ISSN 0027-1349, Moscow University Physics Bulletin, 2024, Vol. 79, Suppl. 2, pp. S750–S764. © Allerton Press, Inc., 2024.

Pointwise and Complex Quality Metrics in Atmospheric Modeling: Methods and Approaches

V. Yu. Rezvov[1,2*], M. A. Krinitskiy[1,2], and M. A. Borisov[1,2]

[1] *Moscow Institute of Physics and Technology (National Research University), Dolgoprudny, Moscow oblast, Russia*

[2] *Shirshov Institute of Oceanology, Russian Academy of Sciences, Moscow, Russia*

Received October 1, 2024; revised October 10, 2024; accepted October 20, 2024

Abstract—In atmospheric sciences, various quantitative indicators, or metrics, are used to describe the quality of modeling results of various flavors including numerical weather prediction, statistical correction, various downscaling products, etc. Metrics provide the accuracy of reproduction of the processes underlying the models and allow comparison of models by assessing the uncertainty of their results. The key importance of metrics lies in a more thorough study of the advantages and disadvantages of classical approaches and in the development of new, more complex assessment methods. This article presents a classification of the most frequently encountered quality metrics in the scientific literature. In addition to assessing traditional pointwise metrics, complex methods considering various aspects of modeling results and special metrics used in climate studies are described. Among the complex metrics, methods with an emphasis on the spatial structure and heterogeneity of the predicted variable fields and probabilistic methods for verifying ensemble forecasts are distinguished. Special attention in this paper is devoted to the growing popularity of object-oriented metrics and metrics based on rare and extreme events. Climate models are assessed by comparing the results of retrospective modeling with historical data, which complicates the choice of metrics. A variety of climate metrics focusing on specific climate processes or integrating several parameters is described. The need for developing more diverse metrics for effective evaluation of climate models is explored. All metrics considered in this article are supplemented by examples in the scientific literature and assessments of their application to atmospheric research.

Keywords: quality metrics, grid data, downscaling, verification, pointwise metrics, complex metrics, ensemble forecasting, object-oriented metrics, climate metrics

DOI: 10.3103/S0027134924702229

1. INTRODUCTION

Climate change significantly impacts human economic activities, that makes accurate weather and climate forecasting crucial both for public safety and for global economy [1, 2]. However, atmospheric research is characterized by a gap between theoretical and applied aspects of predictive models [3]. This dilemma involves balancing the model ability to generalize diverse information with the reliability and uncertainty of the model results for specific tasks. Resolving this dilemma is vital for enhancing the practical significance of atmospheric research [3–7].

To bridge this gap, the justified metrics that represent model results in an easily interpretable form are necessary. Following [3], we define metrics as quantitative indicators that simplify complex, multidimensional datasets into brief numerical or categorical representations.

2. METRICS PROPERTIES

Metrics possess several important properties [3]. In general, they summarize and analyze specific features or aspects of modeling and evaluate variability in datasets or identify differences between datasets.

As shown in [6], quality metrics play a crucial role in reducing the gap between theoretical understanding of the atmosphere and practical application of modeling results. Metrics provide a quantitative description of key climate and weather phenomena and processes, essential for scientific and applied perspectives. These metrics also enable comparison

*E-mail: rezvov.vu@mipt.ru

between weather forecasting or climate models and assessment of their suitability for various tasks [3]. Such versatility advances climate modeling and aids in selecting the most relevant data for specific situations and applications [5, 6].

A primary goal of model evaluation is the quantitative description of uncertainties within the data itself, as shown in [8]. Such uncertainty impacts the informational content of the data [9] and determines its rational use, for example, for hypothesis testing [8]. Effective quantitative assessment of uncertainty also aids in understanding the characteristics of measurement or model errors, distinguishing between systematic and random errors [8, 10, 11].

Using a single metric is desirable from the point of view of optimization problems, yet it is often insufficient. A set of metrics, encompassing multiple components such as spatial distributions, vector fields, or time series, allows for a comprehensive study of spatial and temporal characteristics and interrelationships within the data [3].

Combining complex datasets becomes possible through dimensionality reduction methods applied to a set of multiple metrics, resulting in the generalized interpretable metric [12–14].

Reed et al. [3] also note that transforming atmospheric model output into a set of multiple quantitative metrics creates a universal array of information for both theoretical and practical purposes, addressing specific needs. Such joint study can significantly advance the effective use of research results [15–17].

In [8], it is emphasized that the legitimacy of applying any metrics to assess model quality depends on the assumption of statistical stationarity. When calculating a set of metrics, it is necessary to select appropriate spatial and temporal domains where model error characteristics are considered stationary and to process each such domain separately.

In [18], the importance of quality metrics selection is exemplified by their application to gridded meteorological data. The authors describe gridded meteorological data as meteorological variables, such as precipitation, near-surface temperature, or wind speed, whose values are given at the nodes of spatial grids. Gridded data are typically obtained by combining measurements from meteorological stations, remote sensing data, atmospheric models, and reanalyses [19–21].

The importance of gridded data in atmospheric sciences is defined by their application in various tasks, including climate change analysis [22]. The spatial resolution of gridded data from numerical models ranges from tens to hundreds of kilometers due to computational efficiency [23, 24]. Such resolution is insufficient for detailed regional or local weather modeling [18]. As shown in [25], high spatial resolution is crucial for tasks involving complex computations, such as energy balance.

This problem can be solved through downscaling [26–28]. Downscaling refers to a set of techniques for transforming spatial low-resolution data into high-resolution data [18], enhancing the gridded meteorological data for various applications.

In spatial downscaling of meteorological variables, the importance of selecting quality metrics for objective and quantitative evaluating of the method accuracy is most clearly demonstrated [18].

3. METRICS CLASSIFICATION

In general, quality metrics can be roughly divided into two categories [18].

The first category includes metrics based on pointwise errors between predicted and observed values of variables. Examples of such metrics are bias (B), mean absolute error (MAE), mean squared error (MSE), and root mean square error ($RMSE$). This group also includes peak signal-to-noise ratio ($PSNR$), calculated using the maximum value of the estimated quantity and MSE.

The second group of metrics evaluates different aspects of the model rather than errors at grid points. Examples of such metrics are the correlation coefficient (r), coefficient of determination (R^2), structural similarity index ($SSIM$), Kling–Gupta efficiency (KGE), relative operating characteristics skill score ($ROCSS$), 98th percentile bias ($P98$), continuous ranked probability score ($CRPS$), and fractions skill score (FSS). These metrics assess various aspects of modeling quality, including generalization and predictive ability, especially, for extreme values.

The coefficient of determination, commonly referred to as R^2, is a concept used in analysis of variance and regression analysis [29]. In general, it measures the proportion of variance in the data that is explained by the model $R^2 = 1 - FVU$, where FVU is the fraction of variance unexplained. Therefore, a higher R^2 value indicates that the model provides a better fit for the data.

In a special case, when the data comprises values $x_1, ..., x_N$ of a response variable and a model utilizing predictor variables is applied, R^2 is defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - x_i)^2}{\sum_{i=1}^{N} (x_i - \overline{x})^2},$$

where $\overline{x}$ represents the average of the observations and y_i is the predicted value of x_i based on the fitted model. If there is no relationship between the predictor variables and the response variables, then $\overline{x}$ serves

as the best "model" for explaining the data. Thus, the terms $x_i - \overline{x}$ capture the deviations in the scenario where there is no connection between the predictor variable and the response variable [29].

In [8], the authors do not provide a precise classification of quality metrics, however, referring to [30, 31], they highlight three most commonly used quality metrics. These are *B*, *MSE*, and *r*. It is easy to see that these metrics coincide with the main metrics in the classification described above [18]. In [8], the authors consider the examples of quality metrics derived from the basic ones: conditional and unconditional bias [32], unbiased root mean square error (*ubRMSE*), and skill score (*SS*) [33].

Table 1 lists the main quality metrics described in [8, 18].

3.1. Limitations of Commonly Employed Metrics

The "big three" commonly employed metrics [8] include *B*, *MSE*, and *r*, widely applied in various tasks related to atmosphere [13]. However, the analysis of these basic metrics reveals several fundamental limitations [8].

