Authors Reply to RC4: 'Comment on egusphere-2025-4343', Anonymous Referee #4, 09 Dec 2025

*Reviewer comments appear in **bold**. Authors' responses are in plain text.

Using a single metric to assess the performance of bio-optical algorithms is an interesting but challenging topic. The Euclidean Distance Score (EDS) proposed by the authors could be a useful approach. My major concerns, similar to those raised by other reviewers, involve the choice of individual metrics, their robustness, existing correlations, and inconsistent ranges (i.e., lack of normalization). These issues are critical and should be properly addressed.

From a potential user's perspective, the selection of metrics is somewhat arbitrary, and the results can therefore be misleading. Although the mathematics behind the Euclidean distance is straightforward, the method implicitly assumes that all metrics are equally important, which is not always the case and may vary across application scenarios. For example, in water-quality monitoring, absolute error may be the most important metric, while in time-series studies, bias, which indicates systematic over- or underestimation, may be more relevant. A combined score may hide such differences. This raises the question of how sensitive the EDS is to each input metric. The authors may consider performing a sensitivity analysis to examine whether all selected metrics contribute equally or whether some dominate the score.

Overall, the paper is well written. I was pleased to read it. However, the issues above need to be properly addressed for the paper to contribute to the field.

We thank the reviewer for this thoughtful assessment and for raising concerns that are indeed central to the design of a composite performance metric. These considerations, together with related comments from other reviewers, directly motivated revisions and further development of the EDS framework.

Regarding the perceived arbitrariness in metric selection, the revised manuscript will include a dedicated section detailing the selection process. This includes a characterization of candidate metrics (summarized in Table 3) and a redundancy analysis (more details on Reply to Reviewer 3, D) to ensure that only complementary metrics are retained. Based on this analysis, the EDS was restricted to three dimensionless metrics representing distinct performance aspects: error magnitude (median symmetric accuracy, ϵ), systematic bias (symmetric signed percentage bias, β), and retrieval robustness (valid retrieval ratio, n). Metrics found to be redundant, range-dependent, or sensitive to regression assumptions (e.g., slope and correlation) were removed.

Figure 1 illustrates a geometric representation of the revised EDS in a three-dimensional space, following the exclusion of regression slope and Pearson correlation coefficient from the score . Of the three remaining metrics, the valid retrieval ratio (n) is naturally bounded between 0 and 1 and referenced to an ideal value of unity. Median Symmetric Accuracy (ϵ) and Symmetric Signed Percentage Bias (β) are dimensionless, defined relative to an ideal value of zero, with magnitudes that directly reflect fractional deviations from perfect agreement. Although they are formally unbounded and may exceed unity, we choose to not normalize them so that extreme deviations are strongly penalized rather than compressed through imposed bounds. For retrievals with errors and biases below 100% (corresponding to reasonably performing retrievals in practice), all metrics are of order unity and therefore can contribute comparably to the distance. This is reflected in the approximately isotropic geometry of the high-EDS region, indicating that no single metric is implicitly favoured in that regime. In contrast, values of ϵ or $|\beta|$ exceeding unity correspond to strongly degraded retrievals and dominate the distance, displacing the solution away from the ideal point in the EDS space.

