Authors Reply to RC3: 'Comment on egusphere-2025-4343', Anonymous Referee #3, 21 Nov 2025

*To improve readability the responses below address the reviewer's points directly; the full reviewer comments are not repeated.

We thank the reviewer for this detailed critique. Many of the concerns raised, particularly those related to metric redundancy, dominance of individual components, and interpretation of example cases, highlight limitations of our initial formulation and manuscript. In response to these comments (and related feedback from other reviewers), we have substantially revised the EDS framework to strengthen its conceptual and statistical basis. We also acknowledge that some statements in the original manuscript were imprecise or insufficiently supported by quantitative analysis. These will be revised and clarified accordingly in the updated manuscript.

Below, we provide detailed responses to the reviewer's comments. Related concerns have been grouped and addressed together for clarity, as outlined below.

Author's Reply	**Reviewer comments**
A. Normalization, Sensitivity and relative contributions	1 (partially), 4, 7, 8
B. Misinterpretation of Reduced Major Axis (RMA) Regression	2
C. Ideal point	3
D. Redundancy	1 (partially), 5
E. Behaviour in the Example Cases	6
F – Treatment of Variables	9

A. Normalization, Sensitivity and Relative Contributions

The reviewer is correct in noting that a composite, distance-based score requires careful consideration of the relative scaling, influence and sensitivity of its components. We also agree that our earlier statement that the components "typically weigh equally" was imprecise and will be revised in the manuscript. In response to these concerns, we revised the EDS formulation and explicitly examined how the remaining metrics contribute to the distance calculation both in theory and in practical cases.

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 (see Reply D for details). 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, this property is retained 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 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.