Firstly, classical metrics are interrelated and cannot be considered as independent quality indicators [13]. Such relations may indicate the redundancy of simultaneously applying several related metrics and lead to ambiguity in interpreting their values [8]. Additionally, in [8], the authors point out that different combinations of errors between predicted and actual values of variables can lead to the same metric values. The question of how many metrics are sufficient for an unambiguous and complete description of a model error remains open [8].

Secondly, most of commonly employed traditional metrics are based on the assumption that random error is normally distributed. Otherwise, classical quality metrics can result in misleading or meaningless outcomes [10, 34].

In [8], it is demonstrated that if both actual data x and model forecasts y are known, their joint distribution $p(y, x)$ contains all time-independent information relevant to model forecast verification [35]. Since the joint distribution depends on the conditional distribution $p(y|x)$ and the known distribution of actual data $p(x)$ as $p(y, x) = p(y|x)p(x)$, assessing model quality through the joint distribution is equivalent to assessing the conditional distribution [36]. Thus, parametric decomposition of the model errors $y(x)$ can determine all classical quality metrics based on the parameters of the decomposition [8, 31, 37].

According to several authors, classical metrics are fundamentally inapplicable for nonlinear error models and provide misleading interpretations of quality [10, 34].

3.2. Pointwise Metrics

The first two of the three commonly used traditional metrics described above, and their derivatives, are pointwise metrics, according to the classification given in [18]. However, the legitimacy of their use for evaluating the quality of hydrometeorological research is often not emphasized. The papers that either justify the use of pointwise metrics or highlight the limitations of classical metrics from a statistical (as in [8]) or physical perspective are more interesting, in our view.

MSE is effectively used as a standard statistical indicator for model quality assessment in meteorology and climate science [38–40]. *MAE* is also widely used in model evaluations, although there is no consensus on the most suitable metric for hydrometeorological tasks [40]. The ambiguity of *RMSE*, which gives higher weight to errors with larger absolute values compared to *MAE*, which gives equal weight to all errors, is highlighted in [41, 42].

Other studies have not clarified the superiority of any of two main pointwise metrics. In [40], the authors showed that *MSE* is more appropriate when model errors follow a normal distribution, which is also confirmed in [8]. Furthermore, [40] proved that *RMSE* satisfies the triangle inequality, a necessary condition for a distance metric, refuting the criticism presented in [41]. *MSE* is simpler for gradient calculation and thus more preferable as it can serve as a loss function in optimizing statistical models [40].

A more recent study [43] concludes that comparing *MSE* and *MAE* is pointless since they correspond to different distributions of model errors. In [43], the authors supplement the proposal [40] to use a combination of various indicators, including *RMSE* and *MAE*, for model quality evaluation. Metrics in combination should be weighted by their likelihood to avoid violating the theoretical principles of deriving metrics from the corresponding error distributions [43]. A justified result can also be achieved by transforming data to fit a particular distribution, such as normal, leading to a single pointwise metric.

3.3. Complex Metrics

Variable fields in atmospheric and ocean science typically exhibit coherent spatial structures with some inhomogeneities, such as in precipitation, wind speed or ocean currents fields. Standard pointwise metrics often do not account for the inherent spatial correlation in such fields. As a result, the interpretation of pointwise metrics is often challenging in a physical sense [44].

The study [45] reaches a broader conclusion that pointwise metrics are fundamentally unsuitable for

Table 1. Basic quality metrics. The observations and forecasts are denoted as x and y, respectively. Notations used to denote sample mean and variance for x and y, respectively, are defined as $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$, $\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, $\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \overline{x})^2$, $\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N} (y_i - \overline{y})^2$, where N is the number of grid nodes. FVU is the fraction of variance unexplained

Name	Definition	Range	Ideal value
Bias	$B = \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i)$	$(-\infty; +\infty)$	0
Mean absolute error	$MAE = \frac{1}{N}\sum_{i=1}^{N}\lvert y_i - x_i\rvert$	$[0; +\infty)$	0
Mean squared error	$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i)^2$	$[0; +\infty)$	0
Root mean square error	$RMSE = \sqrt{MSE}$	$[0; +\infty)$	0
Correlation coefficient	$r = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}}$	$[-1; 1]$	1
Coefficient of determination	$R^2 = 1 - FVU$ $R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - x_i)^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2}$	$(-\infty; 1]$	1
Peak signal-to-noise ratio	$PSNR = 10\log_{10}\left(\frac{\left[\max_i\{x_i\}\right]^2}{MSE}\right)$	$(-\infty; +\infty)$	$+\infty$

atmospheric applications. Quantitative assessment of weather forecast model quality represents a more complex task and requires developing multiple indicators focusing on different aspects of forecast quality.

Thus, several scientific works propose moving away from classical metrics towards nonparametric metrics, especially when a parametric error model cannot be determined [36, 46–48], and towards more comprehensive multidimensional metrics [31, 44], including object-oriented metrics [11].

In [49–51], the authors propose various methods for assessing the internal spatial and temporal correlation in meteorological datasets. Other quality metrics can account for the stochastic nature of specific meteorological variables, expressed as binary events, such as the presence or absence of precipitation [52]. The evaluation of extreme events suffers from the "forecaster's dilemma," which discredits skilful forecasts when they are evaluated only under the condition that an extreme event occurred. This results in violating the theoretical assumptions of forecast verification methods [45, 53].

The most frequently used complex metrics in atmospheric research are presented in Table 2 and will be discussed below.

3.3.1. Structural similarity index (*SSIM*). *SSIM* approximates the perceptual difference in spatial structure between two images [54]. This metric ranges from −1 to 1, where 1 indicates a perfect match between two images. Created for comparing natural images, *SSIM* evaluates differences in brightness, contrast, and structure using a sliding window.

Using the notations from Table 1, *SSIM* is calcu-

Table 2. Principal complex metrics

Name	Designation	Range	Ideal value
Structural similarity index	*SSIM*	$[-1; 1]$	1
Kling–Gupta efficiency	*KGE*	$(-\infty; 1]$	1
98th percentile bias	*P98*	$(-\infty; +\infty)$	0
Continuous ranked probability score	*CRPS*	$[0; +\infty)$	0
Fractions skill score	*FSS*	$[0; 1]$	1
Relative operating characteristics skill score	*ROCSS*	$[-1; 1]$	1

lated for each sliding window position:

$$SSIM = \frac{(2\overline{x} \cdot \overline{y} + C_1)(2\sigma_{xy} + C_2)}{(\overline{x}^2 + \overline{y}^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)},$$

where σ_{xy} is the covariance of x and y, and C_1 and C_2 are constants to stabilize the division with weak denominator.

This quality metric and its derivatives are widely used in atmospheric tasks, particularly for spatial downscaling of meteorological variables [55–58]. Gridded data can be represented as images, turning downscaling task into superresolution problem for natural images.

Compared to the actual distribution, various model realizations of meteorological fields may have similar pointwise metric values such as *MSE*, while their spatial coherent structures significantly differ [55]. As a local metric, *SSIM* can account for the overall covariance structure of errors, hypothetically eliminating some common shortcomings of pointwise quality metrics.

Nevertheless, there is growing concern about the insufficient justification for applying *SSIM* to gridded meteorological data, including the arbitrary choice of the sliding window scale [59]. Consequently, some authors do not recommend using *SSIM* as a complete replacement for pointwise metrics, since this index is directly related to *MSE*. *SSIM* definition through mean values, variances, and correlations does not adequately model human visual perception [60].

3.3.2. Kling–Gupta efficiency (*KGE*). As mentioned above, according to [40], it is possible to use pointwise metrics in tasks with gridded data by combining such metrics. [61] proposed another quality metric, Kling–Gupta Efficiency (*KGE*), which has gained popularity in hydrometeorological tasks, including downscaling [62, 63].

KGE is computed as follows:

$$KGE = 1 - \sqrt{(R-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2},$$

where R is the correlation coefficient, α is the variability parameter, and β is the bias parameter.

In the original form suggested in [61], the variability parameter is expressed as the ratio of the standard deviations of the modeled and actual data $\alpha = \frac{\sigma_y}{\sigma_x}$, and the bias parameter is expressed as the ratio of the means of the modeled and actual values $\beta = \frac{\overline{y}}{\overline{x}}$.

KGE takes values from $-\infty$ to 1, with a value of 1 indicating perfect agreement between the model and the observed values [64].

