 2021). A regional climate model (RCM), on the other hand, is a more specialized model that focuses on a smaller geographical area, typically a region within a global model grid. RCMs use higher resolution and a more detailed representation of physical processes compared to GCMs, which allows for a more accurate representation of local and regional climate characteristics including orographic effects. RCMs are often driven by the output of a GCM that provides larger scale boundary conditions for the regional model. By downscaling the GCM output, RCMs can provide more localized information for assessing climate impacts, such as assessing changes in precipitation patterns, temperature extremes or regional climate variability (see e.g. Vrac et al. 2007; Fowler et al. 2007; Johnson and Sharma 2009; Mujumdar et al. 2009).

In summary, GCMs provide a global overview of the Earth's climate, while RCMs focus on smaller regions in a global context and provide more detailed and localized climate information. RCMs are often used in conjunction with GCMs to bridge the gap between global climate projections and regional climate assessments.

As mentioned in the Introduction, for the purpose of this paper a performance evaluation of EURO-CORDEX historical experiment over Southern Italy has been carried out considering historical reference periods of all available combinations of GCMs and RCMs. The investigation focuses on the hydrological variables of precipitation and temperature, by considering four climate indices: annual total precipitation (PRCP), annual maxima of daily precipitation (RX1day), annual minima of daily minimum surface temperatures (TNn) and annual maxima of daily maximum surface temperatures (TXx).

In order to test the historical experiment of EURO-CORDEX over Southern Italy, we used the E-OBS reference dataset (Haylock et al. 2008a, b): a high-resolution gridded dataset of daily climate over Europe, widely used for climate model evaluations. EURO-CORDEX (http://www.EURO-CORDEX.net/) climate model outputs at the highest available resolution (i.e. 0.11°) and for a total number of 55 GCM-RCM combinations (see Fig. 4) were compared with the reference dataset of E-OBS at the same resolution (0.11°), using two types of analysis. The first, which we refer to as "temporal analysis", was conducted by averaging the time series of model outputs and reference data over the entire study area, while preserving the temporal resolution of the original series. The second type of analysis, which we refer to as "spatial analysis", was conducted by averaging the model outputs and reference data over the entire reference period, while preserving the spatial resolution of the original fields.

4 Results

The first variable considered for performance evaluation in this study is surface temperature. The climate indices chosen are the annual minima of daily minimum temperatures, TNn, and the annual maxima of daily maximum temperatures, TXx (i.e. the lowest and highest temperature values from each year, respectively).

Figure 5 shows the climatological spatial pattern of the annual minima of daily minimum surface temperatures TNn (after averaging the corresponding time series over the entire reference period), where different colours (marks) denote different RCM models (GCM drivers). According to Taylor's diagram (Fig. 5a), all GCM–RCM model combinations display high correlations in space, on the order of 0.9. A visual ranking of model performances is feasible, looking at the model distances from the origin. The entirety of models shows a quite homogeneous trend in performance indicators, from better to lower performances, with all of them being very close to the 0.9 radius line. Aras' diagram (Fig. 5b) also indicates high correlations, as all models display short segments, but additional patterns of model performance become also visible. More precisely, no model exhibits total percentage error below 10%, while only 3 out of all GCM-RCM model combinations show total percentage error less than 30%. These 3 combinations are all driven by the same GCM (NCC-NorESM1-M), with best performing RCMs (in

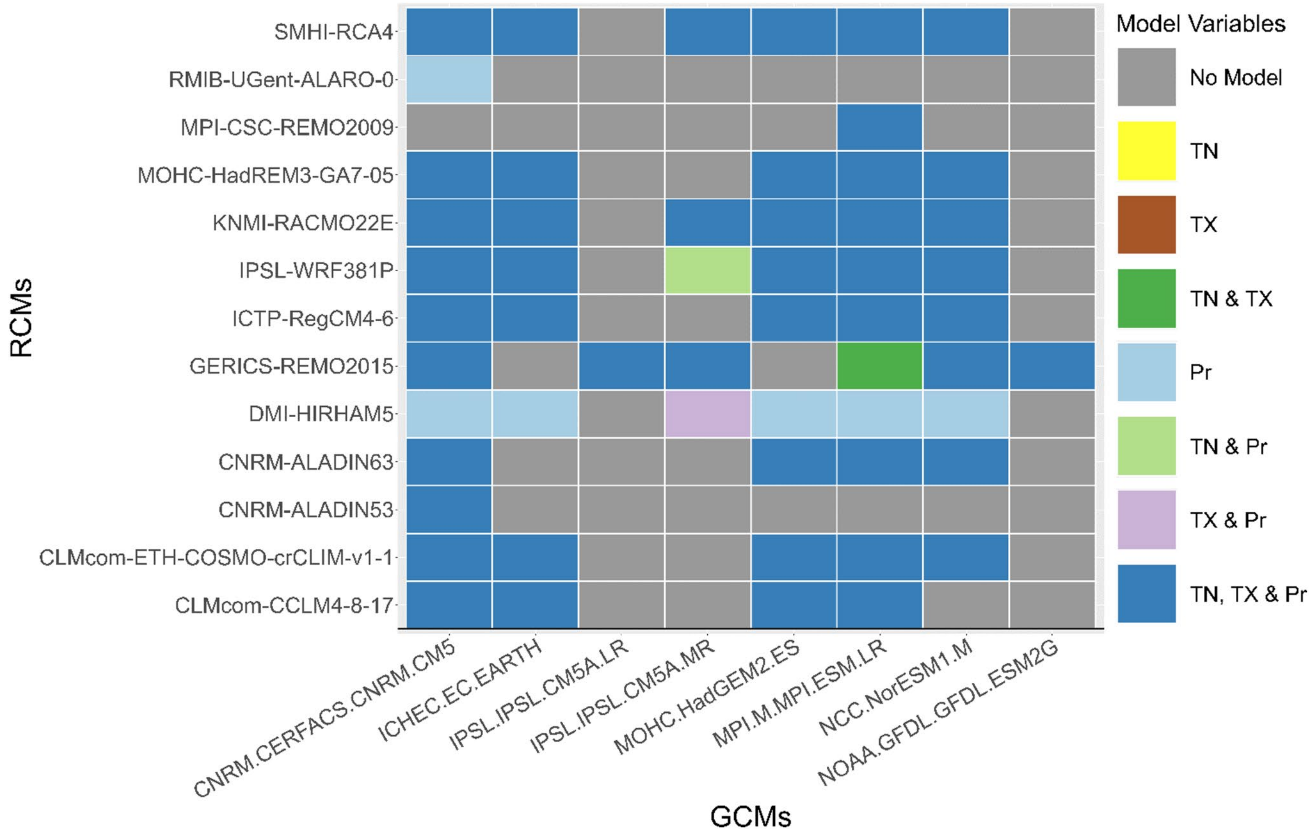

Fig. 4 EURO-CORDEX GCM-RCM combinations used in this study. TN corresponds to daily minimum surface temperature, TX to daily surface maximum temperature and Pr to daily total precipitation

descending performance order): IPSL-WRF381P, CLMcom-ETH-COSMO-crCLIM-v1-1 and ICTP-RegCM4-6. Overall, 15 model combinations display performances with total error lower than 50%. The generally good agreement of model simulations and reference data in terms of correlation in space for all GCM-RCM combinations, leads to a bias-variability percentage error ($E_{\alpha\beta}$ in Figs. 1, 2) that is very close to the total percentage error (i.e. E in Figs. 1, 2). Moreover, almost all models overestimate both the mean and variability (i.e. they are displayed in the first quadrant of the diagram, where $x = \beta$ -1 and $y = \alpha$ -1 coordinates are positive). Exceptions include RCM model CNRM-ALADIN53 driven by GCM model CNRM-CERFACS-CNRM-CM5 (the only one underestimating variability), and three model combinations with negative bias. Two of the latter have the same RCM: GERICS-REMO2015 and are driven by GCMs NCC-NorESM1-M and MPI-M-MPI-ESM-LR.

In general, patterns or clusters of model combinations sharing the same RCMs (same colour but different marks) are also very clear in Aras' diagram. For instance, similar performances are shown by RCMs: MOHC-HadREM3-GA7-05, KNMI-RACMO22E, CNRM-ALADIN63 and SMHI-RCA4, independent of the driving GCM. Worst performances in terms of bias are provided by all models sharing RCM KNMI-RACMO22E regardless of the GCM drivers.

Figure 6 shows diagrams for the spatial climatological pattern of the annual maxima of daily maximum surface temperatures TXx. In this case, the correlation values that range from somewhat below 0.8 to more than 0.9 still display an overall good agreement of model data with observations. This difference in correlation performance is strongly highlighted in Aras' diagram, where the few models characterized by correlation less (or close to) 0.8 are associated with much longer segments. Most of them are driven by the same RCM (ICTP-RegCM4-6) nested in 5 different GCMs. The same 5 models are also among the 7 models showing the lowest bias-variability percentage error (less than 10%).