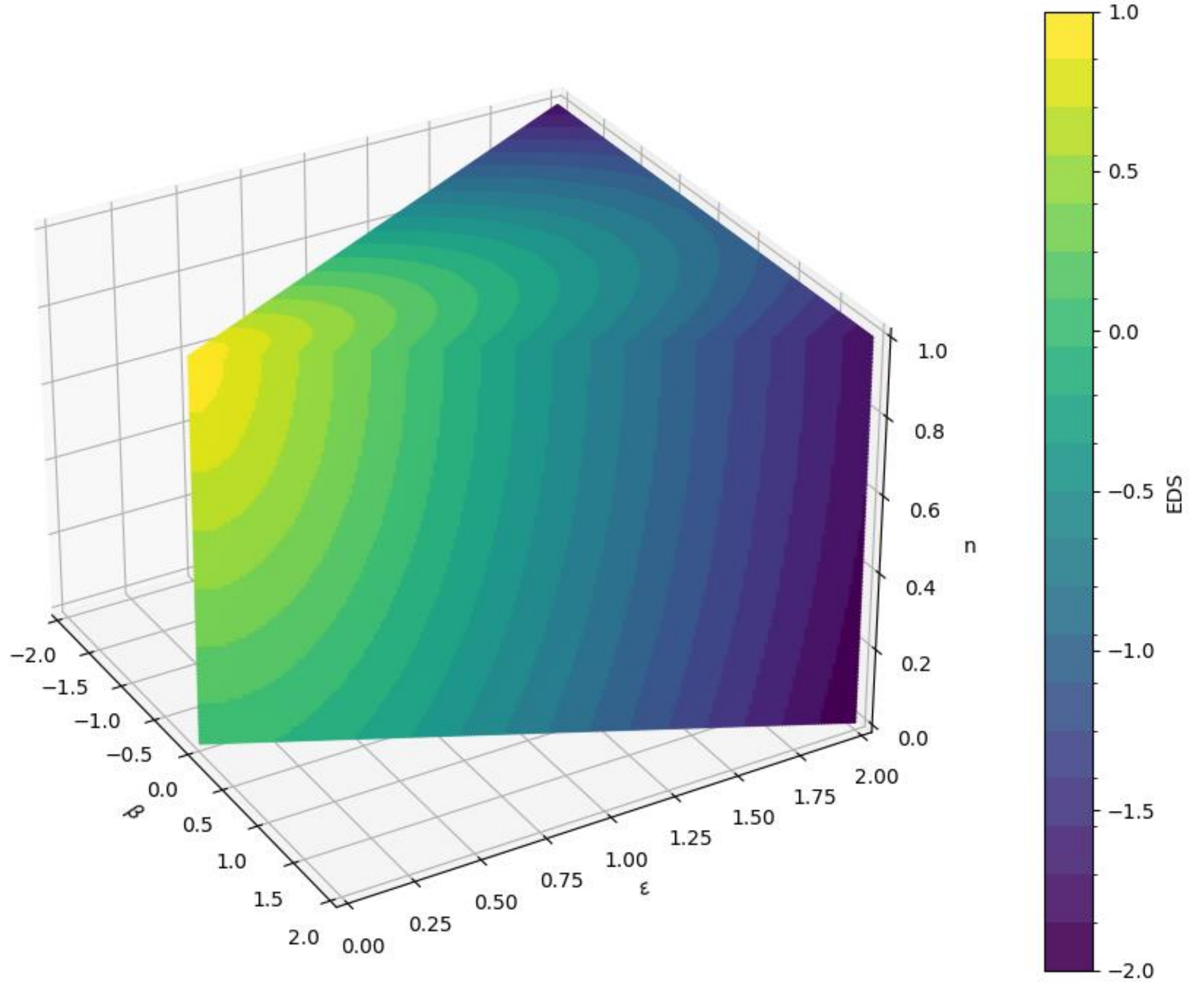

Figure 1. Geometric representation of the Euclidean Distance Score (EDS) in the three-dimensional (β, ϵ, n) *space. The ideal retrieval corresponds to* $(\beta, \epsilon, n) = (0,0,1)$. *The shown domain is restricted to metric combinations satisfying* $|\beta| \leq \epsilon$, *consistent with their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

Nevertheless, having comparable numerical scales does not imply that all metrics exert equal influence on the EDS across its admissible domain. The components exhibit different empirical variances: the agreement-based terms (ϵ and β) may span a wide range depending on retrieval quality, whereas the valid retrieval ratio (n) is bounded and, in most realistic applications, concentrated near its ideal value of unity. To quantify how these differences translate into effective influence on the score, we performed a pointwise sensitivity analysis based on the analytical gradient of the Euclidean distance, identifying the locally dominant direction of score variation at each admissible point.