KGE was modified in [65]. Since the variability and bias parameters can be mutually correlated, it was proposed to use $\alpha_{\text{mod}} = \frac{\sigma_y/\overline{y}}{\sigma_x/\overline{x}} = \frac{\alpha}{\beta}$ instead of α. Despite the existence of other *KGE* modifications, such as nonparametric *KGE* [66], all of them share a common property: nonzero values do not correspond to a specific threshold value, that is viewed as a drawback by some authors [67]. In other words, negative values do not always indicate poor model quality, and positive values do not always indicate good quality, so *KGE* values are meaningful only relative to each other [67].

3.3.3. 98th percentile bias (*P98*). The 98th percentile is a score below which 98% of scores in its frequency distribution falls. These data can be aggregated over the entire time interval for each grid point, over the entire study area, etc.

In [18], it is asserted that the value of the 98th percentile itself can be considered a quality metric. Indeed, the 98th percentile is an important characteristic for understanding and interpreting the distribution of a meteorological variable, especially the "tail" of the distribution with extremely high values.

However, it is more appropriate, in our view, to speak of the 98th percentile bias, which is the difference between the 98th percentile values of the same variable between model results and actual data. This metric, *P98*, is applied for weather downscaling to determine how much the spectrum of high values of the variable is shifted in different parts of the modeling

area relative to the reference distribution of actual data. For example, in [68–70], *P98* evaluates statistical downscaling models of temperature and precipitation, and in [71], for downscaling wind fields.

3.3.4. Continuous ranked probability score (*CRPS*). *CRPS* is one of the most widely used quality metrics for probabilistic or ensemble forecasts of meteorological variables [72, 73]:

$$CRPS\left(\{y_i\}_{e=1,..,E}, x_i\right) = \int_{s=-\infty}^{+\infty} \left[F_i(s) - \theta(s - x_i)\right]^2 \, ds,$$

where $\{y_i\}_{e=1,..,E}$ is the forecasted variable at point i of the grid for E ensemble members, F_i is the cumulative distribution function of the forecasted variable at point i, and θ is the Heaviside function, acting as the cumulative distribution function for the trivial prediction equal to the actual value x_i [73].

Several useful properties of *CRPS* follow from its definition [72]. First, *CRPS* focuses on the entire range of the specific parameter and is considered as a ranked probability assessment with an infinite number of classes, each of which has an infinitesimally small width. Second, this metric is sensitive to the entire continuous range of the variable and does not require the introduction of predefined classes, on which results could depend. Moreover, for deterministic forecasts, *CRPS* is equal to *MAE* and, therefore, has an unambiguous physical interpretation.

CRPS accounts for both bias and uncertainty of the forecast ensemble members. The minimum possible value $CRPS = 0$ corresponds to the ideal case when all ensemble members are identical to the observation. Increasing *CRPS* corresponds to increases in systematic error and forecast ensemble spread [73].

Examples of *CRPS* use are also in [74–76].

3.3.5. Fractions skill score (*FSS*). As noted in [77], unlike classical metrics, which assess the match between modeled and observed values at the same point, there are metrics that evaluate the forecast and/or observation in its vicinity, implying some data processing within a sliding window. The examples include calculating the mean value, the presence of a phenomenon within the sliding window, the frequency of the phenomenon in the window, the distribution of values in the window, etc. [77].

One of the most common neighborhood quality metrics is fractions skill score (*FSS*), initially developed for assessing precipitation modeling [52, 78–80]. *FSS* does not require significant computational costs [81]. *FSS* evaluates the fraction of grid cells within a sliding window filled by the phenomenon in the forecast field P_y and the observation field P_x. For all positions of the sliding window, the root mean square difference of fractions over all grid points is calculated, divided by the worst possible forecast [77]:

$$FSS = 1 - \frac{\sum_{i=1}^{N} (P_{y,i} - P_{x,i})^2}{\sum_{i=1}^{N} {P_{x,i}}^2 + \sum_{i=1}^{N} {P_{y,i}}^2}.$$

One of the most important properties of *FSS* is its ability to determine the spatial scale at which the model can be considered to have acceptable quality [52, 79, 80, 82] by adjusting the size of the sliding window. Additionally, studies have shown that *FSS* is related to the spatial displacement of precipitation in the forecast, which partially mimics the visual assessment of precipitation forecast quality and allows easy interpretation of *FSS* value [78, 83].

A model is considered useful if $FSS > FSS_{\text{thr}}$, where the threshold value $FSS_{\text{thr}} = \frac{1}{2}(1 + f_0)$ is determined by the fraction f_0 of grid cells with the phenomenon in the entire forecast area [77]. An ideal forecast has $FSS = 1$, while the worst modeling scenario is $FSS = 0$ [84].

Although *FSS* is most commonly used to assess precipitation modeling quality [85], it is also used for other variables with small-scale features, such as near-surface air temperature [86] and near-surface wind [87].

3.3.6. Other complex metrics. In the scientific literature dedicated to evaluating the quality of results from various atmospheric models, there are other metric, which are still of interest, while not so widely used.

Relative operating characteristic skill score (*ROCSS*) is a complex measure for classifiers and represents a comparison of the hit rate and false alarm rate [88]. *ROCSS* is often used to assess the quality of probabilistic forecasts [88–91].

This metric ranges from −1 to 1. $ROCSS = 1$ indicates perfect classification, $ROCSS = 0$ indicates model quality equivalent to random guessing, and $ROCSS = -1$ indicates completely incorrect classification [18].

In [92], the concept of an object-oriented measure is presented, based on the identification of "objects" and comparison of their attributes (e.g., intensity, centroid positions) between model data and actual observations. An important property of such measures is their flexibility concerning attributes inherent to different "objects." The authors consider spatial and temporal coordinates to assess the accuracy of modeling precipitation areas [92].

Despite the criticism of metrics based on extreme event verification [45, 53], it is crucial to warn the society about hazardous meteorological phenomena

that can cause significant damage and loss. For such extreme events, it is important to correctly assess the quality of forecasting meteorological variables [44].

In the simplest approach, an extreme event is considered as a binary category in terms of its occurrence or nonoccurrence. The quality of deterministic forecasts of such events is usually assessed using traditional categorical scores for binary events [44]. However, these metrics tend to reach uninformative limits of 0 or 1 for increasingly rare events [93].

In [93], a more sophisticated metric, extreme dependency score (*EDS*), is proposed to assess the quality of deterministic forecasting of rare binary events. One advantage of *EDS* is its lack of explicit dependence on the threshold selection for the occurrence of an extreme event [44].

The low frequency of extreme events creates certain challenges for quality assessment. Firstly, the rarity of events can lead to significant sampling uncertainty. This problem is partially resolved by aggregating observations and forecasts into larger spatial-temporal regions, but larger areas can also lead to nonhomogeneity and nonstationarity [44]. Secondly, the rarity of extreme events can result in small or zero event counts when dividing events into classes. Thirdly, with small sample sizes, outliers can significantly influence the verification of extreme event forecasts [44].

Therefore, to reduce the error in assessing extreme events, it is proposed to use metrics in the approach of probabilistic rather than deterministic forecasts [44, 94].

3.4. Metrics for Climate Models

Physics-based models are essential tools for assessing changes in Earth's climate system due to external forces and internal variability [95]. Nevertheless, evaluating the effectiveness of these models is challenging. Quality metrics for climate models stand apart from the basic set of metrics discussed above.

Typically, climate models are assessed by comparing hindcasts of past climate variables with historical observations. However, development of these hindcast simulations is not always independent from the tuning of parameters that govern unresolved physics [95, 96].

Climate model projections are viewed as predictions based on specific forcing scenarios. These projections represent potential future outcomes when realistic scenarios are employed, and the actual forcings are qualitatively similar to those projected. Comparing model projections with observations serves as a test of model skill, with successful projections enhancing confidence in future forecasting [95].

Efforts like the Coupled Model Intercomparison Project (CMIP) aim to standardize climate model experiments and gather simulation data for intercomparison and evaluation of climate models. CMIP involves numerous global climate models, while quality metrics systematically quantify model biases and behaviors against observations [97].

A diverse range of quality metrics has been established to evaluate climate models [13, 97, 98]. While many metrics focus on quantifying model performance for individual climate characteristics or processes, some assessments adopt a broader perspective by combining multiple diagnostics into a single performance index [97]. Despite the existing advancements, [99] emphasizes the need for a greater variety of quality metrics, as well as the development of an optimal suite from the existing metrics.