Both Taylor's and Aras' diagrams coincide in indicating GERICS-REMO2015 driven by CNRM-CERFACS-CNRM-CM5 as the best performing model, having a total percentage error approximately equal to 10%. This is the best score we obtained in this study, over all indices and models evaluated. Aras' diagram also shows that: a) all models display performances in simulating the spatial climatological pattern of TXx with total percentage error lower than 50%, b) 9 models

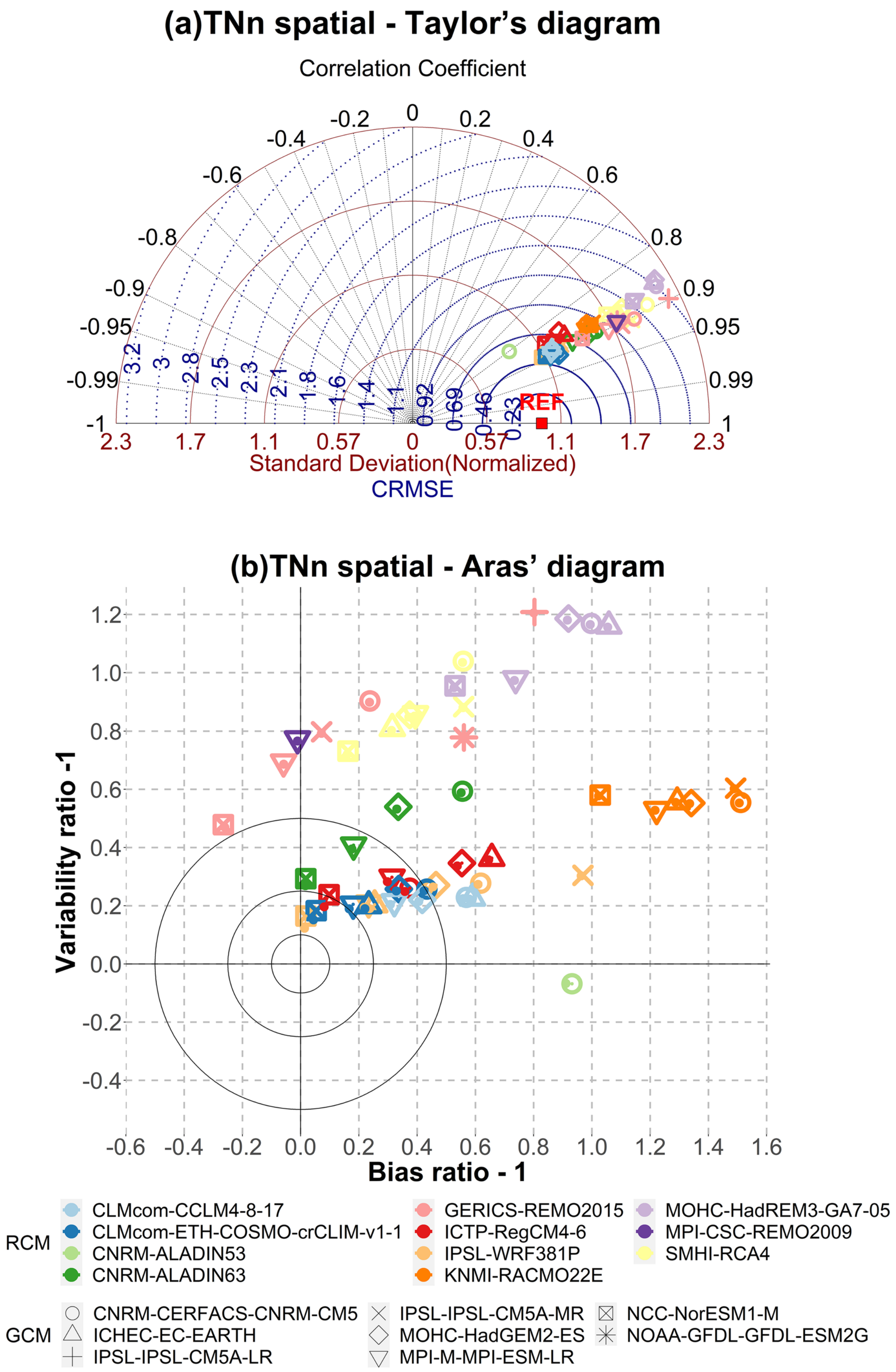

Fig. 5 Spatial climatological pattern analysis for the annual minima of daily minimum surface temperatures using: **a** Taylor's diagam, and **b** Aras' diagram

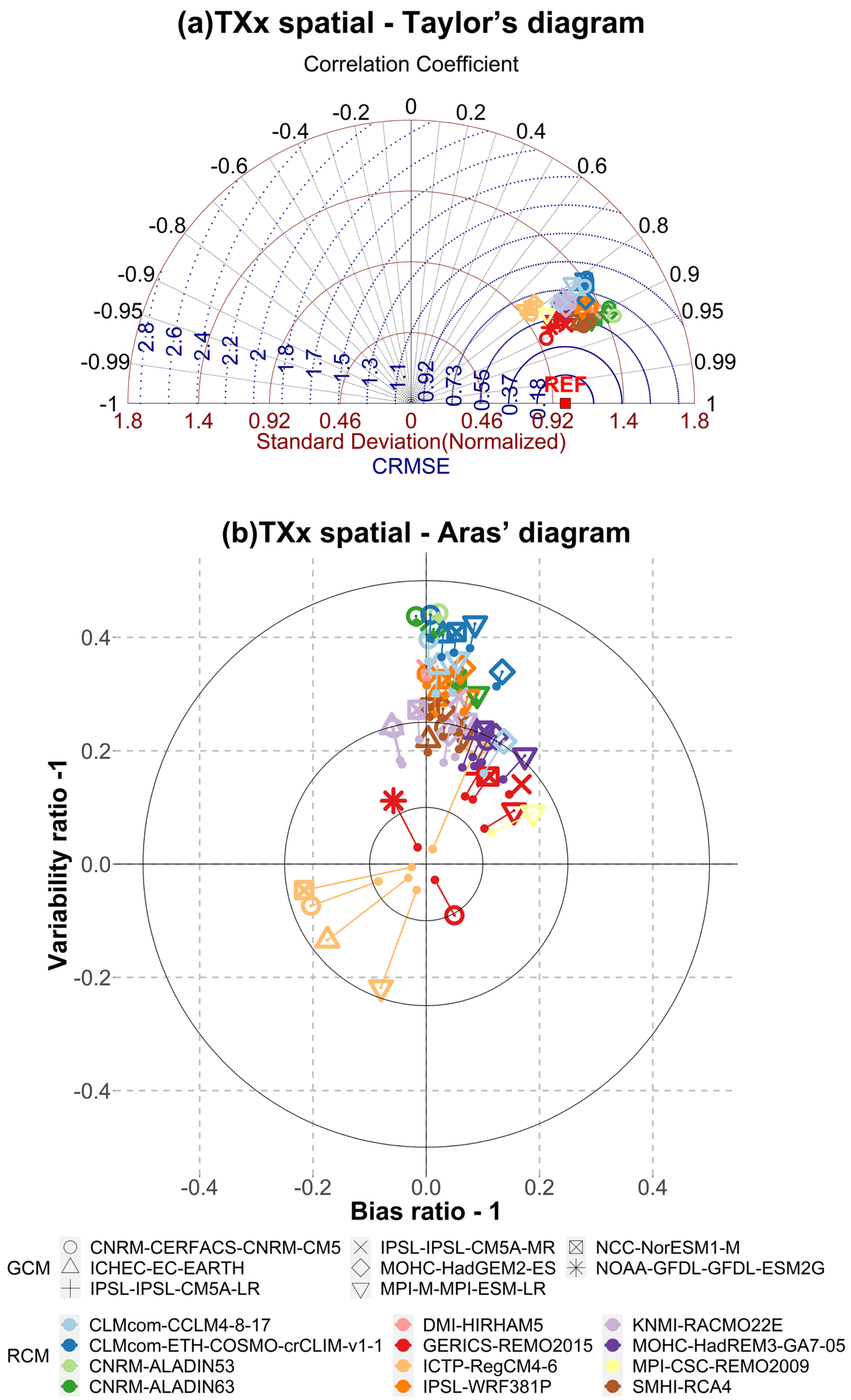

Fig. 6 Spatial climatological pattern analysis for the annual maxima of daily maximum surface temperatures using: **a** Taylor's diagram, and **b** Aras' diagram

exhibit negative bias, and c) all models except for 5 (all of which being among the 7 best performing models regarding bias-variability percentage error) overestimate the spatial variability of TXx field.

