Authors Reply to RC1: 'Comment on egusphere-2025-4343', Richard Stumpf, 20 Oct 2025

*Reviewer comments appear in **bold**. Authors' responses are in plain text and indented for clarity.

The paper proposes a strategy for algorithm comparison/evaluation by designing a single metric to combine multiple metrics. This is a solid progression from previous work (referenced) that looked at metrics for algorithm assessment. The "Euclidan Distance Score" (EDS) is a strong approach to summarize the data. A critical objective of the authors is to identify only the metrics that are relevant, and summarize those, rather than to include lots of (often closely related) metrics and leave it to the reader to make sense of them. I will say that this paper was a pleasure to review, and it will become an excellent paper that should be quite important (and hopefully well used). But it does need revision to make sure it is correct.

A concern with comparing metrics is how to "normalize" those metrics that have quite disparate ranges. This approach addresses it by treating ratios & proportions, and so are unitless. That provides a good approach that is not arbitrary. While it does not force results to be between 0 and 1, it is set up with two strong conditions. An EDS = 1 is "perfect". Any EDS < 0 is unacceptably poor, and each of the input parameters to the EDS are typically going to be between 0 and 1. The ones that are not (proportional slope deviation, proportional error, and proportional bias), are really unacceptable if the values exceed 1.

I have two large concerns that should be directly solvable. First: the parameters to input. Second is whether the configuration of the equation parameters is correct.

> We thank the reviewer for the encouraging evaluation of our proposed method and for the constructive assessment and suggestions. The identification of the key issues is appreciated. We address each group of comments in detail below.

The inputs are R (Pearson correlation coefficient), linear regression slope calculated in log space (m), median ratio error (e ~ epsilon), Median ratio bias (B ~ beta), and valid retrieval ratio (n).

The question is: are these all robust and independent? Of these, e, B, and n are quite good. It is true that e and B are not actually independent, but as there appears to be no robust means of separating the two (de-biasing the error means calculating mean errors, rather than median errors, which gets into non-robust methods), so we will go with it.

As a practical matter a competent product should tend toward a bias ratio of 1. If it does not, then it is punished relatively severely, as e >= B. A biased "low error" model will probably do worse than an unbiased relatively high error model. This should be noted in the paper.

> We agree that, in practice, a good retrieval result should tend toward negligible systematic bias ($\beta \approx 0$; note that β is already rescaled to deviation instead of ratio in our proposed formulation. More details on how we make that more clear in answer below). In the EDS framework, the typical magnitude error (ϵ) and the systematic bias (β) are both derived from the same accuracy ratio $Q = E/O$ and, by construction, $|\beta| \leq \epsilon$, with equality occurring when deviations are purely systematic.

This dependence implies that bias-driven degradations tend to reduce EDS through two pathways. For example, adding a constant offset of $+0.05$ to all log accuracy ratios in a tight, unbiased distribution:

$$\log_{10}(Q_i) = [-0.03, -0.01, 0, 0.01, 0.03],$$

yields:

$$\log_{10}(Q_i)' = [0.02, 0.04, 0.05, 0.06, 0.08],$$

for which β increases from 0 to approximately 12% and ϵ increases from approximately 2% to 12%. Assuming a fixed $n = 0.98$, the corresponding EDS decreases from 0.97 to 0.83.

In contrast, increasing magnitude error through additional scatter without introducing bias, for example:

$$\log_{10}(Q_i)' = [-0.08, -0.06, 0, 0.06, 0.08],$$

increases ϵ from 2% to approximately 15% while β remains zero, resulting in a smaller EDS decrease from 0.97 to 0.85. This illustrates that systematic bias can penalize EDS more strongly than a comparable increase in magnitude error from unbiased scatter, consistent with the reviewer's observation.

We note, however, that this behaviour does not imply that EDS systematically favours unbiased retrievals. For instance, at fixed $n = 0.98$, a case with $\epsilon = \beta = 10\%$ yields a higher EDS (0.86) than a case with $\epsilon = 20\%$ and $\beta = 0\%$ (EDS = 0.80), demonstrating that total error magnitude dominates the score in this case.

We will include this discussion in the revised manuscript, alongside a thorough sensitivity analysis (more details on Reply to Reviewer 3).

At lines 24-28 the paper notes the problem of using root-mean-square error metrics. This is a critical point. Basically, the paper sets out that robust metrics should be used, which is why the paper proposed median e and B. However, Pearson regression and linear regression slope are least squares solutions. Thiel-Sen slope, or an equivalent, should be used for the slope. This is necessary, as many optical models (or for that matter, many models) often deviate at very low or very high values. That statistical leverage will severely alter a least squares regression slope, but not a robust slope metric.