The most frequently used metrics for climate models are presented in Table 3 and will be discussed below.

3.4.1. Pointwise climatologies. To evaluate the performance of the global climate models, the gridded model outputs are usually used as historical simulated data to compare with the historical observed data. The observed data can be either gridded, or station-based [100].

Annual pointwise climatologies are calculated for each gridded dataset. They can be presented in two formats: as deviations from station data and as deviations from the average of the gridded datasets [101].

3.4.1.1. Average T_{max} and T_{min}. The 2-meter temperature is a key variable for analyzing both current climate conditions and climate change. Numerous studies, evaluated in Intergovernmental Panel on Climate Change (IPCC) reports, have compared this variable with observational data [102]. While global mean surface temperature changes serve as a valuable indicator of climate change and variability, examining mean daily minimum (T_{min}) and maximum (T_{max}) temperatures provides a more comprehensive understanding [103]. Moreover, many climate change impacts are more directly associated with fluctuations in T_{min} and T_{max} than with extreme weather events. For instance, growing degree days and accumulated chill hours used in predictive models for crop development are affected differently by T_{min} and T_{max}. Additionally, variations in evapotranspiration and photosynthetic rates are more likely to be influenced by T_{max} rather than T_{min} [104].

To compute these climatologies, for each model and observed data, T_{max} and T_{min} are averaged monthly or seasonally [104]. The resulting bias in T_{max} and T_{min} is computed as the difference between the downscaled and the observed mean values [105].

Table 3. Principal metrics for climate models

Name	Designation	Range	Ideal value
Average T_{max} and T_{min}	ΔT_{max}, ΔT_{min}	$(-\infty; +\infty)$	0
Diurnal temperature range	ΔDTR	$(-\infty; +\infty)$	0
Linear trend		$(-\infty; +\infty)$	0
Variability		$(-\infty; +\infty)$	0
Probability density function (PDF) skill score	S_{score}	$[0; 1]$	1
Cramér–von Mises score	CvM	$[0; +\infty)$	0
Regime-oriented metrics	T_{bias}, P_{bias}	$(-\infty; +\infty)$	0
Heat wave magnitude index daily	$\Delta HWMId$	$(-\infty; +\infty)$	0
Warm spell duration index	$\Delta WSDI$	$(-\infty; +\infty)$	0

3.4.1.2. Diurnal temperature range (DTR). The studies have shown [104] that for most locations, average changes in T_{min} across the models differ from associated changes in T_{max}. As a result, the average diurnal temperature range ($DTR = T_{max} - T_{min}$) is used as a metric indicating the quality of climate modeling.

In [103], the authors indicate that utilizing a collection of diverse climate indices, which are independent of internal climate variability and exhibit a consistent reaction to greenhouse forcing, can enhance the detection of climate change. Furthermore, additional indices, such as *DTR*, can be beneficial in establishing a distinct signature of observed climate change that is less prone to uniform responses from various radiative forcing mechanisms.

Global *DTR* over land is proposed as a potential index of radiative-forced climate change, offering insights that are distinct from those provided by global mean temperature [103].

DTR arises from the intricate interactions among various processes, including those involving the land surface, atmospheric boundary layer, cloud formations, radiative effects of aerosols, and the general atmospheric circulation. Nevertheless, the interactions between clouds and aerosols remain a significant challenge for climate models, complicating the accurate simulation of *DTR* [102].

3.4.1.3. Linear trend. One more example of the pointwise climatologies is linear trend. Climate models should be adept at capturing forced trends when these are present in the boundary conditions driving the models. However, the exclusion or incorrect representation of potential local to regional forcings, such as aerosols or land-use changes, may impede this capability. Given that changes in local or regional forcings are crucial for future regional climate projections, this aspect is evaluated by comparing model-derived trends with corresponding observational data [106]. This metric assesses the model skill in replicating observed trends across all seasons and for annual averages.

Linear trends are calculated for each grid cell by applying least squares linear regression to the complete series of data, e.g., for monthly anomalies. The goal of this analysis is also to underscore discrepancies in trends among the observed and modeled datasets. Significant differences depend on how each dataset addresses data inhomogeneities [101].

One significant limitation is that linear trends may not adequately capture nonlinear changes or abrupt shifts in climate data, which can lead to misleading interpretations. Additionally, linear trend analysis assumes that the relationship between time and the climate variable is constant, which may not hold true in the presence of external factors or climate variability. Furthermore, the presence of outliers can significantly affect the results, potentially skewing the trend estimation.

3.4.1.4. Variability. Temporal variability is a crucial aspect of climate, encompassing systematic changes like the annual cycle, along with residual temporal fluctuations including short-term variations, episodic events, and variability that spans from interannual to long-term trends [107].

To assess data variability, the standard deviation of the complete series of modeled data is calculated for each grid cell. Similarly, variability is determined for station data by utilizing all available series of data. To further explore spatial covariability, empirical orthogonal function (EOF) analysis can be conducted on the entire sequence of data. EOFs, which depict spatial patterns, capture the main modes of spatial

covariability within the domain. The associated principal components (PCs) are time series that reflect the temporal scaling of these patterns. For variability metrics, a comparison is made between the three leading EOFs and their respective principal components [101].

3.4.1.5. Probability density function (PDF) skill score. In global climate model projections, such variables as precipitation often spread over too many days with small amounts, leading to an underestimation of dry days and extreme precipitation events. The metrics estimating the frequency of different event types under climate change scenarios are crucial for understanding climate shifts [108, 109]. *PDF* skill score evaluates the resemblance between two probability density functions. It does so by determining the minimum value between observed and modeled counts within each histogram bin, thereby quantifying the shared area within the distributions [109]:

$$S_{\text{score}} = \sum_{i=1}^{N} \min\left(Z_{m,i}, Z_{o,i}\right),$$

where S_{score} is *PDF* skill score, $Z_{m,i}$ and $Z_{o,i}$ are the frequency of modeled and observed values in a ith bin, respectively, and N is the number of bins used to calculate *PDF*. *PDF* skill score provides an easily-interpretable score for the similarity between the observed and downscaled distribution of values [109].

PDF score is calculated for each individual grid cell. *PDF* scores close to 0 indicate negligible overlap between the reference and the model, and score close to 1 indicates two identical *PDF*s.

Evaluating a climate model using *PDF* score offers a significant advantage. If a model can accurately simulate an entire *PDF*, it demonstrates its ability to replicate rare values that may become more common in the future. When the distribution of values shifts due to climate change, there will likely be considerable overlap between the current and future distributions. Confidence in the model projections diminishes as the overlap between present and future *PDF*s decreases over time. However, until this overlap becomes critically small, a model can be used to simulate the entire *PDF* of a set of variables as a criterion for future impacts assessments. Demonstrating skill in simulating whole *PDF*s is a more rigorous test of a model compared to reproducing just the mean and the standard deviation [110].

A *PDF*-based assessment is not without limitations. It does not assess crucial aspects of model performance, such as periods of sustained high temperatures or heavy rainfall represented by indices [111]. Additionally, the importance of a model's failure to simulate rare events diminishes as the frequency of these events decreases in both the model and observed data. Despite these limitations, a *PDF*-based evaluation of a climate model could effectively replace traditional mean performance evaluations [110].

3.4.2. Cramér–von Mises score. In [112], there was introduced a novel scheme based on a nonparametric statistical test that enables the quantification of differences between two distributions along with their statistical significance. Discrepancies between two distributions can be measured using quadratic statistics, which are less sensitive to singular or outlier values, such as the Cramér–von Mises (CvM) criterion [113]:

$$CvM = \int_{-\infty}^{+\infty} \left[F_n(x) - F(x,\theta)\right]^2 \, dF(x),$$

where $F(\theta)$ is the fitted distribution function and $F_n(x)$ is an empirical distribution function of sample size n [112].

CvM metrics is a measure of the mean squared difference between the empirical and hypothetical cumulative distribution functions. Small values of the CvM statistic indicate a small distance between the two distributions [112, 114].

In climate modeling, CvM statistic serves as a valuable metric for comparing the distributions simulated by models with observed distributions. CvM statistic can be employed, for example, to assess the differences between observed and modeled extreme precipitation distributions across various stations [112].

This simple metric neglects a potential spatial autocorrelation of the test statistic and does not consider field significance [115].