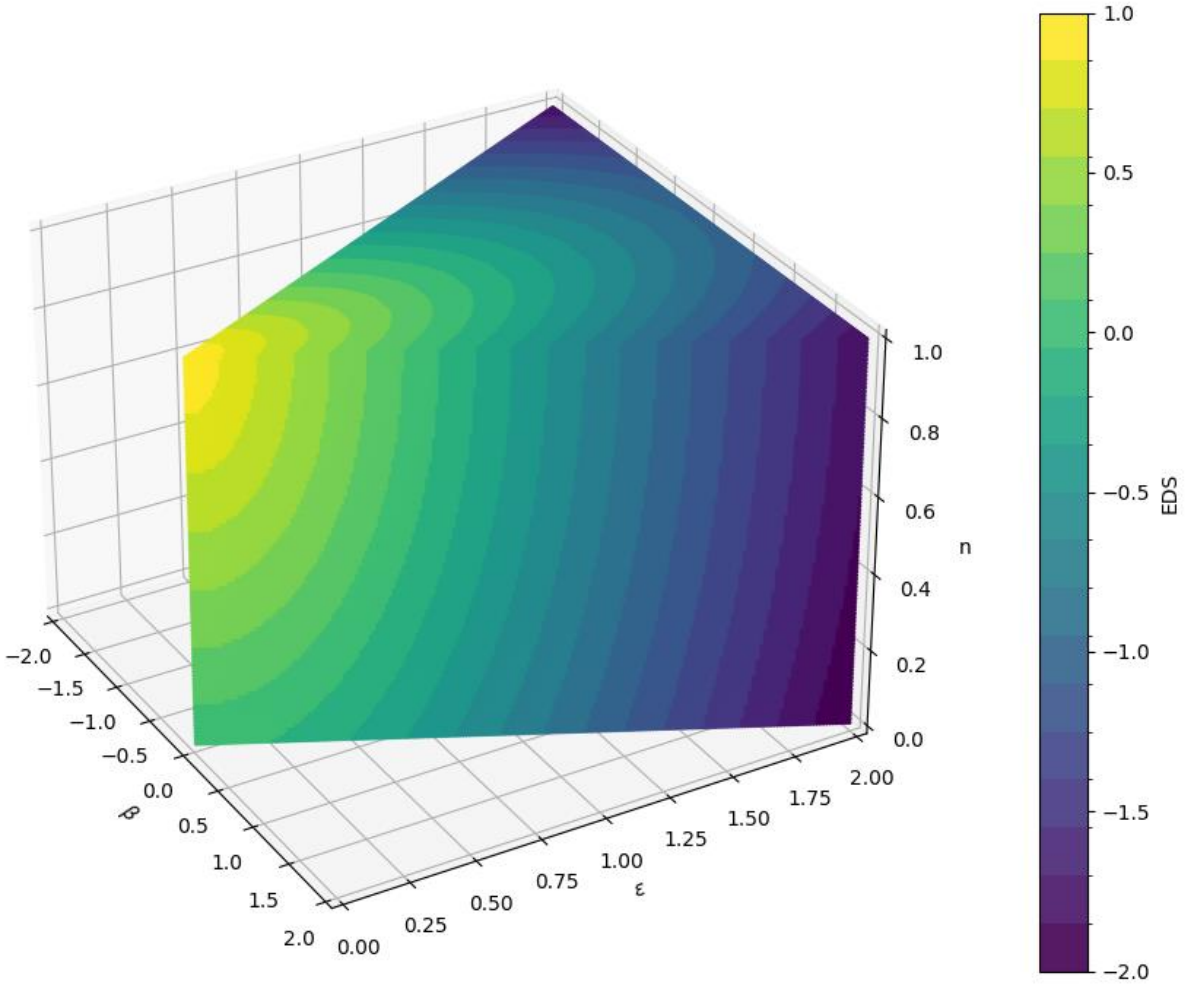

Figure 1. Geometric representation of the Euclidean Distance Score (EDS) in the three-dimensional (β, ϵ, n) space. The ideal retrieval corresponds to $(\beta, \epsilon, n) = (0,0,1)$. The shown domain is restricted to metric combinations satisfying $|\beta| \leq \epsilon$, consistent with their definition. 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. As mentioned by the reviewer, 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.

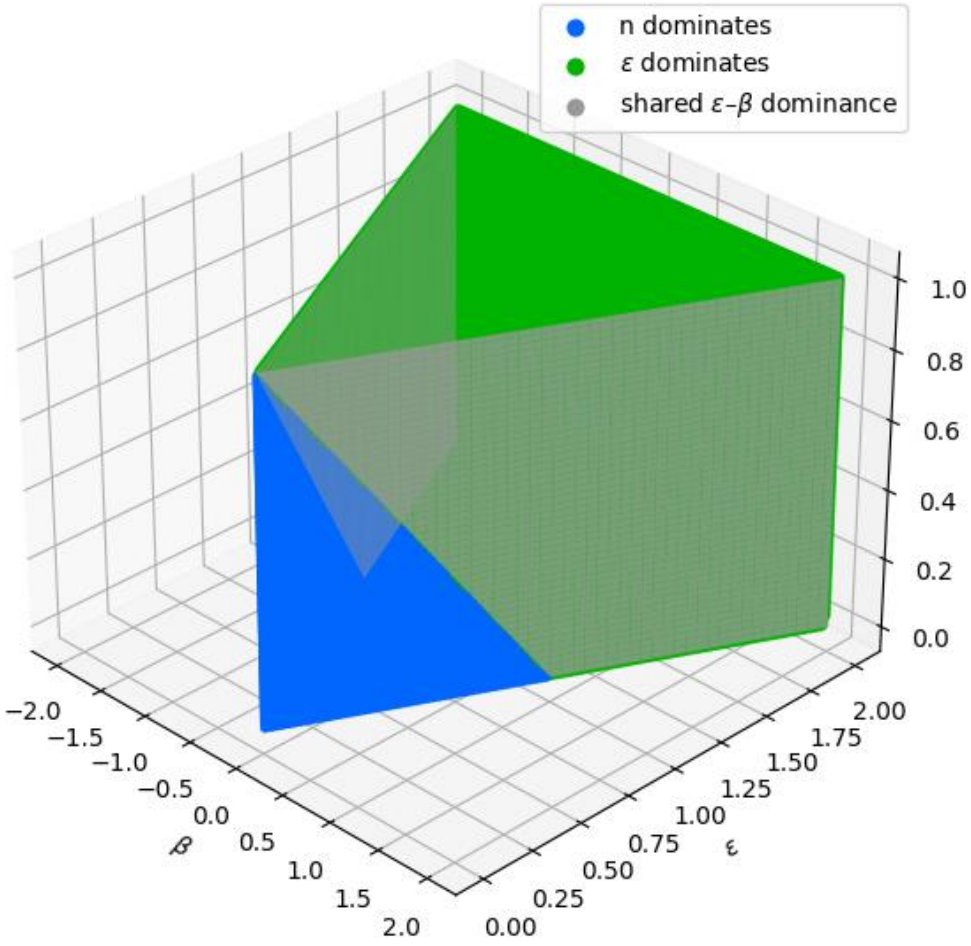

Figure 2. Sensitivity dominance structure of the Euclidean Distance Score (EDS) in the three-dimensional (β, ϵ, n) space under the constraint $|\beta| \leq \epsilon$. Colored regions indicate the metric to which the EDS is locally most sensitive, based on the analytical gradient of the distance

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 Figure 3. 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 ($|\ \Delta EDS\ |$) 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