The resulting dominance structure, illustrated in Figure 2, shows that variations in error magnitude (ϵ) control the sensitivity of the score over most of the admissible space. Systematic bias (β) does not emerge as a dominant sensitivity on its own, but attains equal influence with ϵ along a narrow, well-defined surface where $| \beta |= \epsilon$ and both exceed $n - 1$. Sensitivity to the valid retrieval ratio (n) is comparatively smaller over large portions of the space, but becomes dominant where ϵ and β are low.

When averaged over the explored domain (restricted to EDS > −2), the mean relative sensitivities are 1.106 for ϵ, 0.338 for β, and 0.116 for n. These values describe the average local responsiveness of the EDS to perturbations in each component across the admissible space. Stratifying the analysis by EDS (Table 1) reveals a transition in sensitivity regimes: near-optimal retrievals (EDS > 0.75) are most sensitive to n, whereas increasingly degraded retrievals exhibit progressively stronger sensitivity to ϵ. The comparatively lower sensitivity associated with β does not imply a negligible contribution of bias to the distance. Rather, it reflects the constraint $| \beta |\leq \epsilon$, which limits the independent variability of bias and confines its influence on the distance to specific regions of the space where systematic over- or underestimation occurs.

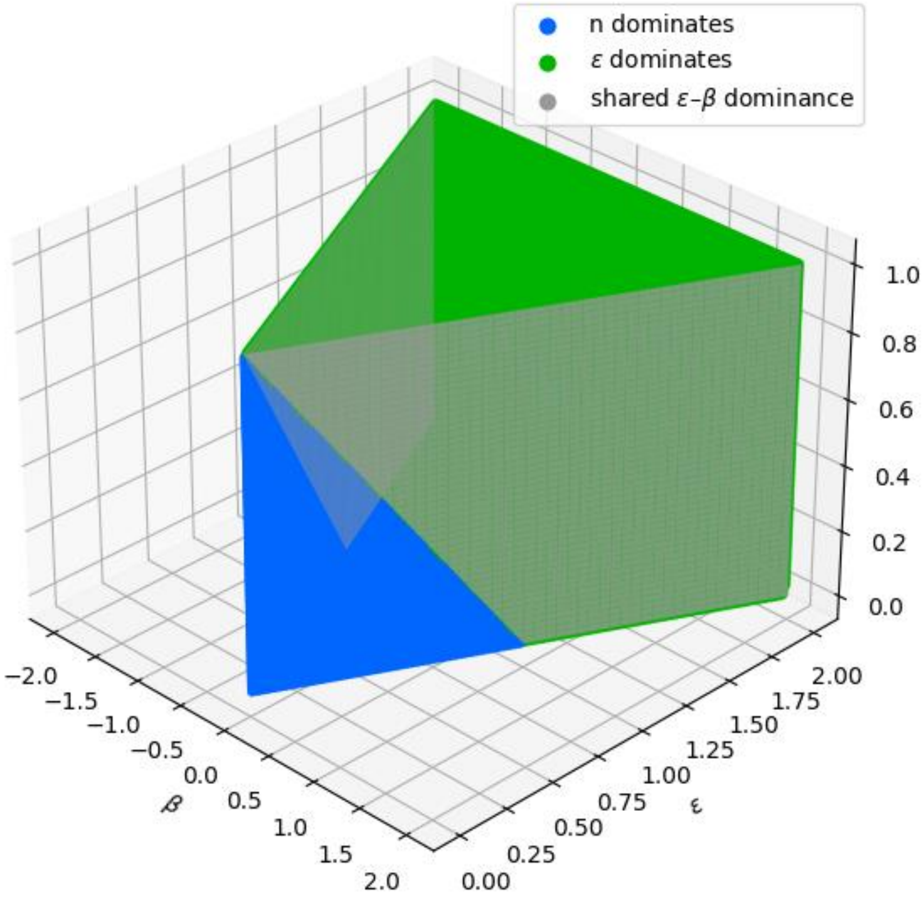

Figure 2. Sensitivity dominance structure of the Euclidean Distance Score (EDS) in the three-dimensional (β, ϵ, n) space under the constraint $| \beta | \leq \epsilon$. Colored regions indicate the metric to which the EDS is locally most sensitive, based on the analytical gradient of the distance

Table 1. Mean relative sensitivities of the Euclidean Distance Score (EDS) with respect to error magnitude (ϵ), systematic bias (β), and valid retrieval ratio (n), computed over non-overlapping EDS bins. Sensitivities quantify the average local response of the score to perturbations in each component.

EDS range	⟨Sn⟩	⟨Sβ⟩	⟨Sϵ⟩	Dominant sensitivity
EDS > 0.75	0.425	0.023	0.103	n
0.50 < EDS ≤ 0.75	0.362	0.049	0.220	n
0.25 < EDS ≤ 0.50	0.284	0.078	0.351	ϵ
0.00 < EDS ≤ 0.25	0.203	0.108	0.485	ϵ
EDS ≤ 0.00	0.098	0.373	1.206	ϵ

