Bergen Metrics: composite error metrics for assessing performance of climate models using EURO-CORDEX simulations

Alok K. Samantaray, Priscilla A. Mooney, Carla A. Vivacqua

We appreciate the comments from the reviewers as they have made some important points that have been addressed in the revised manuscript.

I enjoyed reading the ms. However, I have several concerns listed below:

We thank the reviewer for the comment. In the revised manuscript, we have addressed all the comments suggested by the reviewer.

L65: “38 different metrics” the reader can be curious where are they or how they are obtained from Eq7? Explain in more detail on section 3.3.

We thank the reviewer for the comment. In this study, we employed a clustering approach to group the 38 different error metrics. Randomly selecting one metric from each cluster, we utilized them in Equation 7 to calculate the Bergen metric. For enhanced clarity on this process, we have incorporated a flowchart in the revised draft. This flowchart provides a step-by-step illustration of how the Bergen metric is computed using different error metrics.

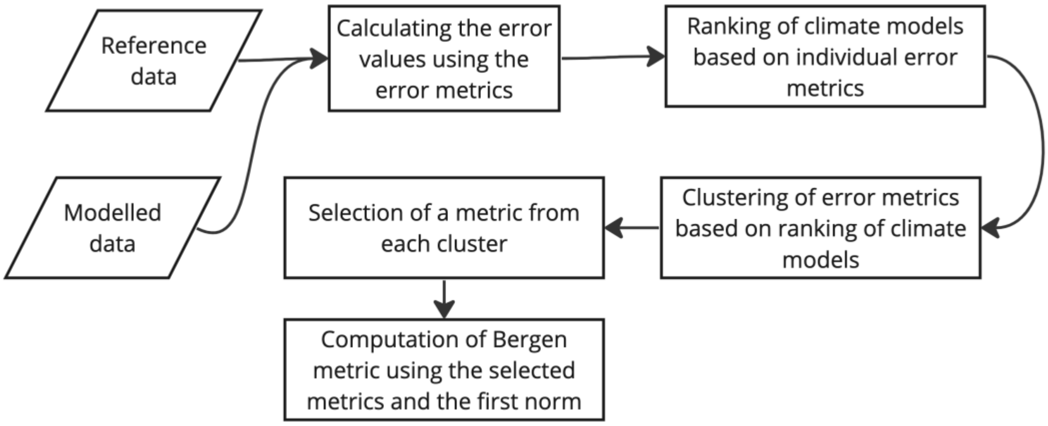
Reference data
Modelled data
Calculating the error values using the error metrics
Ranking of climate models based on individual error metrics
Clustering of error metrics based on ranking of climate models
Selection of a metric from each cluster
Computation of Bergen metric using the selected metrics and the first norm

Figure R1: The flowchart for the calculation of Bergen metric

Eq 4-5-6-7: listing all four forms of the BM is confusing only Eq7 is enough for the reader.

We appreciate the reviewer's comment. As the number of metrics employed in Equation 4, 5 and 6 differs from Equation 7, we have retained Equation 4, and 7 while removing Equation 5 and 6.

Eq7: none of the components are innovative or directly related to distribution. They are all bias-sensitive components and may have overlapping (redundant) information.

We appreciate the reviewer's feedback. In the draft, we explicitly state that our study utilizes established error metrics, examining their focus on specific aspects of model and reference data and how they contribute to the ranking of climate models. A crucial assumption in our study is the recognition of the importance of all error metrics, each carrying its own set of advantages and disadvantages. The study emphasizes the significance of error magnitude, capturing overlapping information, and clustering metrics that target similar error aspects. This allows us to streamline the calculation of the Bergen metric by avoiding redundant contributions from metrics in the same cluster.

In a recent HESS paper, spatial patterns of GCM/RCMs are used to the ranking.

https://doi.org/10.5194/hess-23-4803-2019

We have cited this paper in our modified draft.

Applying cluster analysis doesn’t add much to the novelty of the metric. Each component must add something to the Bergen Metric.