3.4.3. Process-based metrics. The authors of [116] call to an attempt to disentangle the eventual dependency of model performances on the main atmospheric circulations and processes, from large-scale to local scale. They also claim the need for process-based model evaluations, which allow for a better understanding of model errors and their cause.

3.4.3.1. Regime-oriented approach. In the proposed "regime-oriented" approach [116], model results are averaged within categories that describe physically distinct regimes of the system, such as circulation regimes and thermodynamic states, rather than being averaged over space or time (e.g., global averages or yearly mean values) as for pointwise climatic metrics.

In [116], the authors propose a number of the atmospheric regimes, from large- to local-scale, suitable for process-oriented metrics. Large-scale circulation patterns in the North Atlantic/Eurasian region

significantly influence regional atmospheric variability, with the North Atlantic Oscillation (NAO) and atmospheric blocking being prime examples. NAO, the most prominent atmospheric variability pattern in the Northern Hemisphere's mid and high latitudes, strongly affects the European climate. Atmospheric blocking occurs when the westerly flow in midlatitudes is obstructed or redirected northward by a high-pressure system, lasting from several days to weeks. Blocking has been closely studied due to its impact on local weather, being linked to cold extremes in winter and hot extremes in summer, such as the 2010 Russian heat wave.

The example of the regional/local-scale atmospheric regime is the bora wind [116], a gusty downslope flow characterized by its variability due to geographical location and coastal topography. The wind speed of the bora varies spatially due to the formation of alternating jets and wakes linked to mountain gaps, affecting the duration and intensity of the wind.

The important step of the process-oriented metric is detecting a phenomenon. To analyze each of the atmospheric circulation phenomena mentioned, a time series categorizes days as either event or non-event days. For the continuous NAO index, thresholds are applied to generate this binary classification. This allows to estimate the sensitivity of local weather to specific atmospheric circulations for various meteorological variables. For temperature, this sensitivity is measured seasonally by comparing the temperature differences between composites of event days (1) and nonevent days (0) for modeled (mod) and observed (obs) data [116]:

$$T_{\text{bias}} = \left(T_{\text{mod},1} - T_{\text{mod},0}\right) - \left(T_{\text{obs},1} - T_{\text{obs},0}\right).$$

Similarly, for precipitation, sensitivity is determined by calculating the ratio of precipitation on event days to that on nonevent days [116]:

$$P_{\text{bias}} = \frac{P_{\text{mod},1}}{P_{\text{mod},0}} - \frac{P_{\text{obs},1}}{P_{\text{obs},0}}.$$

Small values of bias, close to 0, indicate high process-oriented quality of the climate model.

3.4.3.2. Extreme events. Heat waves significantly impact society, agriculture, and the environment, making them critical extreme climate events. The authors of [117] conduct a comprehensive analysis of future heat waves based on simulations. There is plenty of specific definitions with slight differences for thresholds or number of consecutive days for heat waves. Various definitions of heat waves provide insights into different aspects of their characteristics, such as length (number of days), event frequency (differentiating between multiple short events and fewer lengthy ones with the same total duration), magnitude (total temperature exceedance), and peak temperature during the event. A comprehensive definition of heat waves must encompass both intensity and duration [117]. Commonly utilized indices for measuring these aspects include the Heat Wave Magnitude Index daily (*HWMId*) for intensity [118] and the Warm Spell Duration Index *WSDI* for duration [119].

HWMId is a dimensionless measure designed to account for both the duration and intensity of heat waves. It represents the maximum magnitude of heat waves occurring within a year. A heat wave is identified by the presence of at least three consecutive days where the daily maximum temperature exceeds the 90th percentile for that period, calculated over a 31-day window based on the reference period. The maximum magnitude is determined by summing the daily magnitudes of all the days that constitute the heat wave [117, 118].

WSDI is defined as the number of days per year with at least 6 consecutive days in which the maximum daily temperature is higher than the 90th percentile of the maximum daily temperature in a 5 days moving window during the reference period. This index solely accounts for the duration of heat waves, meaning that two heat waves of identical length are regarded as equally severe, even if their temperature exceedances relative to the reference period vary [117, 119].

The *WSDI* and *HWMId* indices are valuable tools for analyzing and characterizing extreme heat events, showing both consistent patterns and notable differences in their spatial distribution and temporal evolution [117]. It is suggested to employ multiple indices to capture the full scope of heat wave characteristics, including duration, magnitude, and intensity.

3.4.4. Climate change indices. Climate indices are characterized as a collection of relationships among various hydrometeorological variables, structured to evaluate their influence on potential climate change and to articulate its effects. These parameters may include aspects such as the frequency, intensity, or magnitude of precipitation and temperature extremes assessed on a monthly, seasonal, or annual basis, along with projections for future scenarios and their implications for climate change. Such indices serve to distill the complexities of climate change impacts into a more clear and measurable format, thereby enhancing comprehension for research communities across different sectors that are directly affected by these changes [120].

The World Meteorological Organization (WMO) Commission for Climatology, through its World Climate Research Programme (WCRP) project focused

on climate variability (CLIVAR), established an Expert Team on Climate Change, Detection, Monitoring and Indices (ETCCDMI). This team developed a fundamental set of twenty-seven indices based on daily maximum and minimum temperatures, as well as precipitation. Among the twenty-seven indices, sixteen pertain to temperature, while eleven are related to precipitation. These indices can be categorized into several types, including absolute, threshold, duration, and percentile indices, along with others that do not conform to these classifications [120].

4. CONCLUSIONS

In this paper, we have provided an overview of the current state of research on quality metrics in atmospheric tasks. An analysis of the existing scientific literature has shown a significant predominance of the "big three" classical metrics—bias, root mean square error, and correlation coefficient. Despite the universality of application to grid data, these metrics, like other pointwise methods, have a number of significant drawbacks, including the inability to isolate the quality of reproduction of the internal spatial correlation of fields and sensitivity to shifts in inhomogeneities. In addition to the study of classical metrics, we expanded our work with various complex metrics focused on different aspects of modeling.

The shortcomings of classical metrics allowed us to identify some general trends in the development of quality metrics. For example, among complex metrics, we devoted special attention to methods that take into account the spatial structure and inhomogeneities of meteorological fields, as well as data statistics. We also noted the importance of the emergence of new metrics that consider data using sliding windows; the development of methods for probabilistic assessment of the quality of ensemble forecasts; overcoming the problems inherent in extreme event metrics; and the development of algorithms for identifying scientifically important "objects" in spatial fields to create object-oriented metrics. Thus, we believe that all of the existing trends described in the field of quality metrics research are also directions for further research and development of new methods to move away from the widespread use of classical metrics and toward more complex metrics specific to various atmospheric problems.

As physics-based models are vital for assessing changes in Earth's climate system, evaluating their effectiveness is also challenging due to the distinct quality metrics used. Models are typically assessed by comparing hindcasts of past climate variables with historical observations, although this may involve parameter tuning. Climate projections, based on specific forcing scenarios, are compared with observations to test model skill and enhance confidence in future forecasts. Initiatives like the Coupled Model Intercomparison Project (CMIP) standardize experiments and collect data for evaluation. Various quality metrics exist to assess model performance, but a broader range and an optimal suite of these metrics are necessary.

In conclusion, the need for developing more diverse metrics for effective evaluation of climate and weather models is still in the centre of scientific interest.