While the preceding analyses examine the theoretical geometry and sensitivity structure of the EDS, it is also instructive to assess how the score behaves in practical retrieval scenarios. We therefore conducted a structured perturbation analysis across 25 retrieval instances (two algorithms applied to four datasets and multiple variables). ϵ, β and n were independently perturbed by ±5%, ±10%, ±20%, and ±30%, while keeping the remaining components unchanged. The results show that EDS responded smoothly to increasing perturbation magnitude, with the largest sensitivity associated with ϵ, followed by β, and smaller effects for n. For example, a ±20% perturbation yields median absolute EDS changes of approximately 0.054 for ϵ, 0.016 for β, and 0.014 for n. The comparatively smaller influence of n reflects the fact that, for most practical retrieval instances, n is close to its ideal value and contributes little to the total distance. This analysis demonstrates that, in real applications, the sensitivity of the EDS strongly depends on the relative contribution of each component for a given retrieval. For context, the relative contribution of the metrics for the same 25 retrieval instances are shown in **Error! Reference source not found.**. The score is primarily driven by ϵ, systematic bias (β) generally plays a secondary role, and contribution of the valid retrieval ratio (n) is small for most retrievals.

Table 2. Sensitivity of the Euclidean Distance Score (EDS) to metric-level perturbations. Minimum, median, mean, and maximum absolute changes in EDS (| ΔEDS |) resulting from ±5%, ±10%, ±20%, and ±30% perturbations

Metric	Perturbation	Min | ΔEDS |	Median | ΔEDS |	Mean | ΔEDS |	Max | ΔEDS |
$\boldsymbol{\beta}$	5%	3.38E-06	3.89E-03	8.05E-03	6.10E-02
$\boldsymbol{\beta}$	10%	6.58E-06	7.78E-03	1.61E-02	1.23E-01
$\boldsymbol{\beta}$	20%	1.25E-05	1.55E-02	3.21E-02	2.52E-01
$\boldsymbol{\beta}$	30%	1.77E-05	2.29E-02	4.78E-02	3.85E-01
$\boldsymbol{\epsilon}$	5%	1.18E-03	1.38E-02	1.83E-02	6.10E-02
$\boldsymbol{\epsilon}$	10%	2.34E-03	2.75E-02	3.66E-02	1.23E-01
$\boldsymbol{\epsilon}$	20%	4.59E-03	5.38E-02	7.31E-02	2.52E-01
$\boldsymbol{\epsilon}$	30%	6.73E-03	7.87E-02	1.09E-01	3.85E-01
$\boldsymbol{n}$	5%	0.00E+00	2.56E-03	4.36E-03	2.65E-02
$\boldsymbol{n}$	10%	0.00E+00	8.07E-03	1.18E-02	7.16E-02
$\boldsymbol{n}$	20%	0.00E+00	1.44E-02	3.33E-02	1.69E-01
$\boldsymbol{n}$	30%	0.00E+00	2.28E-02	6.10E-02	2.68E-01