Shifting to the larger picture, a general observation one can make is that the overall EURO-CORDEX experiment provides better results for the maximum than for the minimum daily surface temperature in space (compare Figs. 6, 5), with the majority of model combinations overestimating both the mean and variability of the reference data.

The annual minima of daily minimum temperatures TNn, and the annual maxima of daily maximum temperatures TXx averaged over the entire study area were also subjected to temporal analysis with results shown in Figs. 7, 8, respectively.

The first impression one gets from both diagrams is that performances in terms of correlation in time are well below those in space for all models and indices analyzed (compare Figs. 5, 6 to Figs. 7, 8). The two Taylor diagrams in Figs. 7a, 8a show that correlation is always below 0.3–0.4 and a significant number of models exhibit negative correlation with the reference data in time. The insufficient overall performance in terms of correlation error is also shown in Aras' diagram, where all models display long segments, with those exhibiting negative correlations being marked with an empty circle at their endpoints $E_{\alpha\beta}$.

No model in both Aras' diagrams shown in Fig. 7b and Fig. 8b exhibits total percentage error below 50%, with most of the insufficient performance being due to errors in correlation. In fact, when considering the bias-variability percentage error $E_{\alpha\beta}$, a good number of models (i.e. about half for the minimum and almost all for the annual maximum temperatures) are encompassed within the 50% error circle.

More precisely, with specific regard to annual minima of daily minimum surface temperatures TNn (see Fig. 7b), 22 models exhibit bias-variability percentage errors (i.e. $E_{\alpha\beta}$) less than 50%, 7 models less than 25% and 1 model less than 10%. The latter, corresponds to RCM CLMcom-ETH-COSMO-crCLIM-v1-1 driven by GCM ICHEC-EC-EARTH. All except for 2 models overestimate the interannual variability, and one model (CNRM-ALADIN63 driven by MOHC-HadGEM2-ES) reproduces the interannual variability almost perfectly. All except 6 models exhibit positive bias, and a general tendency is observed towards overestimation of both the mean and variability of annual minimum daily temperatures.

The best model according to Aras' diagram is different from the best one indicated by Taylor's diagram; i.e. according to Taylor's diagram, RCM CNRM-ALADIN53 driven by GCM CNRM-CERFACS-CNRM-CM5 performs best (see Fig. 7a), whereas Aras' diagram indicates RCM CNRM-ALADIN63 nested in GCM MOHC-HadGEM2-ES as the best performing model combination (see Fig. 7b). This difference is attributed to the fact that, contrary to Taylor's diagram, Aras' diagram accounts for model biases, which in the case of annual minima of daily surface minimum temperatures play an important role; i.e. 117% for the best model according to Taylor's diagram, and 87% for the best model according to Aras' diagram (see Fig. 7b).

With specific regard to annual maxima of daily maximum temperatures (see Fig. 8), TXx, Taylor's and Aras' diagrams indicate the same best performing model: RCM GERICS-REMO2015 driven by GCM IPSL-IPSL-CM5A-MR, with total percentage error of 62% mostly due to low correlation.

When looking at the bias-variability percentage error, 11 models exhibit errors below 10%, 4 of them sharing the same RCM (ICTP-RegCM4-6) and other 3 sharing RCM IPSL-WRF381P. Also, 35 models exhibit bias-variability percentage errors below 25%.

A general tendency to overestimate both the mean and variability of maximum daily temperatures is also evident: 82% of the models have positive bias, and 78% of the models overestimate interannual variability with respect to observations.

With regard to precipitation over southern Italy, the EURO-CORDEX models show a wide range of performances in terms of both the spatial and temporal patterns of annual precipitation totals (PRCP) and annual maxima of daily rainfall (Rx1day). Results from the spatial analysis are shown in Figs. 9, 10.

The Taylor diagrams in Fig. 9a and Fig. 10a show that all models exhibit positive correlation with the reference data, indicating better performance for PRCP relative to Rx1day. This is also visible from the length of the segments and their filled endpoints shown in Aras' diagrams in Figs. 9b, 10b.

With regard to PRCP, the Aras diagram in Fig. 9b indicates that only one model exhibits total error less than 25% (i.e. RCM CLMcom-CCLM4-8–17 driven by GCM ICHEC-EC-EARTH). The same model is the second best in terms of bias-variability error (i.e. $E_{\alpha\beta}$ less than 10%), outperformed only by the model sharing the same RCM and driven by GCM MOHC-HadGEM2-ES. Also, all combinations of CLMcom-ETH-COSMO-crCLIM-v1-1 RCM perform well (i.e. bias-variability percentage error $E_{\alpha\beta}$ less than 25%), independent of the driving GCM. Worst performances in terms of overestimation of variability are associated with RCM DMI-HIRHAM5. Overall, 48 out of 55 models overestimate the variability of the reference data and 45 out of 55 models overestimate the observed mean.

Regarding the annual maximum daily precipitation, Rx1day, the Aras diagram in Fig. 10b shows that all models exhibit total errors above 50%, and positive bias with respect to the spatial climatological pattern averaged over time. When considering only the bias-variability percentage error $E_{\alpha\beta}$, best models (with error under 25%) are those resulting from RCM ICTP-RegCM4-6 driven by

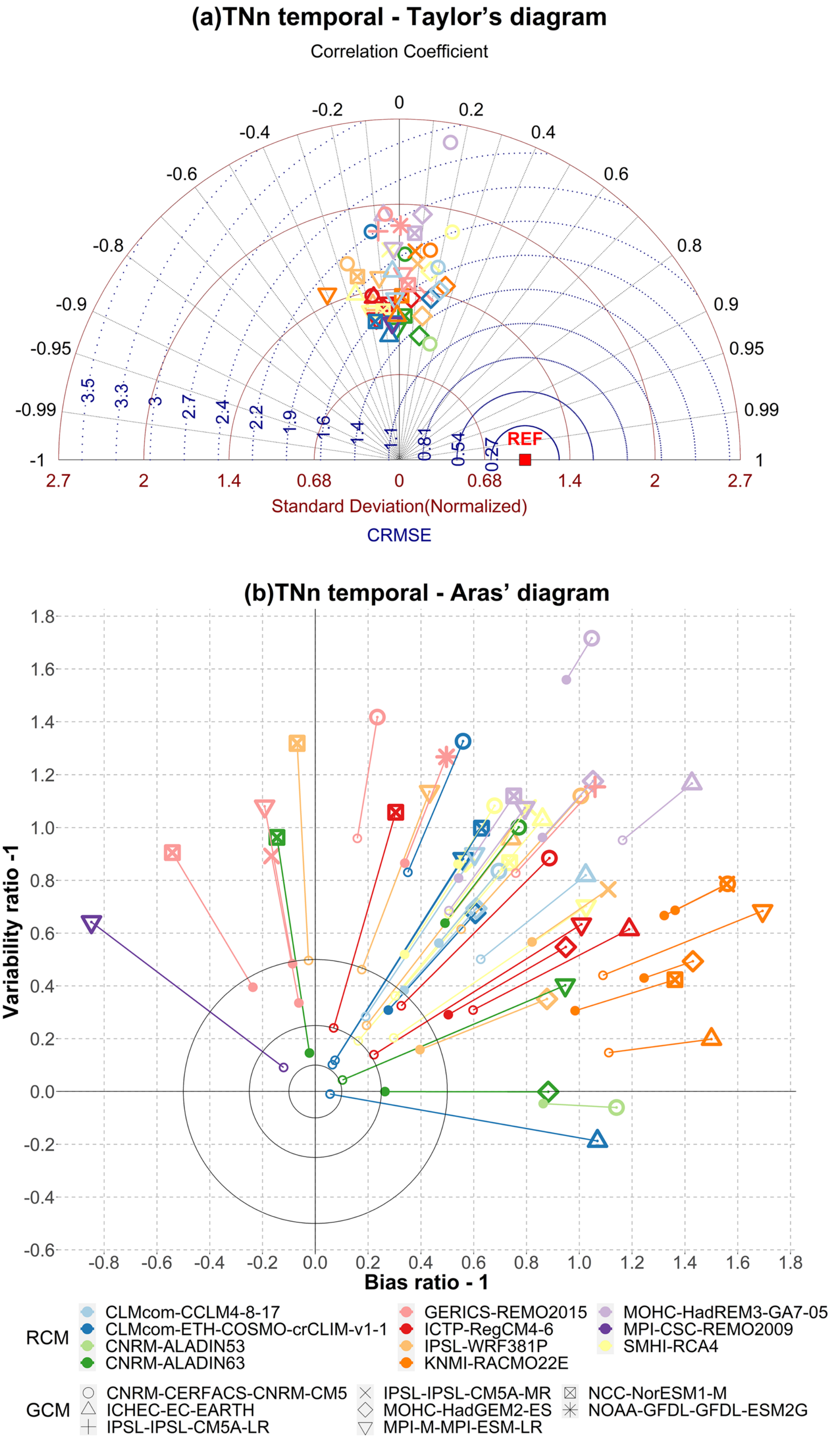

Fig. 7 Temporal analysis for the annual minima of daily minimum temperatures using: **a** Taylor's diagram, and **b** Aras' diagram