FUNDING

This research was funded by the program PRIORITY 2030 of the Moscow Institute of Physics and Technology (National Research University). The review of quality metrics applied for climate models was carried out as part of the work under Agreement no. 075-03-2024-117 dated January 17, 2024.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

1. V. Yu. Rezvov, M. Krinitskiy, and S. Gulev, in *Proceedings of The 6th International Workshop on Deep Learning in Computational Physics—PoS(DLCP2022)* (2022), Vol. 429, p. 023. https://doi.org/10.22323/1.429.0023
2. H. Yang, J. Yan, Yo. Liu, and Z. Song, Global Energy Interconnect. **5**, 217 (2022). https://doi.org/10.1016/j.gloei.2022.04.018
3. K. A. Reed, N. Goldenson, R. Grotjahn, W. J. Gutowski, K. Jagannathan, A. D. Jones, L. R. Leung, S. A. McGinnis, S. C. Pryor, A. K. Srivastava, P. A. Ullrich, and C. M. Zarzycki, WIREs Clim. Change **13**, 799 (2022). https://doi.org/10.1002/wcc.799
4. J. J. Barsugli, G. Guentchev, R. M. Horton, A. Wood, L. O. Mearns, X. Liang, J. A. Winkler, K. Dixon, K. Hayhoe, R. B. Rood, L. Goddard, A. Ray, L. Buja, and C. Ammann, EOS, Trans. Am. Geophys. Union **94**, 424 (2013). https://doi.org/10.1002/2013eo460005
5. L. Briley, R. Kelly, E. D. Blackmer, A. V. Troncoso, R. B. Rood, J. Andresen, and M. C. Lemos, Bull. Am. Meteorol. Soc. **101**, E1709 (2020). https://doi.org/10.1175/bams-d-19-0099.1
6. K. Jagannathan, A. D. Jones, and I. Ray, Bull. Am. Meteorol. Soc. **102**, E1579 (2020). https://doi.org/10.1175/bams-d-19-0296.1

7. R. H. Moss, S. Avery, K. Baja, M. Burkett, A. M. Chischilly, J. Dell, P. A. Fleming, K. Geil, K. Jacobs, A. Jones, K. Knowlton, J. Koh, M. C. Lemos, J. Melillo, R. Pandya, T. C. Richmond, L. Scarlett, J. Snyder, M. Stults, A. M. Waple, J. Whitehead, D. Zarrilli, B. M. Ayyub, J. Fox, A. Ganguly, L. Joppa, S. Julius, P. Kirshen, R. Kreutter, A. McGovern, R. Meyer, J. Neumann, W. Solecki, J. Smith, P. Tissot, G. Yohe, and R. Zimmerman, Weather, Clim., Soc. **11**, 465 (2019). https://doi.org/10.1175/wcas-d-18-0134.1
8. Yu. Tian, G. S. Nearing, Ch. D. Peters-Lidard, K. W. Harrison, and L. Tang, Mon. Weather Rev. **144**, 607 (2016). https://doi.org/10.1175/mwr-d-15-0087.1
9. E. T. Jaynes, *Probability Theory: The Logic of Science*, Ed. by G. L. Bretthorst (Cambridge University Press, Cambridge, 2003).
10. A. G. Barnston and J. L. Thomas, J. Appl. Meteorol. Climatol. **22**, 2038 (1983). https://doi.org/10.1175/1520-0450(1983)022<2038:RMAIFA>2.0.CO;2
11. E. E. Ebert and J. L. McBride, J. Hydrol. **239**, 179 (2000). https://doi.org/10.1016/s0022-1694(00)00343-7
12. N. Collier, F. M. Hoffman, D. M. Lawrence, G. Keppel-aleks, Ch. D. Koven, W. J. Riley, M. Mu, and J. T. Randerson, J. Adv. Model. Earth Syst. **10**, 2731 (2018). https://doi.org/10.1029/2018ms001354
13. K. E. Taylor, J. Geophys. Res.: Atmos. **106**, 7183 (2001). https://doi.org/10.1029/2000jd900719
14. M. F. Wehner, Clim. Dyn. **40**, 59 (2013). https://doi.org/10.1007/s00382-012-1393-1
15. M. Franco-Torres, B. C. Rogers, and R. M. Ugarelli, Environ. Innovation Societal Transitions **36**, 34 (2020). https://doi.org/10.1016/j.eist.2020.04.010
16. S. Sarkki, H. I. Heikkinen, T. Komu, M. Partanen, K. Vanhanen, and É. Lépy, Sci. Public Policy **47**, 161 (2020). https://doi.org/10.1093/scipol/scz055
17. E. Turnhout, Sci. Public Policy **36**, 403 (2009). https://doi.org/10.3152/030234209x442007
18. Yo. Sun, K. Deng, K. Ren, J. Liu, Ch. Deng, and Yo. Jin, ISPRS J. Photogramm. Remote Sens. **208**, 14 (2024). https://doi.org/10.1016/j.isprsjprs.2023.12.011
19. S. Yeggina, R. S. V. Teegavarapu, and S. Muddu, Theor. Appl. Climatol. **140**, 1495 (2020). https://doi.org/10.1007/s00704-020-03175-7
20. H. Zandler, T. Senftl, and K. A. Vanselow, Sci. Rep. **10**, 22446 (2020). https://doi.org/10.1038/s41598-020-79480-y
21. L. Song, Ch. Xu, Yu. Long, X. Lei, N. Suo, and L. Cao, Remote Sens. **14**, 6039 (2022). https://doi.org/10.3390/rs14236039
22. M. A. Gilabert, A. Moreno, F. Maselli, B. Martínez, M. Chiesi, S. Sánchez-Ruiz, F. J. García-Haro, A. Pérez-Hoyos, M. Campos-Taberner, O. Pérez-Priego, P. Serrano-Ortiz, and A. Carrara, ISPRS J. Photogramm. Remote Sens. **102**, 184 (2015). https://doi.org/10.1016/j.isprsjprs.2015.01.017
23. H. Kikumoto, R. Ooka, Yu. Arima, and T. Yamanaka, Sustainable Cities Soc. **14**, 404 (2015). https://doi.org/10.1016/j.scs.2014.08.007
24. F. Z. Maina, E. R. Siirila-Woodburn, and P. Vahmani, Hydrol. Earth Syst. Sci. **24**, 3451 (2020). https://doi.org/10.5194/hess-24-3451-2020
25. P. Leng, Z.-L. Li, S.-B. Duan, M.-F. Gao, and H.-Y. Huo, ISPRS J. Photogramm. Remote Sens. **131**, 40 (2017). https://doi.org/10.1016/j.isprsjprs.2017.07.013
26. B. C. Hewitson and R. G. Crane, Clim. Res. **7**, 85 (1996). https://doi.org/10.3354/cr007085
27. R. L. Wilby and T. M. L. Wigley, Prog. Phys. Geogr.: Earth Environ. **21**, 530 (1997). https://doi.org/10.1177/030913339702100403
28. J. T. Schoof, Geogr. Compass **7**, 249 (2013). https://doi.org/10.1111/gec3.12036
29. A. Di Bucchianico, in *Encyclopedia of Statistics in Quality and Reliability* (Wiley, 2008), Vol. 1. https://doi.org/10.1002/9780470061572.eqr173
30. S. R. A. Fisher, *Statistical Methods for Research Workers* (Hafner Publishing, 1958).
31. D. S. Wilks, *Statistical Methods in the Atmospheric Sciences*, 3rd ed. (Academic, Amsterdam, 2011).
32. T. R. Stewart, Weather Forecast. **5**, 661 (1990). https://doi.org/10.1175/1520-0434(1990)005<0661:ADOTCC>2.0.CO;2
33. A. H. Murphy and E. S. Epstein, Mon. Weather Rev. **117**, 572 (1989). https://doi.org/10.1175/1520-0493(1989)117<0572:SSACCI>2.0.CO;2
34. E. Habib, W. F. Krajewski, and G. J. Ciach, J. Hydrometeorology **2**, 621 (2001). https://doi.org/10.1175/1525-7541(2001)002<0621:EORIC>2.0.CO;2
35. A. H. Murphy, B. G. Brown, and Y. S. Chen, Weather Forecast. **4**, 485 (1989). https://doi.org/10.1175/1520-0434(1989)004<0485:DVOTF>2.0.CO;2
36. G. S. Nearing and H. V. Gupta, Water Resour. Res. **51**, 524 (2015). https://doi.org/10.1002/2014wr015895
37. R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu, *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd ed. (Chapman and Hall/CRC, New York, 2006). https://doi.org/10.1201/9781420010138
38. S. A. McKeen, J. Wilczak, G. Grell, I. Djalalova, S. Peckham, E.-Y. Hsie, W. Gong, V. Bouchet, S. Menard, R. Moffet, J. McHenry, J. McQueen, Y. Tang, G. R. Carmichael, M. Pagowski, A. Chan,

