Regarding the response to the reviewers' comment on our manuscript entitled "pyESDv1.0.1: An open-source Python framework for empirical-statistical downscaling of climate information" by Daniel Boateng and Sebastian G. Mutz.

Dear Dr. Charles Onyutha,

We would like to thank the editor for their constructive comments and time for highlighting parts of the manuscript that require further changes and improvement. We have addressed each of the comments and suggestions and included the relevant references. We reply to the comments below:

COMMENT 1

The metrics RMSE and MAE in Eqs. (13)-(14) all depend on the squares of the residuals. This renders both RMSE and MAE susceptible to the effects of outliers in the data. This setback should be highlighted to the readers. Importantly, to conform to the definition of MAE which you provided in Eq. (14), the brackets in front of the summation sign should not be squared and you need to replace the brackets with the modulus symbol, '| |' to indicate that you are dealing with absolute values. Furthermore, it would be good to recommend that the interpretation of RMSE and MAE should be made with respect to the mean of the observations.

We thank the editor for raising such important points regarding the limitations associated with the RMSE and MAE metrics. We have highlighted these limitations in the text to inform readers. Furthermore, we appreciate the editor for accurately pointing out errors in the MAE equation. The equation that was initially presented was for the mean squared error, which is also included in pyESD. However, in this manuscript, we exclusively employed the MAE. We have made the necessary adjustments to the equation accordingly (see lines 445-458).

Additional metrics such as the Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), Maximum Error, Adjusted R-squared (Miles, 2014), and Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe, 1970) are included in pyESD. However, the predicted values from the trained model and their corresponding observed values can be evaluated using other metrics not included in pyESD. For example, additional metrics like the model skill score E and the revised R-squared (RRS), which combines correlation, bias measure, and the capacity to capture variability, can be used (Onyutha, 2021). We highlight that the limitations and assumptions underpinning these metrics should be considered when interpreting performance metrics. For example, the RMSE is sensitive to outliers because the squaring of errors assigns more weight to large errors. This implies that a single outlier can bias its estimate and lead to a misinterpretation of extreme data points in the predictand. Although MAE is less sensitive to outliers compared to RMSE, its treatment of all errors with equal weight may not adequately account for the impact of extreme errors on model performance. Consequently, both metrics should be interpreted with respect to the mean of the observed values. On the other hand, the Pearson Correlation Coefficient (PCC) assumes a linear relationship between the predicted and observed values and a bivariate normal distribution. However, distance correlation (Székely et al., 2007), which is more computationally demanding and makes no assumptions about the relationship or distribution, can be considered. Chaudhuri and Hu (2019) demonstrated a fast algorithm that can be used to compute the distance correlation.

COMMENT 2

The metric PCC in Eq. (12) assumes that the values of the two variables have linear relationship. This assumption needs to be checked using a simple scatter plot to check the suitability of a linear model to fit through the points. If the assumption is violated, the metric PCC gets affected in terms of its accuracy. Furthermore, such a classical measure such as PCC cannot detect dependence between the two variables under consideration even if they are highly dependent (Székely et al. 2007). Here, distance correlation (Székely et al. 2007) can be a plausible option since its accuracy is regardless of the assumed requirement of a linear relationship. A fast computation algorithm of Chaudhuri and Hu (2019) can be recommended for computing the distance correlation.

Székely G. J., Rizzo M. L. & Bakirov N. K. 2007 Measuring and testing independence by correlation of distances. The Annals of Statistics 35 (6), 2769–2794.

Chaudhuri A. & Hu W. 2019 A fast algorithm for computing distance correlation. Computational Statistics & Data Analysis 135, 15–24.

We appreciate the editor for bringing this to our attention. We would like to emphasize that the Pearson Correlation Coefficient (PCC) is determined using a scatter plot, as illustrated in Figure 6. It is not actually used as the model evaluation metric beyond the training and validation period. The PCC assesses the linear relationship between continuously predicted and observed values, as depicted in Figure 6 for the model validation period, but it was not employed as a model evaluation metric for the independent testing datasets. However, we have acknowledged its limitations and highlighted the best alternative, as suggested in the manuscript (see lines 445-458).

COMMENT 3

There are several variants of coefficient of determination (R-squared) as elaborated by Kvålseth (1985). Actually, the version of the R-squared presented as Eq. (11) in this manuscript corresponds to the well-known Nash Sutcliffe Efficiency (NSE) (Nash and Sutcliffe, 1970) especially in hydrology. The authors should cite the correct original paper for the version of the R-squared they applied in their manuscript.

Kvålseth, T. O. 1985 Cautionary note about R2. The American Statistician 39 (4), 279–285.

Nash, J. E. & Sutcliffe, J. V. 1970 River flow forecasting through conceptual models part I – a discussion of principles. Journal of Hydrology 10, 282–290.

We acknowledge that there are variations of R-squared, as described in Kvålseth (1985). However, we want to emphasize that the metric reported in this manuscript is the fraction of variance explained, which is calculated as 1 minus the ratio of the variance of model residuals (Sum Squared Error) to the variance of the observed values (Sum Squared Totals). This metric is similar to the Nash-Sutcliffe Efficiency (NSE) when the bias is zero. In our analysis, we have included both metrics, and our preliminary results (not shown) indicate no significant difference between them. Additionally, in the manuscript, we have highlighted the inclusion of other metrics in pyESD and

mentioned that alternative metrics like the model skill score (E) and the revised R-squared (RRS) suggested by Onyutha (2021) can also be used as performance metrics.

Additional private note (visible to authors and reviewers only):
COMMENT 4
The setbacks of NSE are well-known in literature (see e.g. Garrick et al. 1978; Legates and McCabe G. J. 1999). Again, for instance, read the abstract of the paper in Hydrology Research (2022) 53 (1): 51–64. In this paper two versions of R-squared were eventually proposed to address the issues of the original version of the R-squared. Such an information is relevant for the readers or users of pyESD. The relevant codes can be obtained in the supplementary material of the said paper (or DOI: 10.5281/zenodo.6570904).

Garrick M., Cunnane C. & Nash J. E. 1978 A criterion of efficiency for rainfall-runoff models. Journal of Hydrology 36, 375–381.

Legates D. R. & McCabe G. J. 1999 Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. Water Resources Research 35, 233–241.

We thank the editor for the additional references on evaluation metrics, which we have recommended them in the manuscript (see lines 445-450).

The submission file consists of the revised manuscript (with tracked changes) specifying all the modifications made in accordance with the editor's comments.

Please contact us if further clarifications are required.

Sincerely,
Daniel Boateng and S. G. Mutz