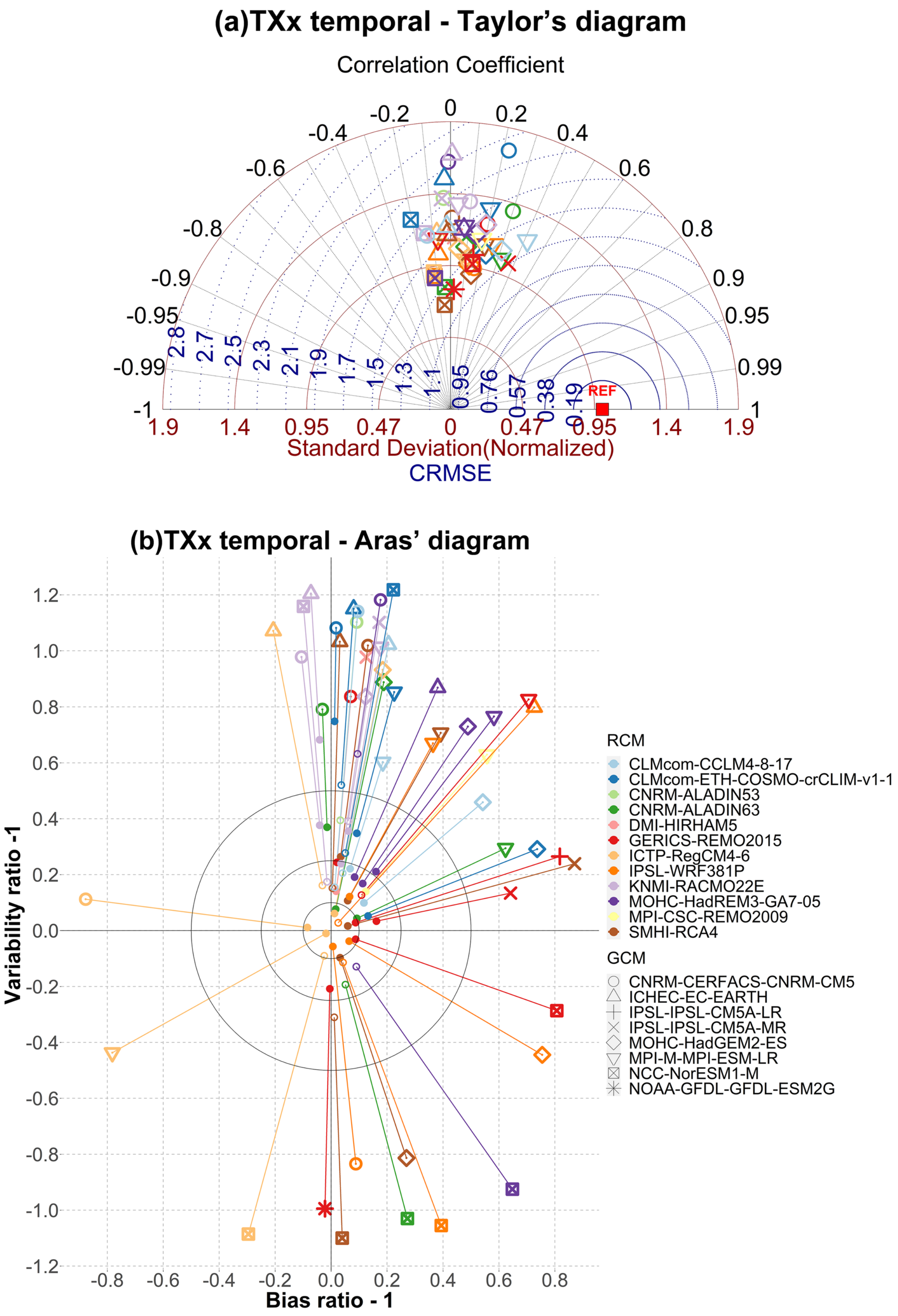

Fig. 8 Temporal analysis for the annual maxima of daily maximum surface temperatures using: **a** Taylor's diagram, and **b** Aras' diagram

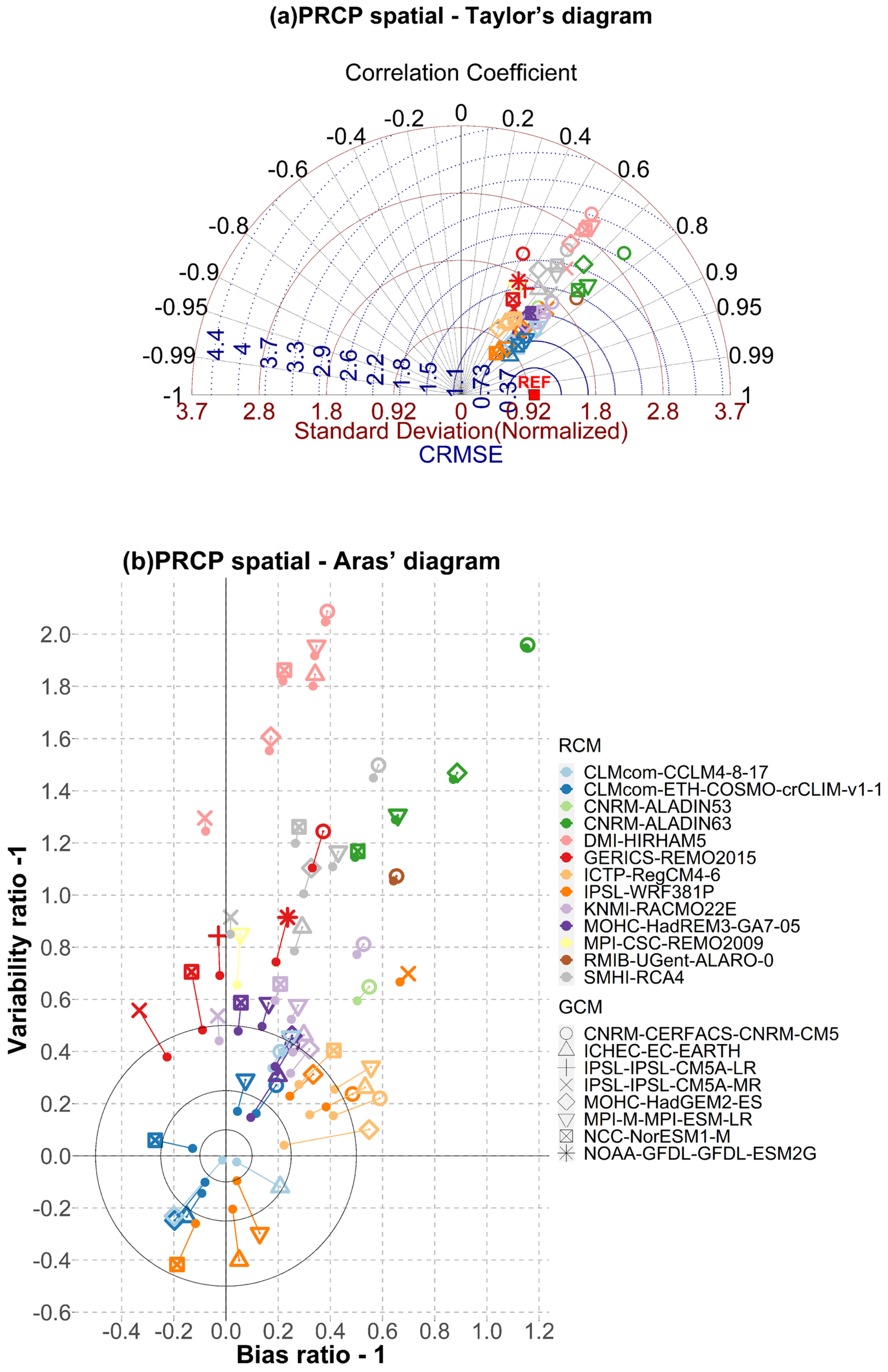

Fig. 9 Spatial analysis of annual total precipitation using: **a** Taylor's diagram, and **b** Aras' diagram