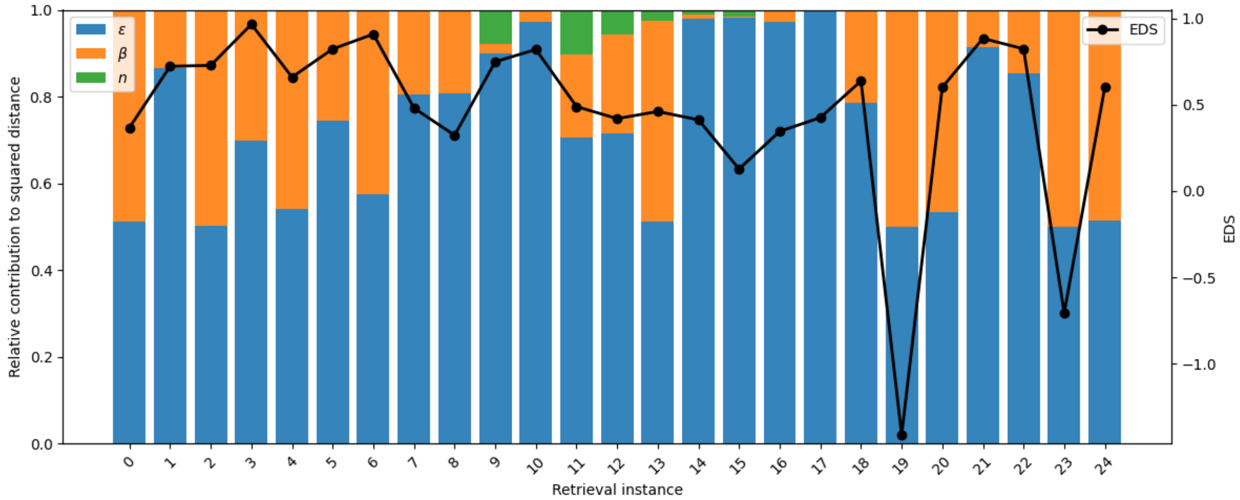

Figure 3. Fractional contribution of the EDS components to the squared distance for each retrieval instance (stacked bars), with the corresponding EDS shown on the secondary axis. Contributions are shown for ϵ, β, and n.

We also acknowledge that different applications may prioritize different performance aspects (e.g., error magnitude versus bias), thank you for raising this point. The EDS was conceptualized for typical algorithm validation in aquatic remote sensing. We do however highlight that the EDS formulation could be adapted to give more weight to different aspects according to the needs of the user. In the revised manuscript we will highlight this more explicitly.

Table 3. Summary of candidate performance metrics considered for the evaluation of bio-optical retrieval algorithms. For each metric, the table reports its mathematical definition, metric class, key characteristics, and main limitations when applied to bio-optical variables. In all definitions, E_i and O_i denote the estimated and observed values of the i-th retrieval, respectively, and $Q_i = E_i/O_i$ is the corresponding accuracy ratio.

