Authors Reply to RC2: 'Comment on egusphere-2025-4343', Anonymous Referee #2, 19 Nov 2025

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

The manuscript addresses an issue of great importance, which is the assessment and comparison of remote sensing algorithms, given the non-normal distributions of most bio-optical variables. Generally, the text is brief, and the points made by the authors are clear and relevant. I have recognized the need for a robust assessment method for some time, and I see the advantages of the proposed one. However, I think the manuscript is too brief at times, and some sections may benefit from more in-depth explanation.

> We thank you for the encouraging and constructive review of our paper. Your suggestions are much appreciated. We address each in detail below.

Firstly, regarding the assumptions made, the reduced major axis regression is mentioned several times, but I find that additional information is needed to clarify the significance of the problem. To illustrate my point, I found an interesting publication by Bilal et al. (2022) in the Encyclopedia of Mathematical Geosciences (https://doi.org/10.1007/978-3-030-26050-7_270-1). This work discusses the presence of errors in both the dependent and independent variables in geosciences, which is exactly what I find missing in this text to highlight the value of this study.

> We thank the reviewer for this suggestion and for providing a helpful reference. We agree that this context was not sufficiently articulated in the original manuscript.
>
> In the revised manuscript, we will add a dedicated section discussing several commonly used metrics in the field, with their characteristics and limitations, including regression-based diagnostics:
>
> "*Regression parameters are commonly reported to assess the proportionality and offset between estimated and observed values. The slope is more widely reported to evaluate performance, as it describes how variability in the observations is scaled by the estimates, indicating whether the dynamic range is compressed or expanded. The intercept represents an additive offset at the reference origin*.
>
> *Several regression approaches can be used to estimate these parameters. Ordinary least squares (OLS, type-1 regression) minimizes vertical residuals and implicitly assumes that the observed values are error-free. To account for uncertainty in both estimated and observed quantities, type-2 regressions such as reduced major axis (RMA) are frequently employed in aquatic remote-sensing studies. RMA minimizes the perpendicular distance to the regression line and is symmetric, such that the slope of* E *(estimated) regressed on 0 (observed) is the inverse of 0 regressed on* E*. Nevertheless, both type-1 and type-2 regressions are least-squares methods and therefore remain sensitive to the statistical distribution of the data, outliers and leverage points. More robust alternatives, including Theil–Sen (asymmetric) and Passing–Bablok (symmetric) regression methods, reduce sensitivity to outliers by relying on median-based estimators.*

Nevertheless, regardless of the method employed, regression parameters are sensitive to the range of the evaluated data. 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 perturbations in the data can lead to large variations in the estimated slope. As a result, regression parameters are not always reliable and directly comparable across datasets or stratifications."

We note, however, that considering the mentioned limitations and identified redundancy with bias metrics (see Reply to Reviewer 3 for more details), regression slope (and Pearson correlation coefficient, more details on the following answer) is no longer retained a component of the composite EDS. This revision removes reliance on regression assumptions in the score formulation itself, but we suggest that regression diagnostics are still reported, where relevant, as descriptive information.

Furthermore, I am curious as to why the Pearson correlation coefficient was selected instead of the Mann-Kendall test, which does not have such strict assumptions, particularly when not all variables have ideal log-normal distributions and log-transformation does not always ensure normality.

We thank the reviewer for this question and for highlighting the limitations of Pearson correlation in the presence of non-normal and heteroscedastic data.

In the revised manuscript, we will add a dedicated section discussing several commonly used metrics in the field, including the characteristics and limitations of Pearson r:

"The Pearson correlation coefficient (r) and the coefficient of determination (r^2) are widely used metrics to represent the goodness of fit between estimated and observed values. While these metrics are useful for characterizing linear association, they do not quantify agreement or accuracy per se (Bland & Altman, 1986). In particular, they are insensitive to systematic bias: an algorithm may consistently over- or under-estimate observations while still yielding a high correlation coefficient if the relative ordering of values is preserved. Furthermore, both r and r^2 are sensitive to outliers and strongly dependent on the dynamic range of the data. When values span a wide range, high correlation coefficients may be obtained even in the presence of substantial absolute or relative errors, whereas restricted ranges can yield low correlation despite good agreement."

Considering these limitations (some of which also apply to other rank-based association measures, such as the Mann–Kendall statistic), and in light of the redundancy identified between correlation and error-magnitude metrics, the correlation coefficient is no longer retained as a component of the EDS in the revised formulation.

Secondly, a paper of this nature, aiming to establish a certain assessment standard, should provide a broader explanation of the somewhat arbitrary nature of the logarithm selection mentioned in line 106. I remember being quite confused about this when I was a beginning researcher, and I believe that a methods paper should explain it more thoroughly. Similarly, the

definition of the number of valid retrievals in Equation 6 seems rather vague. I would expect a more specific definition of what "valid" means here and how it may affect the results.

We thank the reviewer for this comment and agree that these points require clearer explanation in a methods-oriented paper.

Regarding the logarithm selection, we will expand the revised manuscript to explicitly clarify that the choice of logarithm base is mathematically arbitrary:

"*It should be noted that although Morley et al. (2018) formulated these metrics using the natural logarithm, the authors note that the choice of logarithm base is arbitrary, provided that the corresponding antilogarithm is applied consistently. In practice, using different logarithm bases does not change the meaning of the metric, as long as the same base is used throughout the calculation.*"

We will also expand the definition of the valid retrieval ratio to more clearly specify what constitutes a "valid" retrieval:

"*The valid retrieval ratio (n) is a measure of retrieval robustness. It is defined as the fraction of cases for which an algorithm produces a valid estimate relative to the number of available reference observations. A valid retrieval is defined as one for which the algorithm converges and returns physically plausible parameter values within predefined bounds. These bounds are established a priori based on instrument uncertainty, known physical limits, and extreme values reported in the literature. Retrievals falling outside these bounds are considered non-valid and excluded from the valid retrieval count*.

The number of valid retrievals is commonly reported in the aquatic remote-sensing literature, but is often treated as a descriptive statistic rather than being explicitly incorporated into the quantitative evaluation or ranking of algorithm performance. Recognizing the importance of retrieval feasibility, particularly for inversion-based algorithms that may fail to converge or produce non-physical solutions, some studies have started explicitly integrated measures of retrieval success into their scoring or ranking frameworks (Brewin et al., 2015; Müller et al., 2015; Seegers et al., 2018)."

Lastly, I appreciate presenting real-life examples. However, I believe that adding a few more commonly used metrics, such as the root mean squared error, and discussing their limitations could help illustrate why the proposed approach is more robust.

We thank the reviewer for this suggestion and agree that including commonly used metrics can help contextualize the proposed approach. In the revised manuscript, we will expand the example section to report additional traditional metrics alongside the EDS results and to explicitly discuss their behaviour, limitations and how they compare.

As mentioned in the previous answers, we have developed a dedicated section discussing commonly used metrics in the field, with their characteristics and limitations (a summary table is shown in the Reply to Reviewer 4). This will provide a good basis for discussion when we present the examples.

To summarize, I find this work to be much needed and valuable, and it is already well-written. However, to convince sceptics and encourage broader application, I recommend providing additional explanations for those entering the field who may not understand the jargon or have not yet grasped all the challenges related to assessing optical algorithms.

We thank the reviewer for this encouraging assessment. We believe the suggested additions will help broaden the applicability and understanding of the proposed approach.