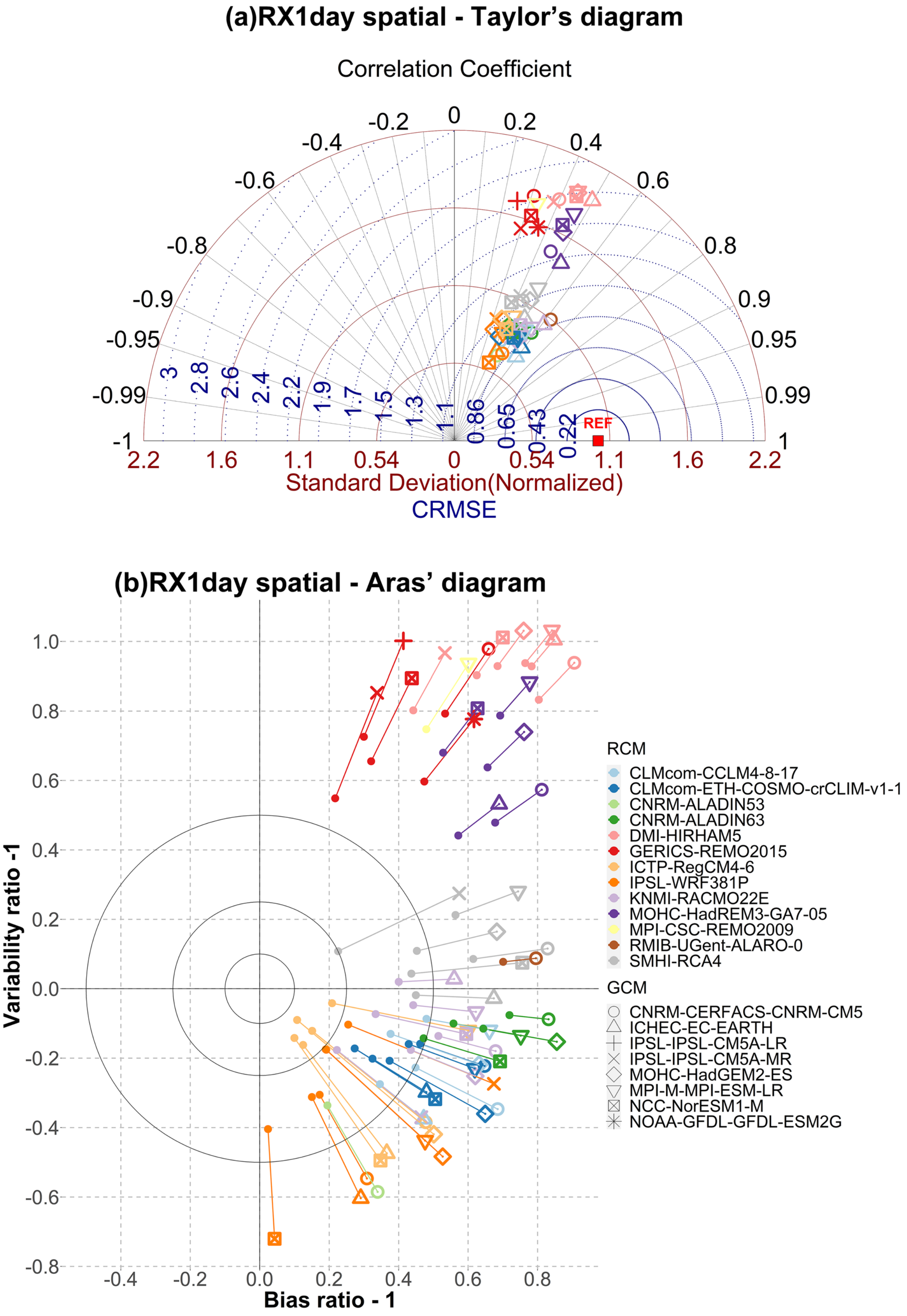

Fig. 10 Spatial analysis of annual maxima of precipitation using: **a** Taylor's diagram, **b** Aras' diagram

5 different GCMs. This can be seen as an evidence that regional climate model selection is of major importance when interest is in the spatial distribution of annual maxima of daily rainfall.

The results of the temporal analysis of the two indices related to precipitation (i.e. PRCP and Rx1day) are shown in Figs. 11, 12. The Taylor diagrams in Figs. 11a, 12a are somewhat similar to those of temperature related indices (Figs. 7a, 8a): all models show low correlation (always below 0.5) with an important number exhibiting negative correlations. The Aras diagrams in Figs. 11b, 12b confirm this and, in addition, indicate that none of the available GCM-RCM combinations exhibits total percentage error lower than 50%, for both the annual rainfall totals (PRCP) and the annual maxima of daily rainfall (Rx1day).

With specific regard to PCRP (Fig. 11), 7 models exhibit bias-variability percentage error $E_{\alpha\beta}$ below 10%, and 15 models below 25%. The best performing model in term of bias-variability percentage error corresponds to RCM CLMcom-ETH-COSMO-crCLIM-v1-1 driven by GCM MOHC-HadGEM2-ES, the second and third best models in terms of $E_{\alpha\beta}$ are RCMs CLMcom-ETH-COSMO-crCLIM-v1-1 and GERICS-REMO2015, both nested in GCM NCC-NorESM1-M. The majority of models (i.e. 48 out of 55) exhibit positive bias, and 41 out of 55 overestimate variability. The worst performing model according to Taylor's diagram is RCM CNRM-ALADIN63 driven by GCM NCC-NorESM1-M, while according to Aras' diagram the worst model combination results from the same RCM (i.e. CNRM-ALADIN63) nested in CNRM-CERFACS-CNRM-CM5. This difference is attributed to the fact that, differently from Taylor's diagram, Aras' diagram accounts for model biases, which in the latter case exceed 150%.

Results relative to temporal analysis of Rx1day averaged over Southern Italy (Fig. 12b), show that all models are associated with positive biases and, in addition, overestimate the interannual variability. No model exhibits total error below 50%, with the best model in terms of bias-variability error (below 25%) resulting from the combination of RCM ICTP-RegCM4-6 nested in GCM NCC-NorESM1-M. Models with $E_{\alpha\beta}$ below 50% include: RCM ICTP-RegCM4-6 driven by GCMs ICHEC-EC-EARTH, CNRM-CERFACS-CNRM-CM5 and MPI-M-MPI-ESM-LR, and RCM IPSL-WRF381P driven GCM NCC-NorESM1-M, and RCM CNRM-ALADIN53 driven by GCM CNRM-CERFACS-CNRM-CM5. The aforementioned same-RCM-clusters, regardless of the driving GCM, indicate, once more, the critical role of RCMs when it comes to modeling rainfall maxima. The worst model (in both diagrams) in terms of interannual variability of daily rainfall maxima is RCM MOHC-HadREM3-GA7-05 driven by GCM CNRM-CERFACS-CNRM-CM5, with associated error that exceeds 200%.

5 Discussion

Application of Aras' diagram to four indices related to extreme temperature and precipitation highlight the added value of including in a single 2-dimensional graphical tool three important and independent metrics for performance evaluation of climate models: correlation, bias and variability ratios. By displaying the total percentage error and the percentage error due to bias and variability, it is possible to easily identify models with similar performances, as well as the sources of observed discrepancies. The three homocentric circles corresponding to different values of relative error (i.e. 10%, 25% and 50%) ease selection of models with better overall performance, whereas decomposition of the observed error to bias-, variability- and correlation-related components becomes visually possible.

By visually detecting clusters of model combinations with similar performances (see Sect. 4), we found a general dominance of RCM selection in model performance; i.e. models using the same RCM driven by different GCMs tend to perform similarly. Also, similarly to other studies (see e.g., Deidda et al. 2013; Sillmann et al. 2013; Kotlarski et al. 2014; Mascaro et al. 2018; Vautard et al. 2021), we found that climate models performance for temperature is better than that for precipitation, in particular when looking at spatial patterns, see discussion on Figs. 5–12. The component of total error due to lack of correlation proved to be one of the most important sources of poor model performance for temporal analysis of all three variables (precipitation, minimum temperature, and maximum temperature), see Figs. 7, 8, 11, 12. Very good results in terms of correlation were obtained by all models of the EURO-CORDEX experiment in the analysis of the spatial distribution of maximum and minimum temperatures. Not so high but still positive correlations were obtained by all EURO-CORDEX models in the spatial analysis of annual rainfall totals (PRCP) and annual maxima of daily rainfall (Rx1day); see discussion on Figs. 9, 10. Lack of correlation becomes the most important error source in the temporal analysis, with about half of the models exhibiting negative correlations for all studied variables. Another general result, is that we could not identify a single best performing model for all studied variables.