Regression as a metric has an additional critical flaw: it normalizes to the standard deviation of the data. Therefore, an exact subset of a population that has a smaller range will have a lower R value than the population. (Worse, as observed in Seegers et al., a low error method with a small range of data will have a lower R values than a higher error method with a much larger range of data.) This problem is also seen in Figure 3. Oligotrophic water has the smallest error, but a low R value. The problem is the narrow range of data. Conversely if the range is large enough, R provides no useful information, both good and poor models can have high R values. Because of this problem, including R means that EDS values are not be comparable across the different data sets. (There is a good discussion of the problem of R by a top statistician https://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/10/lecture-10.pdf).

By the way, R and linear regression slope are not independent, slope = S_y/S_x * R.

As to the input metrics, based on appropriate and consistently robust metrics, the appropriate ones would then appear to be

1 median (Thiel=Sen) slope, to capture whether the data generally behaves well across the range. (I will say that I don't really like slope, but I do not see a better option, as that would involve more complex partitioning alternatives that are difficult to standardize.)

2 median error

3 median bias

4 retrievals n.

We thank the reviewer for raising this important discussion and for directing us to relevant literature. We acknowledge the concerns raised regarding regression-based diagnostics, in particular their sensitivity to data range and leverage effects, which can compromise comparability across datasets.

These considerations directly motivated the revisions to the EDS framework. In the revised formulation, neither regression slope nor correlation coefficient are retained as components of the composite score. We adopted a conservative approach and restricted the EDS to three metrics: a robust measure of error magnitude (ϵ), a robust measure of systematic bias (β), and a term capturing retrieval feasibility (n).

While robust slope estimators such as the Theil–Sen method mitigate sensitivity to outliers and leverage, slope estimates remain inherently dependent on the range of the evaluated data regardless of the regression method employed. When values span a narrow range (e.g., within a single optical water type), slope estimates become poorly conditioned and associated with increased uncertainty, such that small data perturbations can lead to large variations in the estimated slope. Consequently, slopes are not always reliable as performance metrics or directly comparable across datasets or stratifications. In addition, our redundancy analysis (see reply to reviewer 3) indicated that including any slope-based diagnostic introduces overlap with bias-related metrics.

By excluding regression-based diagnostics and focusing on robust, range-independent measures, the revised EDS avoids the limitations highlighted by the reviewer and improves comparability across datasets and stratifications. We have also expanded the discussion about the rationale for metric selection and this will be included the revised manuscript.

Median error and bias do not appear to be correctly specified in EDS equation (7). As these are ratios, shouldn't they be $(e - 1)^2$ and $(B-1)^2$? Both are defined as a ratio of E/O (expected/observed), so a value of 1, is perfect, and should reduce to zero. Equation would be:

EDS = 1 – sqrt [$(m-1)^2$ + $(e-1)^2$ + $(B-1)^2$ + $(n-1)^2$]

As defined in Equations (3) and (5), ϵ and β are expressed as proportional deviations from unity (ratio minus one) and are therefore zero at perfect agreement. This formulation allows zero-centred deviations to be combined directly in Equation (7) without additional transformation.

The ratio-minus-one form was chosen to preserve interpretability: for example, a value of 0.10 directly indicates a 10% deviation, avoiding potentially confusing statements such as “an error of 1.1 corresponds to a 10% deviation.”

To be as clear as possible, we will make it more explicit in the revised manuscript that ϵ and β are defined as zero-centred proportional deviations. The proposed description of the metrics to be included is the following:

“*Building on the concept of the accuracy ratio introduced by Tofallis (2015), Morley et al. (2018) proposed a set of metrics designed for variables that span several orders of magnitude. These metrics are based on the logarithm of the accuracy ratio:*

$$log(Q_i) = log\left(\frac{E_i}{O_i}\right)$$

To quantify typical error magnitude while ensuring symmetry between over- and under-estimation, the absolute value of the logarithmic accuracy ratio is considered. Interchanging estimated and observed values therefore yields the same error magnitude. These values are aggregated across all estimation–observation pairs using the median, providing robustness to skewed distributions and outliers:

$$M = median(|log_{10}(Q_i)|)$$

The aggregated value is exponentiated to return to multiplicative space and shifted relative to the ideal ratio of unity by subtracting one, yielding the Median Symmetric Accuracy (ϵ):

$$\epsilon = 10^M - 1$$

This formulation produces an unsigned, ***zero-centred measure of typical proportional deviation from perfect agreement****, directly interpretable as a fractional (or percentage) error.*

Using the same underlying quantity, systematic bias is quantified by taking the median of the signed logarithmic accuracy ratio:

$$M' = median(log_{10}(Q_i))$$

And defining the Symmetric Signed Percentage Bias (β) as:

$$\beta = sign(M)\left(10^{|M'|} - 1\right)$$

where the sign indicates systematic over- or under-estimation and the magnitude reflects the ***typical proportional bias relative to the ideal value of zero****.*”

The authors might ponder thought experiments as examples (suggestion only). I did only one. An algorithm that has all results on an exact line with a slope of 1, but is severely biased. Error (e ~ epsilon) and (B ~ beta) will be equal. If the bias is 2x, which is a low performance, the EDS would return a value of zero.