Metric	Definition	Metric Class	Key characteristics	Limitations for bio-optical variables
MAPE	$\frac{100}{n}\sum_{i=1}^{n}\left\lvert\frac{E_i - O_i}{O_i}\right\rvert$	Agreement (multiplicative, deviation-based)	- Average deviation in percentage format - Range-independent interpretability	- Unstable for small observed values - Asymmetric (overestimation penalized more than underestimation)
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(E_i - O_i)^2}$	Agreement (additive, deviation-based)	- Deviation in original units - Quadratic penalty emphasizes large errors	- Highly sensitive to outliers and heteroscedasticity - Interpretability affected by the range of the data
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert E_i - O_i\rvert$	Agreement (additive, deviation-based)	- Average deviation in original units - Less sensitive to outliers than RMSE	- Sensitive to heteroscedasticity - Interpretability affected by the range of the data
Bias	$\frac{1}{n}\sum_{i=1}^{n}(E_i - O_i)$	Agreement (additive, signed deviation-based)	- Average signed deviation in original units - Measure of systematic over- /under estimation	- Sensitive to heteroscedasticity - Interpretability affected by the range of the data
MAE-ratio	$10^{\frac{1}{n}\sum_{i=1}^{n}\lvert\log_{10}(Q_i)\rvert}$	Agreement (multiplicative, ratio-based)	- Average deviation in ratio-based format - Range-independent interpretability - Suitable for log-normally distributed variables	- Mean aggregation remains sensitive to outliers and heteroscedasticity that could remain even in log space
Bias-ratio	$10^{\frac{1}{n}\sum_{i=1}^{n}\log_{10}(Q_i)}$	Agreement (multiplicative, signed ratio-based)	- Average signed deviation in ratio form - Measure of systematic over- /under estimation - Range-independent interpretability	- Mean aggregation remains sensitive to outliers and heteroscedasticity that could remain even in log space

Metric	Definition	Metric Class	Key characteristics	Limitations for bio-optical variables
			- Suitable for log-normally distributed variables	
RMSE-ratio	$10^{\frac{1}{n}\sum_{i=1}^{n}(\log_{10}(Q_i))^2}$	Agreement (multiplicative, ratio-based)	- Deviation in a ratio format - Quadratic penalty emphasizes large errors - Range-independent interpretability - Suitable for log-normally distributed variables	- Mean aggregation remains sensitive to outliers and residual heteroscedasticity that could remain even in log space
Median Symmetric Accuracy (ϵ)	$10^{median(\lvert\log_{10}(Q_i)\rvert)} - 1$	Agreement (multiplicative, deviation-based)	- Median proportional deviation - Range-independent interpretability - Suitable for log-normally distributed variables - Median aggregation yields robustness to outliers and residual heteroscedasticity	- Less familiar metric
Symmetric Signed Percentage Bias (β)	$\mathrm{sign}(M)\left(10^{\lvert M\rvert} - 1\right)$ where $M = \mathrm{median}(\log_{10}(Q_i))$	Agreement (multiplicative, signed deviation-based)	- Median signed proportional deviation (systematic bias) - Range-independent interpretability - Suitable for log-normally distributed variables - Median aggregation yields robustness to outliers and residual heteroscedasticity	- Less familiar metric
Pearson correlation coefficient (r)	$r = \frac{cov(E,O)}{\sigma_E \sigma_O}$	Association	- Strength of linear association (typically in log space)	- Does not quantify agreement - Lacks response to bias

Metric	Definition	Metric Class	Key characteristics	Limitations for bio-optical variables
				- Sensitive to outliers, leverage points and data range
Regression slope	Regression method dependent	Association	- Describes how variability in the observations is scaled by the estimates (typically in log space)	- Does not quantify agreement - Sensitive to outliers and leverage points - Poorly conditioned for narrow data ranges
Intercept	Regression method dependent	Association	- Describes the offset between estimated and observed values at the origin (typically in log space)	- Limited physical interpretability
Valid retrieval ratio (n)	$\frac{N_{\mathrm{E,valid}}}{N_{\mathrm{O}}}$	Algorithm robustness	- Fraction of cases for which a valid retrieval is produced - Captures algorithm convergence and ability to produce physically plausible outputs	- Depends on the definition of the valid retrieval range, which can be subjective