We exploited Aras' diagram also to gain information about underestimation/overestimation of the mean and variability of the reference data. A systematic overestimation of the mean and variability can be observed in all analyses we performed; see Figs. 5–12 and their discussion. In particular, in the spatial analysis of minimum and maximum temperatures and annual rainfall totals, almost all models overestimate both the mean and variability of observed fields. In the spatial analysis of annual maxima

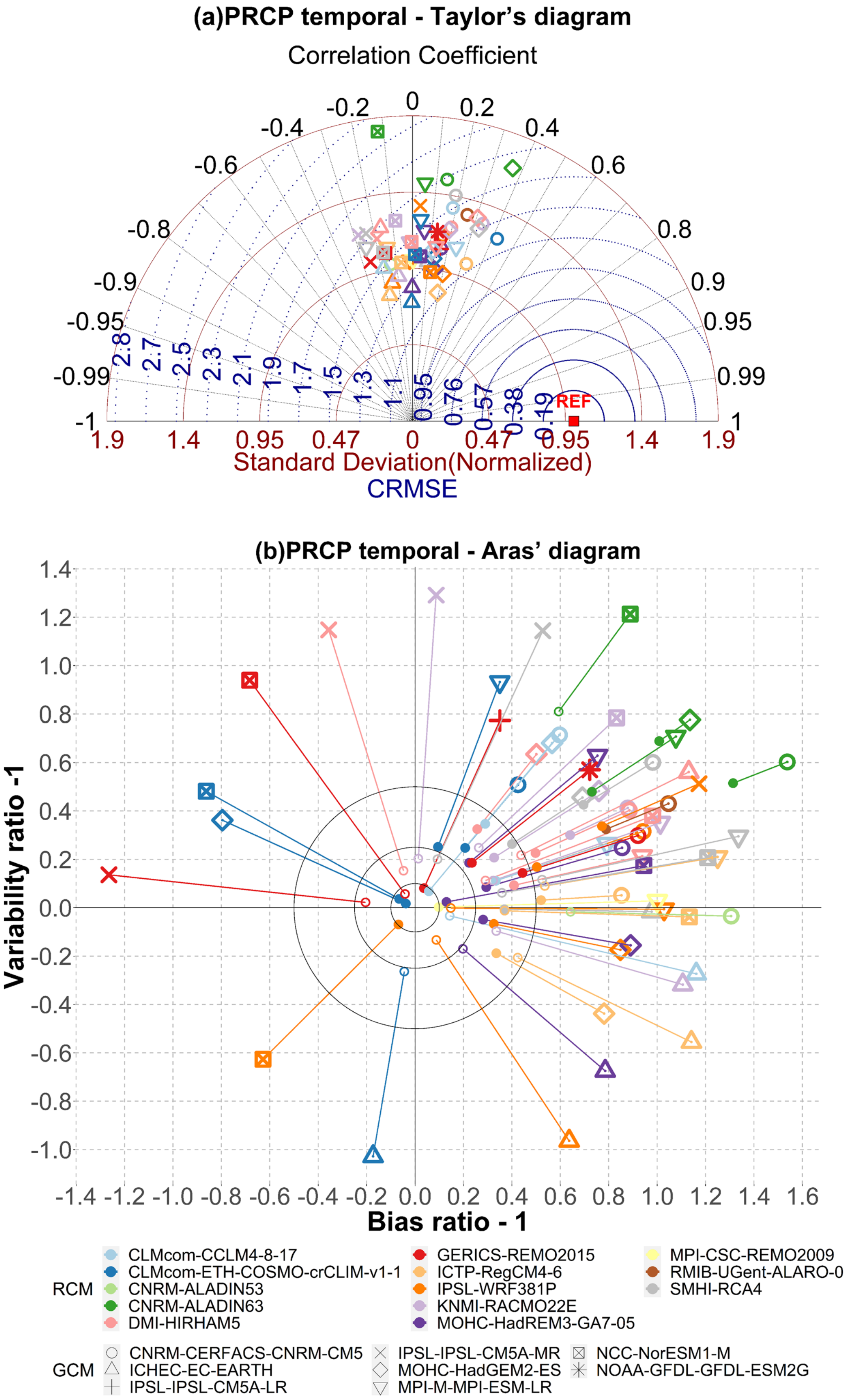

Fig. 11 Temporal analysis of annual total precipitation using: **a** Taylor's diagram, and **b** Aras' diagram

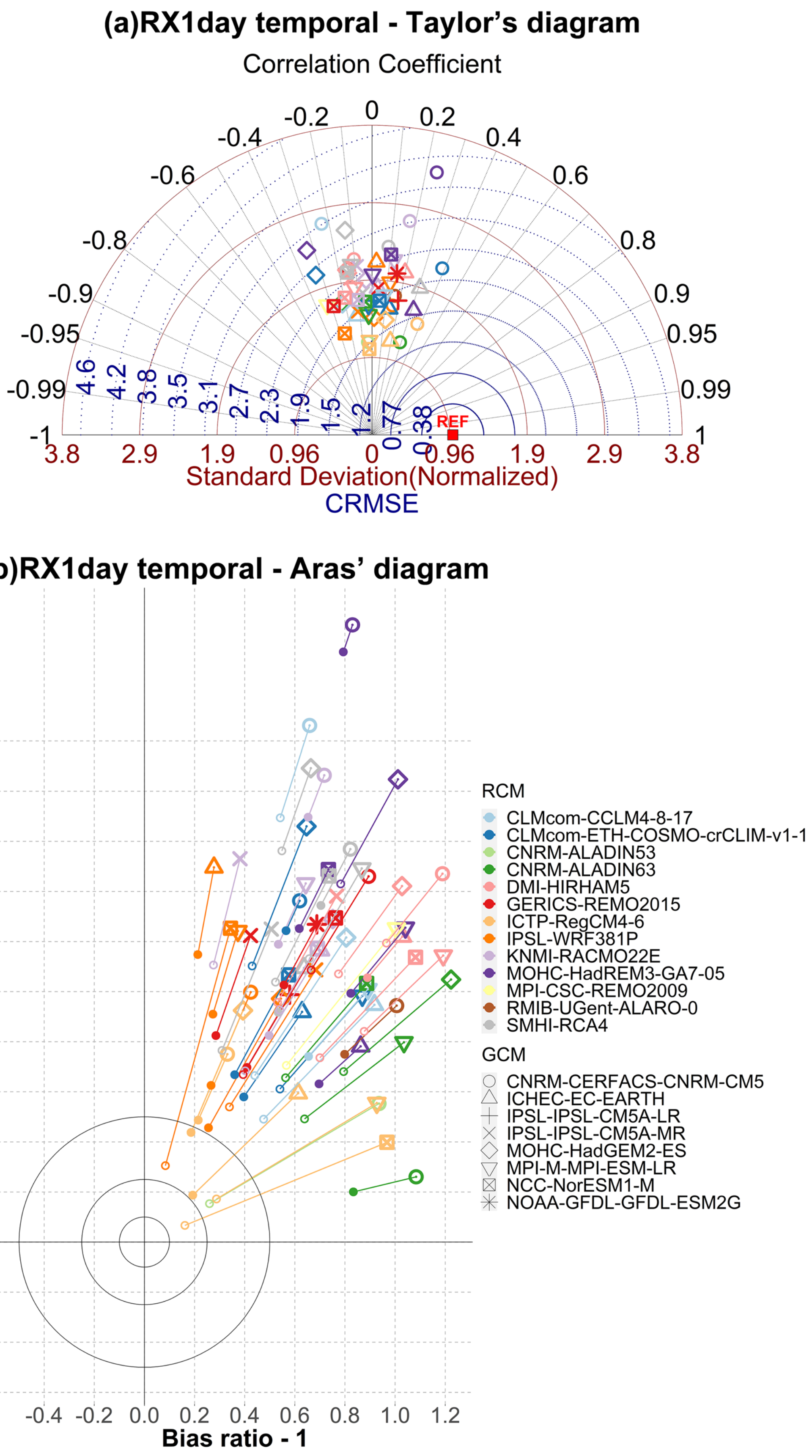

Fig. 12 Temporal analysis of annual maxima of daily precipitation using: **a** Taylor's diagram, and **b** Aras' diagram

of daily rainfall all models show a strong positive bias. In the temporal analysis of minimum temperatures almost all models overestimate variability, while for the maximum temperatures almost all models exhibit positive bias. In the temporal analysis of annual maxima of daily rainfall all models without exception strongly overestimate both the mean and variability of the spatial averages.

Taylor's and Aras' diagrams allow for different views of model performances. For instance, with regard to temporal analysis of annual rainfall totals (Fig. 11), best performing models according to Taylor's diagram are: RCM ICTP-RegCM4-6 driven by CNRM-CERFACS-CNRM-CM5, and RCMs MOHC-HadREM3-GA7-05, CLMcom-ETH-COSMO-crCLIM-v1-1 nested in CNRM-CERFACS-CNRM-CM5. Among those, Aras' diagram allows to clearly detect RCM CLMcom-ETH-COSMO-crCLIM-v1-1 driven by CNRM-CERFACS-CNRM-CM5 as the best performing one. This is due to the value added by accounting for biases in models' total error.