T. Dye, G. Frost, P. Lee, and R. Mathur, J. Geophys. Res.: Atmos. **110**, D21307 (2005). https://doi.org/10.1029/2005jd005858
39. N. H. Savage, P. Agnew, L. S. Davis, C. Ordóñez, R. Thorpe, C. E. Johnson, F. M. O'Connor, and M. Dalvi, Geosci. Model. Dev. **6**, 353 (2013). https://doi.org/10.5194/gmd-6-353-2013
40. T. Chai and R. R. Draxler, Geosci. Model. Dev. **7**, 1247 (2014). https://doi.org/10.5194/gmd-7-1247-2014
41. C. J. Willmott and K. Matsuura, Clim. Res. **30**, 79 (2005). https://doi.org/10.3354/cr030079
42. C. J. Willmott, K. Matsuura, and S. M. Robeson, Atmos. Environ. **43**, 749 (2009). https://doi.org/10.1016/j.atmosenv.2008.10.005
43. T. O. Hodson, Geosci. Model. Dev. **15**, 5481 (2022). https://doi.org/10.5194/gmd-15-5481-2022
44. B. Casati, L. J. Wilson, D. B. Stephenson, P. Nurmi, A. Ghelli, M. Pocernich, U. Damrath, E. E. Ebert, B. G. Brown, and S. Mason, Meteorol. Appl. **15**, 3 (2008). https://doi.org/10.1002/met.52
45. M. G. Schultz, C. Betancourt, B. Gong, F. Kleinert, M. Langguth, L. H. Leufen, A. Mozaffari, and S. Stadtler, Philos. Trans. R. Soc., A **379**, 20200097 (2021). https://doi.org/10.1098/rsta.2020.0097
46. S. V. Weijs, G. Schoups, and N. Van De Giesen, Hydrol. Earth Syst. Sci. **14**, 2545 (2010). https://doi.org/10.5194/hess-14-2545-2010
47. W. Gong, H. V. Gupta, D. Yang, K. Sricharan, and A. O. Hero, Water Resour. Res. **49**, 2253 (2013). https://doi.org/10.1002/wrcr.20161
48. J. A. Vrugt and M. Sadegh, Water Resour. Res. **49**, 4335 (2013). https://doi.org/10.1002/wrcr.20354
49. E. E. Ebert, Meteorol. Appl. **15**, 51 (2008). https://doi.org/10.1002/met.25
50. M. Weniger, F. Kapp, and P. Friederichs, Q. J. R. Meteorol. Soc. **143** (702), 120 (2017). https://doi.org/10.1002/qj.2881
51. S. Buschow, J. Pidstrigach, and P. Friederichs, Geosci. Model. Dev. **12**, 3401 (2019). https://doi.org/10.5194/gmd-12-3401-2019
52. N. M. Roberts and H. W. Lean, Monthly Weather Review **136**, 78 (2008). https://doi.org/10.1175/2007mwr2123.1
53. S. Lerch, T. L. Thorarinsdottir, F. Ravazzolo, and T. Gneiting, Stat. Sci. **32**, 106 (2017). https://doi.org/10.1214/16-sts588
54. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, IEEE Trans. Image Process. **13**, 600 (2004). https://doi.org/10.1109/tip.2003.819861
55. A. M. Ebtehaj, E. Foufoula-georgiou, and G. Lerman, J. Geophys. Res.: Atmos. **117**, D08107 (2012). https://doi.org/10.1029/2011jd017057
56. A. Geiss, S. J. Silva, and J. C. Hardin, Geosci. Model. Dev. **15**, 6677 (2022). https://doi.org/10.5194/gmd-15-6677-2022
57. D. Cho, J. Im, and S. Jung, Q. J. R. Meteorol. Soc. **150** (760), 1222 (2024). https://doi.org/10.1002/qj.4643
58. X. Zhong, F. Du, L. Chen, Zh. Wang, and H. Li, Q. J. R. Meteorol. Soc. **150** (758), 275 (2024). https://doi.org/10.1002/qj.4596
59. C. Robertson, J. A. Long, F. S. Nathoo, T. A. Nelson, and C. C. F. Plouffe, Geogr. Anal. **46**, 53 (2014). https://doi.org/10.1111/gean.12028
60. R. Dosselmann and X. D. Yang, Signal, Image Video Process. **5**, 81 (2011). https://doi.org/10.1007/s11760-009-0144-1
61. H. V. Gupta, H. Kling, K. K. Yilmaz, and G. F. Martinez, J. Hydrol. **377**, 80 (2009). https://doi.org/10.1016/j.jhydrol.2009.08.003
62. Sh. A. Shahriar, M. A. M. Siddique, and Sh. M. A. Rahman, Meteorol. Atmos. Phys. **133**, 1409 (2021). https://doi.org/10.1007/s00703-021-00817-x
63. B. Böker, P. Laux, P. Olschewski, and H. Kunstmann, Int. J. Climatol. **43**, 5130 (2023). https://doi.org/10.1002/joc.8136
64. F. Wang, D. Tian, L. Lowe, L. Kalin, and J. Lehrter, Water Resour. Res. **57**, e2020WR029308 (2021). https://doi.org/10.1029/2020wr029308
65. H. Kling, M. Fuchs, and M. Paulin, J. Hydrol. **424–425**, 264 (2012). https://doi.org/10.1016/j.jhydrol.2012.01.011
66. S. Pool, M. Vis, and J. Seibert, Hydrol. Sci. J. **63**, 1941 (2018). https://doi.org/10.1080/02626667.2018.1552002
67. W. J. M. Knoben, J. E. Freer, and R. A. Woods, Hydrol. Earth Syst. Sci. **23**, 4323 (2019). https://doi.org/10.5194/hess-23-4323-2019
68. E. Hertig, D. Maraun, J. Bartholy, R. Pongracz, M. Vrac, I. Mares, J. M. Gutiérrez, J. Wibig, A. Casanueva, and P. M. M. Soares, Int. J. Climatol. **39**, 3846 (2019). https://doi.org/10.1002/joc.5469
69. J. Baño-Medina, R. Manzanas, and J. M. Gutiérrez, Clim. Dyn. **57**, 2941 (2021). https://doi.org/10.1007/s00382-021-05847-0
70. A. Vaughan, W. Tebbutt, J. S. Hosking, and R. E. Turner, Geosci. Model. Dev. **15**, 251 (2022). https://doi.org/10.5194/gmd-15-251-2022
71. A. C. O'Neill, L. H. Erikson, and P. L. Barnard, Earth Space Sci. **4**, 314 (2017). https://doi.org/10.1002/2016ea000193
72. H. Hersbach, Weather Forecast. **15**, 559 (2000). https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2
73. H. Jin, W. Jiang, M. Chen, M. Li, K. Sh. Bakar, and Q. Shao, Stochastic Environ. Res. Risk Assess. **37**, 3185 (2023). https://doi.org/10.1007/s00477-023-02444-x