We thank the reviewer for this constructive suggestion. We agree that thought experiments can be valuable for illustrating the behaviour and interpretation of the EDS.

To support this, we have generated a geometric representation of the revised EDS in a three-dimensional (β, ϵ, n) space (Figure 1), following the exclusion of regression slope and Pearson correlation coefficient from the score. This representation explicitly shows the domain of admissible metric combinations and how the EDS varies as a function of error magnitude, systematic bias, and retrieval robustness.

The reviewer's example of an algorithm producing estimates that lie on an exact line with slope one but are severely biased corresponds to the case $\epsilon = |\beta| = 1$ i.e. a systematic multiplicative bias of a factor of two. For this configuration, considering $n = 1$, the revised EDS would be approximately −0.41. This value reflects a retrieval that is strongly inaccurate and can also be used to motivate practical benchmark value, for instance illustrating that an EDS around −0.4 corresponds to a consistently biased retrieval with errors on the order of 100%, which would generally be regarded as poor performance in practice.

In the revised manuscript, we will explicitly include such thought experiments derived from the geometric representation to illustrate how different performance regimes map onto the EDS.

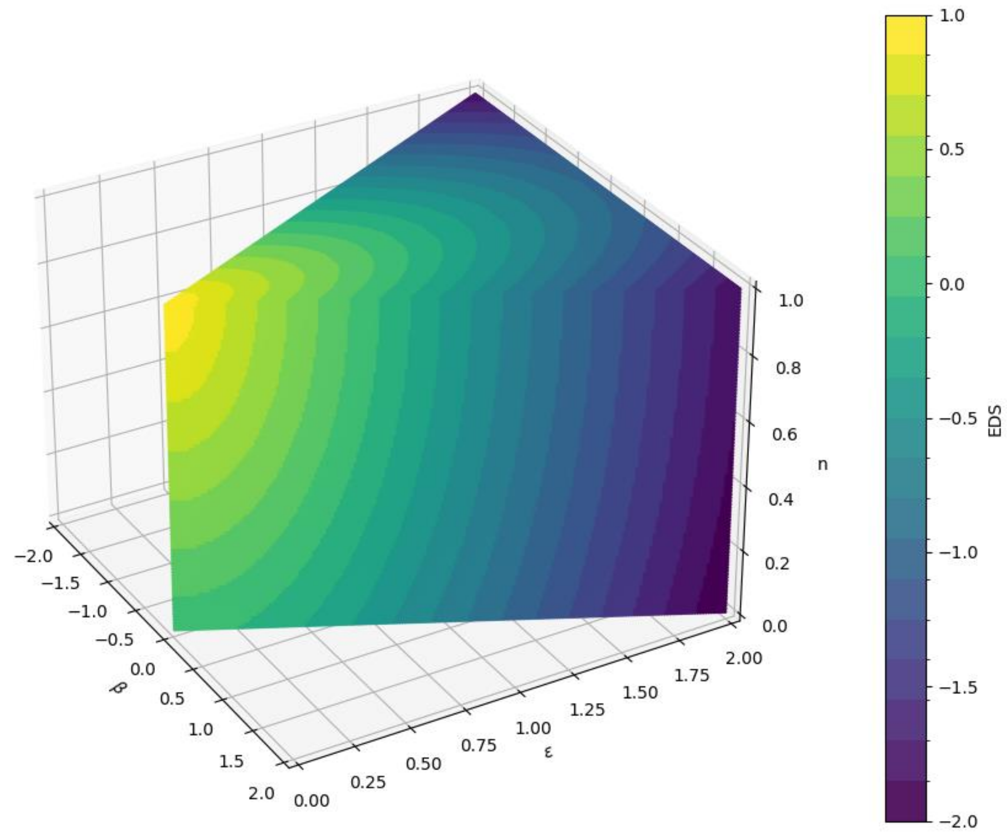

Figure 1. Geometric representation of the Euclidean Distance Score (EDS) in the three-dimensional (β, ϵ, n) space. The ideal retrieval corresponds to $(\beta, \epsilon, n) = (0,0,1)$. The shown domain is restricted to metric combinations satisfying $|\beta| \leq \epsilon$, consistent with their definition. For visualization purposes, EDS values are displayed over the range $[-2,1]$.