6 Conclusions

Different types of diagrams may be used to evaluate climate models. Among those, the most popular one is Taylor's diagram. While very useful, this diagram does not include information on bias-induced errors. In this context, we propose a two-dimensional diagram, referred to as Aras' diagram, which allows for visual evaluation of the relative performances of complex models. This is done by exploiting the Kling–Gupta efficiency (KGE) through a decomposition to its three main components: correlation, bias and variability ratios, allowing for quick assessments of possible overestimation or underestimation of both the mean and variability of observed fields.

Aras' diagram can be used in many aspects of model evaluation and diagnosis, for any type of variable, index, temporal or spatial climatological pattern, seasonality, etc. In this paper Aras' diagram has been used successfully for performance evaluation of climate models over Southern Italy for the hydrological variables of precipitation and temperature, by considering four climate indices: annual total precipitation, annual maxima of daily precipitation, annual minima of daily minimum temperatures, and annual maxima of daily maximum temperatures, in both space and time. The proposed diagram, based on the Kling–Gupta efficiency equation and its decomposition, has shown interesting properties which may assist climate model evaluations based on their accuracy in reproducing the climatological patterns observed in time and space.

An important note to be made here is that, as Aras' diagram is based on the standardization used in deriving the KGE index (see Sect. 2), it may lead to inconclusive results in cases when the observations exhibit low mean value and/or standard deviation (see e.g. Clark et al. 2021), resulting in high values of the bias and variability ratios, respectively. In such cases, Aras' and Taylor's should be used in conjunction, as their combined use should allow for a better overall assessment of the accuracy of model results relative to observations.

Appendix

Denote by (μ_m, σ_m) and (μ_o, σ_o) the first two statistical moments (i.e. means and standard deviations) of model results and observations, respectively. As shown by Gupta et al. (2009), the mean-squared-error (MSE) and the Nash–Sutcliffe efficiency (NSE) are linked through the following strictly monotonically decreasing relationship:

$$NSE = 1 - \frac{MSE}{\sigma_o^2} \tag{17}$$

In addition, Gupta et al. (2009) showed that NSE can be decomposed as:

$$NSE = 2a\rho - a^2 - b^2 \tag{18}$$

where $a = \sigma_m/\sigma_o$, $b = (\mu_m - \mu_o)/\sigma_o$, and ρ stands for the correlation coefficient between observations and model results.

By combining equations (17) and (18), one obtains:

$$MSE = \sigma_o^2(1 - 2a\rho + a^2 + b^2) \tag{19}$$

By taking the 1st and 2nd derivatives of equation (19) with respect to a, and considering that σ_o^2 is a positive constant, one obtains:

$$\frac{d(MSE)}{da} = 2\sigma_o^2(a - \rho), \frac{d^2(MSE)}{da^2} = 2\sigma_o^2 > 0 \tag{20}$$

According to equation (20), the minimum MSE (corresponding to $d(\mathrm{MSE})/da = 0$) is attained when $a = \sigma_m/\sigma_o = \rho \leq 1$. Therefore, unless $\rho = 1$, model selection through minimization of MSE and/or RMSE $=$ MSE$^{1/2}$, would favor models that tend to underestimate the variability of observations.

Acknowledgements The Authors greatly acknowledge the comments and suggestions of two anonymous reviewers that improved the presented work.

Author contributions All authors contributed to the study conception and design, Material preparation, data collection and analysis

Funding Open access funding provided by Politecnico di Bari within the CRUI-CARE Agreement. This work is partially funded by the projects entitled "Ricognizione delle disponibilità della risorsa idrica in Puglia mediante utilizzo di strumenti probabilistici nel contesto dei cambiamenti climatici" (Puglia regional programme "Research for Innovation") and “Dottorati di ricerca in Puglia XXXIII XXXIV XXXV ciclo” in the framework of the Puglia Regional Operational Programme FESR FSE 2014-2020.

Declarations

Conflict of interest The authors declare no competing interests.

References

Agyekum J, Annor T, Quansah E, Lamptey B, Okafor G (2022) Extreme precipitation indices over the Volta basin: CMIP6 model evaluation. Scientific African 16:e01181. https://doi.org/10.1016/j.sciaf.2022.e01181

Ahmed K, Sachindra DA, Shahid S, Demirel MC, Chung ES (2019) Selection of multi-model ensemble of general circulation models for the simulation of precipitation and maximum and minimum temperature based on spatial assessment metrics. Hydrol Earth Syst Sci 23(11):4803–4824. https://doi.org/10.5194/hess-23-4803-2019

Baker NC, Taylor PC (2016) A framework for evaluating climate model performance metrics. J Clim 29(5):1773–1782. https://doi.org/10.1175/JCLI-D-15-0114.1

Biondi D, Freni G, Iacobellis V, Mascaro G, Montanari A (2012) Validation of hydrological models: conceptual basis, methodological approaches and a proposal for a code of practice. Phys Chem Earth 42–44:70–76. https://doi.org/10.1016/j.pce.2011.07.037

Brunner MI, Melsen LA, Wood AW, Rakovec O, Mizukami N, Knoben WJM, Clark MP (2021) Flood spatial coherence, triggers, and performance in hydrological simulations: large-sample evaluation of four streamflow-calibrated models. Hydrol Earth Syst Sci 25(1):105–119. https://doi.org/10.5194/hess-25-105-2021

Castaneda-Gonzalez M, Poulin A, Romero-Lopez R, Arsenault R, Brissette F, Chaumont D, Paquin D (2018) Impacts of regional climate model spatial resolution on summer flood simulation. EPiC Ser Eng 3:372–362. https://doi.org/10.29007/hd8l

Chaulya SK, Prasad GM (2016) Application of cloud computing technology in mining industry. Elsevie, Sensing and Monitoring Technologies for Mines and Hazardous Areas

Clark MP, Vogel RM, Lamontagne JR, Mizukami N, Knoben WJM, Tang G, Gharari S, Freer JE, Whitfield PH, Shook KR, Papalexiou SM (2021) The abuse of popular performance metrics in hydrologic modeling. Water Resour Res 57(9):1–16. https://doi.org/10.1029/2020WR029001

Deidda R, Marrocu M, Caroletti G, Pusceddu G, Langousis A, Lucarini V, Puliga M, Speranza A (2013) Regional climate models’ performance in representing precipitation and temperature over selected Mediterranean areas. Hydrol Earth Syst Sci 17(12):5041–5059. https://doi.org/10.5194/hess-17-5041-2013

Emmanouil S, Langousis A, Nikolopoulos EI, Anagnostou EN (2021) An ERA-5 Derived CONUS-Wide high-resolution precipitation dataset based on a refined parametric statistical downscaling framework. Water Resour Res 57(6):1–17. https://doi.org/10.1029/2020WR029548

Emmanouil S, Langousis A, Nikolopoulos EI, Anagnostou EN (2022) The spatiotemporal evolution of rainfall extremes in a changing climate: a CONUS-wide assessment based on multifractal scaling arguments. Earth’s Future 10(3):1–16. https://doi.org/10.1029/2021EF002539

Emmanouil S, Langousis A, Nikolopoulos EI, Anagnostou EN (2023) Exploring the future of rainfall extremes over CONUS: the effects of high emission climate change trajectories on the intensity and frequency of rare precipitation events. Earth’s Future 11(4):1–21. https://doi.org/10.1029/2022EF003039

Flato et al. (2013). Evaluation of climate models, falsche info. Climate Change 2013 the Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.

Fowler HJ, Blenkinsop S, Tebaldi C (2007) Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. Int J Climatol 27(12):1547–1578. https://doi.org/10.1002/joc.1556

Gleckler PJ, Taylor KE, Doutriaux C (2008) Performance metrics for climate models. J Geophys Res Atmos 113(6):1–20. https://doi.org/10.1029/2007JD008972

Gupta HV, Kling H, Yilmaz KK, Martinez GF (2009) Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J Hydrol 377(1–2):80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Gupta HV, Wagener T, Liu Yuqiong (2008) Reconciling theory with observations: elements of a diagnostic approach to model evaluation. Hydrological Processes 22(18):3802–3813

Haylock MR, Hofstra N, Klein Tank AMG, Klok EJ, Jones PD, New M (2008a) A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. J Geophys Res Atmos. https://doi.org/10.1029/2008JD010201

Haylock MR, Hofstra N, Klein Tank AMG, Klok EJ, Jones PD, New M (2008b) A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. J Geophys Res Atmos 113(20):1–12. https://doi.org/10.1029/2008JD010201

Hu Z, Chen X, Zhou Q, Chen D, Li J (2019) DISO: a rethink of Taylor diagram. Int J Climatol 39(5):2825–2832. https://doi.org/10.1002/joc.5972

Jacob D, Petersen J, Eggert B, Alias A, Christensen OB, Bouwer LM, Braun A, Colette A, Déqué M, Georgievski G, Georgopoulou E, Gobiet A, Menut L, Nikulin G, Haensler A, Hempelmann N, Jones C, Keuler K, Kovats S, Yiou P (2014) EURO-CORDEX: new high-resolution climate change projections for European impact research. Reg Environ Change 14(2):563–578. https://doi.org/10.1007/s10113-013-0499-2