74. J. Brocker, Q. J. R. Meteorol. Soc. **138** (667), 1611 (2012).
https://doi.org/10.1002/qj.1891
75. M. J. Alessi and A. T. DeGaetano, Meteorol. Appl. **28**, e1976 (2021).
https://doi.org/10.1002/met.1976
76. L. Harris, A. T. T. McRae, M. Chantry, P. D. Dueben, and T. N. Palmer, J. Adv. Model. Earth Syst. **14**, e2022MS003120 (2022).
https://doi.org/10.1029/2022ms003120
77. A. Yu. Bundel, A. V. Muraviev, and E. D. Olkhovaya, Gidrometeorologicheskie Issledovaniya i Prognozy, No. 4, 30 (2021).
https://doi.org/10.37162/2618-9631-2021-4-30-49
78. N. M. Roberts, Meteorol. Appl. **15**, 163 (2008).
https://doi.org/10.1002/met.57
79. E. E. Ebert, L. Wilson, A. Weigel, M. Mittermaier, P. Nurmi, P. Gill, M. Göber, S. Joslyn, B. Brown, T. Fowler, and A. Watkins, Meteorol. Appl. **20**, 130 (2013).
https://doi.org/10.1002/met.1392
80. G. Skok and N. Roberts, Q. J. R. Meteorol. Soc. **144** (711), 414 (2018).
https://doi.org/10.1002/qj.3212
81. N. Faggian, B. Roux, P. Steinle, and B. Ebert, Mausam **66**, 457 (2015).
https://doi.org/10.54302/mausam.v66i3.555
82. M. Mittermaier, N. Roberts, and S. A. Thompson, Meteorol. Appl. **20**, 176 (2013).
https://doi.org/10.1002/met.296
83. G. Skok, Meteorol. Appl. **22**, 477 (2015).
https://doi.org/10.1002/met.1478
84. J. A. Weyn and D. R. Durran, Q. J. R. Meteorol. Soc. **145** (S1), 57 (2019).
https://doi.org/10.1002/qj.3367
85. L. Glawion, J. Polz, H. Kunstmann, B. Fersch, and Ch. Chwala, Earth Space Sci. **10**, e2023EA002906 (2023).
https://doi.org/10.1029/2023ea002906
86. L. L. De Haan, M. Kanamitsu, F. De Sales, and L. Sun, Theor. Appl. Climatol. **122**, 47 (2015).
https://doi.org/10.1007/s00704-014-1278-9
87. E. S. Gentile, S. L. Gray, and H. W. Lewis, Q. J. R. Meteorol. Soc. **148** (743), 685 (2022).
https://doi.org/10.1002/qj.4225
88. V. V. Kharin and F. W. Zwiers, J. Clim. **16**, 4145 (2003).
https://doi.org/10.1175/1520-0442(2003)016<4145:OTRSOP>2.0.CO;2
89. R. Buizza and T. N. Palmer, Mon. Weather Rev. **126**, 2503 (1998).
https://doi.org/10.1175/1520-0493(1998)126<2503:IOESOE>2.0.CO;2
90. S. J. Mason and N. E. Graham, Weather Forecast. **14**, 713 (1999).
https://doi.org/10.1175/1520-0434(1999)014<0713:CPROCA>2.0.CO;2
91. R. Manzanas, M. D. Frías, A. S. Cofiño, and J. M. Gutiérrez, J. Geophys. Res.: Atmos. **119**, 1708 (2014).
https://doi.org/10.1002/2013jd020680
92. Ch. Davis, B. Brown, and R. Bullock, Mon. Weather Rev. **134**, 1772 (2006).
https://doi.org/10.1175/mwr3145.1
93. D. B. Stephenson, B. Casati, C. A. T. Ferro, and C. A. Wilson, Meteorol. Appl. **15**, 41 (2008).
https://doi.org/10.1002/met.53
94. A. H. Murphy, Weather Forecast. **6**, 302 (1991).
https://doi.org/10.1175/1520-0434(1991)006<0302:POAFOR>2.0.CO;2
95. Z. Hausfather, H. F. Drake, T. Abbott, and G. A. Schmidt, Geophys. Res. Lett. **47**, e2019GL085378 (2020).
https://doi.org/10.1029/2019gl085378
96. A. Gettelman, C. Hannay, J. T. Bacmeister, R. B. Neale, A. G. Pendergrass, G. Danabasoglu, J.-F. Lamarque, J. T. Fasullo, D. A. Bailey, D. M. Lawrence, and M. J. Mills, Geophys. Res. Lett. **46**, 8329 (2019).
https://doi.org/10.1029/2019gl083978
97. N. C. Baker and P. C. Taylor, J. Clim. **29**, 1773 (2016).
https://doi.org/10.1175/jcli-d-15-0114.1
98. C. J. Willmott, Bull. Am. Meteorol. Soc. **63**, 1309 (1982).
https://doi.org/10.1175/1520-0477(1982)063<1309:SCOTEO>2.0.CO;2
99. R. Knutti, Clim. Change **102**, 395 (2010).
https://doi.org/10.1007/s10584-010-9800-2
100. A. Rahman and S. Pekkat, Sci. Rep. **14**, 3076 (2024).
https://doi.org/10.1038/s41598-024-52275-1
101. D. Walton and A. Hall, J. Clim. **31**, 3789 (2018).
https://doi.org/10.1175/jcli-d-17-0410.1
102. J. Lindvall and G. Svensson, Clim. Dyn. **44**, 405 (2015).
https://doi.org/10.1007/s00382-014-2144-2
103. K. Braganza, D. J. Karoly, and J. M. Arblaster, Geophys. Res. Lett. **31**, L13217 (2004).
https://doi.org/10.1029/2004gl019998
104. D. B. Lobell, C. Bonfils, and P. B. Duffy, Geophys. Res. Lett. **34**, L05715 (2007).
https://doi.org/10.1029/2006gl028726
105. J. M. Gutiérrez, D. Maraun, M. Widmann, R. Huth, E. Hertig, R. Benestad, O. Roessler, J. Wibig, R. Wilcke, S. Kotlarski, D. San Martín, S. Herrera, J. Bedia, A. Casanueva, R. Manzanas, M. Iturbide, M. Vrac, M. Dubrovsky, J. Ribalaygua, J. Pórtoles, O. Räty, J. Räisänen, B. Hingray, D. Raynaud, M. J. Casado, P. Ramos, T. Zerenner, M. Turco, T. Bosshard, P. Štěpánek, J. Bartholy, R. Pongracz, D. E. Keller, A. M. Fischer, R. M. Cardoso, P. M. M. Soares, B. Czernecki, and C. Pagé, Int. J. Climatol. **39**, 3750 (2019).
https://doi.org/10.1002/joc.5462

106. J. H. Christensen, E. Kjellström, F. Giorgi, G. Lenderink, and M. Rummukainen, Clim. Res. **44**, 179 (2010).
https://doi.org/10.3354/cr00916
107. D. Maraun, R. Huth, J. M. Gutiérrez, D. S. Martín, M. Dubrovsky, A. Fischer, E. Hertig, P. M. M. Soares, J. Bartholy, R. Pongrácz, M. Widmann, M. J. Casado, P. Ramos, and J. Bedia, Int. J. Climatol. **39**, 3786 (2019).
https://doi.org/10.1002/joc.5222
108. S. E. Perkins, A. J. Pitman, and S. A. Sisson, Int. J. Climatol. **33**, 1153 (2013).
https://doi.org/10.1002/joc.3500
109. A. D. Polasky, J. L. Evans, and J. D. Fuentes, Environ. Modell. Software **165**, 105712 (2023).
https://doi.org/10.1016/j.envsoft.2023.105712
110. S. E. Perkins, A. J. Pitman, N. J. Holbrook, and J. McAneney, J. Clim. **20**, 4356 (2007).
https://doi.org/10.1175/jcli4253.1
111. D. Kiktev, D. M. H. Sexton, L. Alexander, and C. K. Folland, J. Clim. **16**, 3560 (2003).
https://doi.org/10.1175/1520-0442(2003)016<3560:COMAOT>2.0.CO;2
112. Yv. Tramblay, W. Badi, F. Driouech, S. El Adlouni, L. Neppel, and E. Servat, Global Planet. Change **82–83**, 104 (2012).
https://doi.org/10.1016/j.gloplacha.2011.12.002
113. D. A. Darling, Ann. Math. Stat. **28**, 823 (1957).
https://doi.org/10.1214/aoms/1177706788
114. F. Laio, Water Resour. Res. **40**, W09308 (2004).
https://doi.org/10.1029/2004wr003204
115. S. Kotlarski, P. Szabó, S. Herrera, O. Räty, K. Keuler, P. M. Soares, R. M. Cardoso, T. Bosshard, Ch. Pagé, F. Boberg, J. M. Gutiérrez, F. A. Isotta, A. Jaczewski, F. Kreienkamp, M. A. Liniger, C. Lussana, and K. Pianko-kluczyńska, Int. J. Climatol. **39**, 3730 (2019).
https://doi.org/10.1002/joc.5249
116. P. M. M. Soares, D. Maraun, S. Brands, M. W. Jury, J. M. Gutiérrez, D. San-martín, E. Hertig, R. Huth, A. Belušić Vozila, R. M. Cardoso, S. Kotlarski, P. Drobinski, and A. Obermann-hellhund, Int. J. Climatol. **39**, 3868 (2019).
https://doi.org/10.1002/joc.5911
117. M. O. Molina, E. Sánchez, and C. Gutiérrez, Sci. Rep. **10**, 8801 (2020).
https://doi.org/10.1038/s41598-020-65663-0
118. S. Russo, J. Sillmann, and E. M. Fischer, Environ. Res. Lett. **10**, 124003 (2015).
https://doi.org/10.1088/1748-9326/10/12/124003
119. X. Zhang, L. Alexander, G. C. Hegerl, P. Jones, A. K. Tank, T. C. Peterson, B. Trewin, and F. W. Zwiers, WIREs Clim. Change **2**, 851 (2011).
https://doi.org/10.1002/wcc.147
120. W. Ahmad, International Journal for Research in Applied Science and Engineering Technology **8**, 876 (2020).
https://doi.org/10.22214/ijraset.2020.31604

AI tools may have been used in the translation or editing of this article.