We appreciate the reviewer's input. The primary objective of our study is to delve into the characteristics of error metrics, examining how they address diverse aspects of the relationship between observed and reference data errors. The disparity in error magnitudes and model rankings, as demonstrated in the study, can potentially lead to confusion among users. With an abundance of metrics available in the literature, our study does not seek to introduce a new one but rather aims to streamline the existing metrics by data analysis. Clustering plays a pivotal role in this process, especially when dealing with metrics that, despite targeting similar error types, exhibit significant data-oriented variations. Take, for instance, various versions of root mean squared error – in the absence of outliers, they remain in the same cluster, while the presence of outliers assigns them to different clusters. The study operates under the assumption that all error metrics are crucial. By integrating them into a composite metric without overlap (achieved through clustering) and subsequently analyzing individual error metrics, we can gain enhanced insights into the relationship between observed and modeled data.

One of the components in Eq7 should be histogram match (overlap %) for better discrimination power of the metric.

We appreciate the valuable feedback provided by the reviewer. Acknowledging the significance of histogram matching as an essential metric for model evaluation, we would like to highlight that Equation 7 (Bergen metric) has a flexible framework. This flexibility allows the incorporation of new metrics into the evaluation process. Recognizing the abundance of metrics available in the literature, the inclusion of additional metrics in the clustering analysis is anticipated to enhance our understanding of model performance.

Also the metric should be insensitive to the unit-differences since observation and simulation may have different units. GCM to GCM comparison can be smooth and not unit issue; however, observed AET from MODIS is watt/m2 whereas hydrologic model flux simulations are in mm/day. Then ,Bergen metric cannot be applied to other hydroclimatologic problems.

We appreciate the reviewer's comment. Firstly, it's important to note that the Bergen metric is unitless, as the error metrics used in its computation are normalized across all models. This characteristic enables its application to any hydroclimatologic problem, provided that the necessary error metrics can be calculated.

The unit of Actual Evapotranspiration (AET) is mm/day, while the latent heat flux has a unit of watt/m^2, which is derived from various remotely sensed data. The linear relationship between latent heat flux and water flux (AET) involves a constant, which is the product of the latent heat of vaporization of water and water density. The interchangeability of units is possible due to this linear relationship. However, if the units are not linearly related, it could lead to complex issues, as there would be no basis for comparison.

It's worth noting that not all metrics need to be independent of units. Many widely used error metrics in climate studies, such as Mean Squared Error and Root Mean Squared Error, are unit-dependent. Both unit-dependent and unit-independent error metrics have their advantages and disadvantages, as highlighted by Hyndman (2006). Therefore, in this study, we consider all error metrics to be important and suggest incorporating them based on the specific requirements and underlying data of individual studies.

Hyndman, R. J. (2006). Another look at forecast-accuracy metrics for intermittent demand. Foresight: The International Journal of Applied Forecasting, 4(4), 43-46.

EOF, SSIM, FSS, SPEM and SPAEF metrics should be covered in the literature review.

We thank the reviewer for the comment. These methods have been referenced in the modified draft.

‘In addition to this, researchers have employed various characteristics of climatic parameters as measures to assess and compare climate models with observed datasets. Metrics encompassing the frequency of days with precipitation over 1 mm and over 15 mm, the 90% quantile of the frequency distribution, and the maximum number of consecutive dry days, along with parameters such as daily mean, daily maximum, daily minimum, yearly maximum, length of the frost-free period, growing degree days (> 5°C), cooling degree days (> 22°C), heating degree days (< 15.5°C), days with RR (> 99th percentile of daily amounts for all days), ratio of spatial variability, pattern correlation, ratio of interannual variability, temporal correlation of interannual variability, number of summer days, number of frost days, consecutive dry days, and ratio of yearly amplitudes, have been utilized for the validation of Euro-CORDEX data (Kotlarski et al., 2014; Giot et al., 2016; Smiatek et al., 2016; Torma, 2019; Vautard et al., 2021). Other studies have employed the empirical orthogonal functions (Rasmus et al., 2023), structural similarity index metric (Wang & Bovik, 2002), fractions skill score (Roberts & Lean, 2008), spatial pattern efficiency metric (Dembélé et al., 2020), spatial efficiency metric (Demirel, 2018) and probability distribution function (Perkins et al., 2007; Boberg et al., 2009; Boberg et al., 2010; Masanganise et al., 2014) to evaluate climate models.’