Johnson F, Sharma A (2009) Measurement of GCM skill in predicting variables relevant for hydroclimatological assessments. J Clim 22(16):4373–4382. https://doi.org/10.1175/2009JCLI2681.1

Jolliff JK, Kindle JC, Shulman I, Penta B, Friedrichs MAM, Helber R, Arnone RA (2009) Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment. J Mar Syst 76(1–2):64–82. https://doi.org/10.1016/j.jmarsys.2008.05.014

Kaleris V, Langousis A (2017) Comparison of two rainfall–runoff models: effects of conceptualization on water budget components. Hydrol Sci J 62(5):729–748. https://doi.org/10.1080/02626667.2016.1250899

Kirchmeier-Young MC, Zhang X (2020) Human influence has intensified extreme precipitation in North America. Proc Natl Acad Sci USA 117(24):13308–13313. https://doi.org/10.1073/pnas.1921628117

Knutti R (2010) The end of model democracy? Clim Change 102(3):395–404. https://doi.org/10.1007/s10584-010-9800-2

Kotlarski S, Keuler K, Christensen OB, Colette A, Déqué M, Gobiet A, Goergen K, Jacob D, Lüthi D, Van Meijgaard E, Nikulin G, Schär C, Teichmann C, Vautard R, Warrach-Sagi K, Wulfmeyer V (2014) Regional climate modeling on European scales: a joint standard evaluation of the EURO-CORDEX RCM ensemble. Geosci Model Dev 7(4):1297–1333. https://doi.org/10.5194/gmd-7-1297-2014

Krause P, Boyle DP, Bäse F (2005) Comparison of different efficiency criteria for hydrological model assessment. Adv Geosci 5:89–97. https://doi.org/10.5194/adgeo-5-89-2005

Kundzewicz ZW, Robson AJ (2004) Change detection in hydrological records: a review of the methodology. Hydrol Sci J 49(1):7–19. https://doi.org/10.1623/hysj.49.1.7.53993

Lamontagne JR, Barber CA, Vogel RM (2020) Improved estimators of model performance efficiency for skewed hydrologic data. Water Resour Res 56(9):1–25. https://doi.org/10.1029/2020WR027101

Langousis A, Kaleris V (2014) Statistical framework to simulate daily rainfall series conditional on upper-air predictor variables. Water Resour Res 50(5):3907–3932. https://doi.org/10.1002/2013WR014936

Langousis A, Mamalakis A, Deidda R, Marrocu M (2016) Assessing the relative effectiveness of statistical downscaling and distribution mapping in reproducing rainfall statistics based on climate model results. Water Resour Res RES 52:471–494. https://doi.org/10.1002/2015WR017556

Legates DR, McCabe GJ (1999) Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour Res 35(1):233–241. https://doi.org/10.1029/1998WR900018

Liu D (2020) A rational performance criterion for hydrological model. J Hydrol 590(September):125488. https://doi.org/10.1016/j.jhydrol.2020.125488

Mamalakis A, Langousis A, Deidda R, Marrocu M (2017) A parametric approach for simultaneous bias correction and high-resolution downscaling of climate model rainfall. Water Resour Res RES 53:2149–2170. https://doi.org/10.1002/2016WR019578

Mascaro G, Viola F, Deidda R (2018) Evaluation of precipitation from EURO-CORDEX regional climate simulations in a small-scale Mediterranean site. J Geophys Res: Atmos 123(3):1604–1625. https://doi.org/10.1002/2017JD027463

Moriasi DN, Gitau MW, Pai N, Daggupati P (2015) Hydrologic and water quality models: performance measures and evaluation criteria. Trans ASABE 58(6):1763–1785. https://doi.org/10.13031/trans.58.10715

Moriasi et al (2012) Hydrologic and water quality models: use, calibration, and validation. Am Soc Agric Biol Eng 55(4):1241–1247

Moustakis Y, Papalexiou SM, Onof CJ, Paschalis A (2021) Seasonality, intensity, and duration of rainfall extremes change in a warmer climate. Earth's Future 9(3):1–15. https://doi.org/10.1029/2020EF001824

Mujumdar P, Ghosh S, Raje D (2009) Hydro-meteorological predictions from GCM simulations: downscaling techniques and uncertainty modelling. IAHS AISH Publ 333:165–175

Mwangi HM, Julich S, Patil SD, McDonald MA, Feger KH (2016) Modelling the impact of agroforestry on hydrology of Mara river basin in East Africa. Hydrol Process 30(18):3139–3155. https://doi.org/10.1002/hyp.10852

Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I: a discussion of principles. J Hydrol 10(3):282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Palatella L, Miglietta MM, Paradisi P, Lionello P (2010) Climate change assessment for Mediterranean agricultural areas by statistical downscaling. Nat Hazards Earth Syst Sci 10(7):1647–1661. https://doi.org/10.5194/nhess-10-1647-2010

Paul A, Afroosa M, Baduru B, Paul B (2023) Showcasing model performance across space and time using single diagrams. Ocean Model 181:102150. https://doi.org/10.1016/j.ocemod.2022.102150

Pechlivanidis IG, Arheimer B (2015) Large-scale hydrological modelling by using modified PUB recommendations: The India-HYPE case. Hydrol Earth Syst Sci 19(11):4559–4579. https://doi.org/10.5194/hess-19-4559-2015

Perra E, Viola F, Deidda R, Caracciolo D, Paniconi C, Langousis A (2020) Hydrologic impacts of surface elevation and spatial resolution in statistical correction approaches: case study of Flumendosa basin, Italy. J Hydrol Eng. https://doi.org/10.1061/(asce)he.1943-5584.0001969

Pool S, Vis M, Seibert J (2018) Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency. Hydrol Sci J 63(13–14):1941–1953. https://doi.org/10.1080/02626667.2018.1552002

Ritter A, Muñoz-Carpena R (2013) Performance evaluation of hydrological models: statistical significance for reducing subjectivity in goodness-of-fit assessments. J Hydrol 480:33–45. https://doi.org/10.1016/j.jhydrol.2012.12.004

Sáenz J, Carreno-Madinabeitia S, Esnaola G, González-Rojí SJ, Ibarra-Berastegi G, Ulazia A (2020) The Sailor diagram: a new diagram for the verification of two-dimensional vector data from multiple models. Geosci Model Dev 13(7):3221–3240. https://doi.org/10.5194/gmd-13-3221-2020

Sillmann J, Kharin VV, Zhang X, Zwiers FW, Bronaugh D (2013) Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. J Geophys Res Atmos 118(4):1716–1733. https://doi.org/10.1002/jgrd.50203

Ta Z, Yu Y, Sun L, Chen X, Mu G, Yu R (2018) Assessment of precipitation simulations in Central Asia by CMIP5 climate models. Water (Switzerland). https://doi.org/10.3390/w10111516

Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. J Geophys Res: Atmos 106:7183–7192

Tsonis AA, Deyle ER, Ye H, Sugihara G (2017) Convergent cross mapping: theory and an example. In Adv Nonlinear Geosci. https://doi.org/10.1007/978-3-319-58895-7_27

Tsonis AA, Kirwan AD (2023) Geoengineering from the standpoint of uncertainty and related risks: science or science fiction? Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-023-02454-9

Vautard R, Kadygrov N, Iles C, Boberg F, Buonomo E, Bülow K, Coppola E, Corre L, van Meijgaard E, Nogherotto R, Sandstad M, Schwingshackl C, Somot S, Aalbers E, Christensen OB, Ciarlo JM, Demory ME, Giorgi F, Jacob D, Wulfmeyer V (2021) Evaluation of the Large EURO-CORDEX regional climate model ensemble. J Geophys Res: Atmos 126(17):1–28. https://doi.org/10.1029/2019JD032344

Vrac M, Marbaix P, Paillard D, Naveau P (2007) Non-linear statistical downscaling of present and LGM precipitation and temperatures over Europe. Climate of the past 3(4):669–682. https://doi.org/10.5194/cp-3-669-2007

Wetterhall F, Bárdossy A, Chen D, Halldin S, Xu CY (2009) Statistical downscaling of daily precipitation over Sweden using GCM output. Theoret Appl Climatol 96(1–2):95–103. https://doi.org/10.1007/s00704-008-0038-0

Xu Z, Hou Z, Han Y, Guo W (2016) A diagram for evaluating multiple aspects of model performance in simulating vector fields. Geosci Model Dev 9(12):4365–4380. https://doi.org/10.5194/gmd-9-4365-2016

Zhou Q, Chen D, Hu Z, Chen X (2021) Decompositions of Taylor diagram and DISO performance criteria. Int J Climatol 41(12):5726–5732. https://doi.org/10.1002/joc.7149