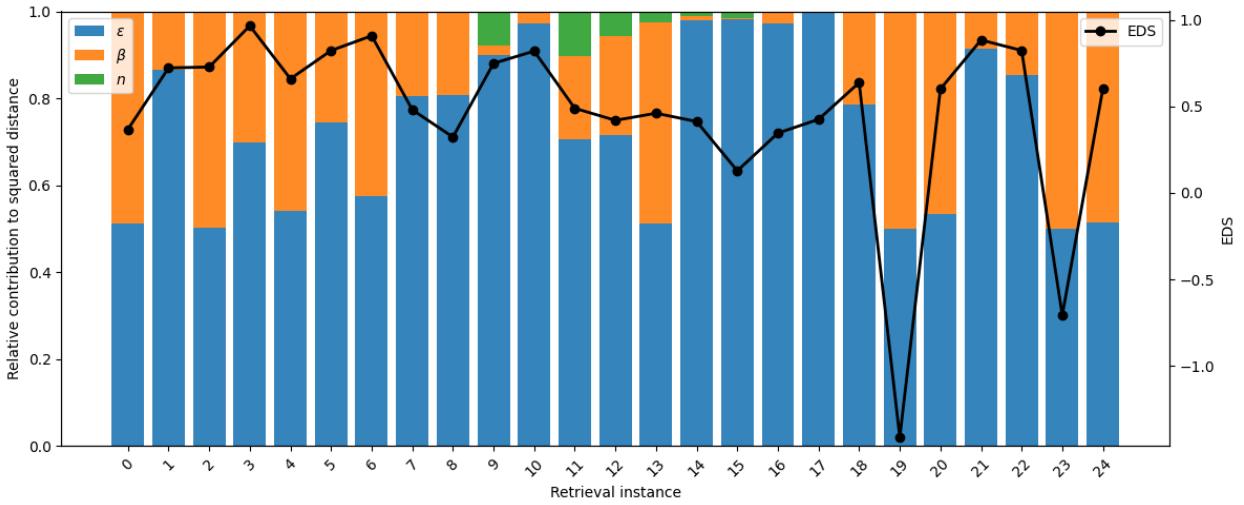

Figure 3. Fractional contribution of the EDS components to the squared distance for each retrieval instance (stacked bars), with the corresponding EDS shown on the secondary axis. Contributions are shown for ϵ, β, and η.

B. Misinterpretation of Reduced Major Axis (RMA) Regression

We thank the reviewer for this comment. In the revised EDS framework, regression slope and the Pearson correlation coefficient are no longer included in the score due to identified issues such as the statistical dependency between metrics (see Reply D for more details on redundancy). As a result, the revised EDS no longer treats slope and correlation as independent dimensions, thereby addressing the concern raised by the reviewer.

C. Ideal Point

We thank the reviewer for this comment. The ideal point in the EDS is not intended to represent a physically attainable retrieval result, but rather a reference used to define a distance-based measure of relative performance. In the revised formulation, the ideal point is defined solely in terms of zero error, zero bias, and full retrieval success, which serve as consistent reference values for quantifying deviations among algorithms, not as targets expected to be reached in practice.

D. Redundancy

We thank the reviewer for raising the concern about metric redundancy. We agree that jointly retaining metrics that capture the same performance aspect can lead to redundancy and violate the assumptions underlying Euclidean aggregation. This concern directly motivated the revision of the EDS metric set.

To explicitly assess redundancy among candidate metrics, we analysed pairwise inter-metric relationships considering 25 retrieval instances (two algorithms applied to four datasets and different variables, including $a_{ph}(\lambda)$, $b_{bp}(\lambda)$, , $a_{dg}(\lambda)$, Chla, SPM, kd(λ) and Secchi depth). Inter-metric relationships were quantified using Spearman's rank correlation coefficient (Figure 4), which is appropriate for assessing monotonic associations in small samples. Pairwise scatterplots of the raw metric values provide a visual check of the corresponding relationships.

This analysis indicates that several candidate metrics exhibit substantial dependence and should not be jointly retained in a distance-based score. In particular, regression slope and the Pearson correlation coefficient (r) show systematic associations with error magnitude- and bias-based metrics respectively and were therefore considered redundant. A complementary rank-based variance inflation factor (VIF) analysis further supports this conclusion: while the original formulation exhibited elevated VIF values (up to ~2.7 for ε and ~2.6 for r), indicating shared variance among metrics, all retained components in the revised EDS exhibit consistently low VIF values (≤ ~1.8), confirming minimal multicollinearity.

Based on these results, regression slope and correlation were removed from the EDS, and only one representative metric of error magnitude (ε), one of systematic bias (β), and one of retrieval robustness (n) were retained. The revised EDS therefore avoids double-counting of performance aspects and satisfies the requirement that each dimension captures complementary information, addressing the redundancy concern raised by the reviewer. Detailed correlation and redundancy analyses will be provided in a revised manuscript.

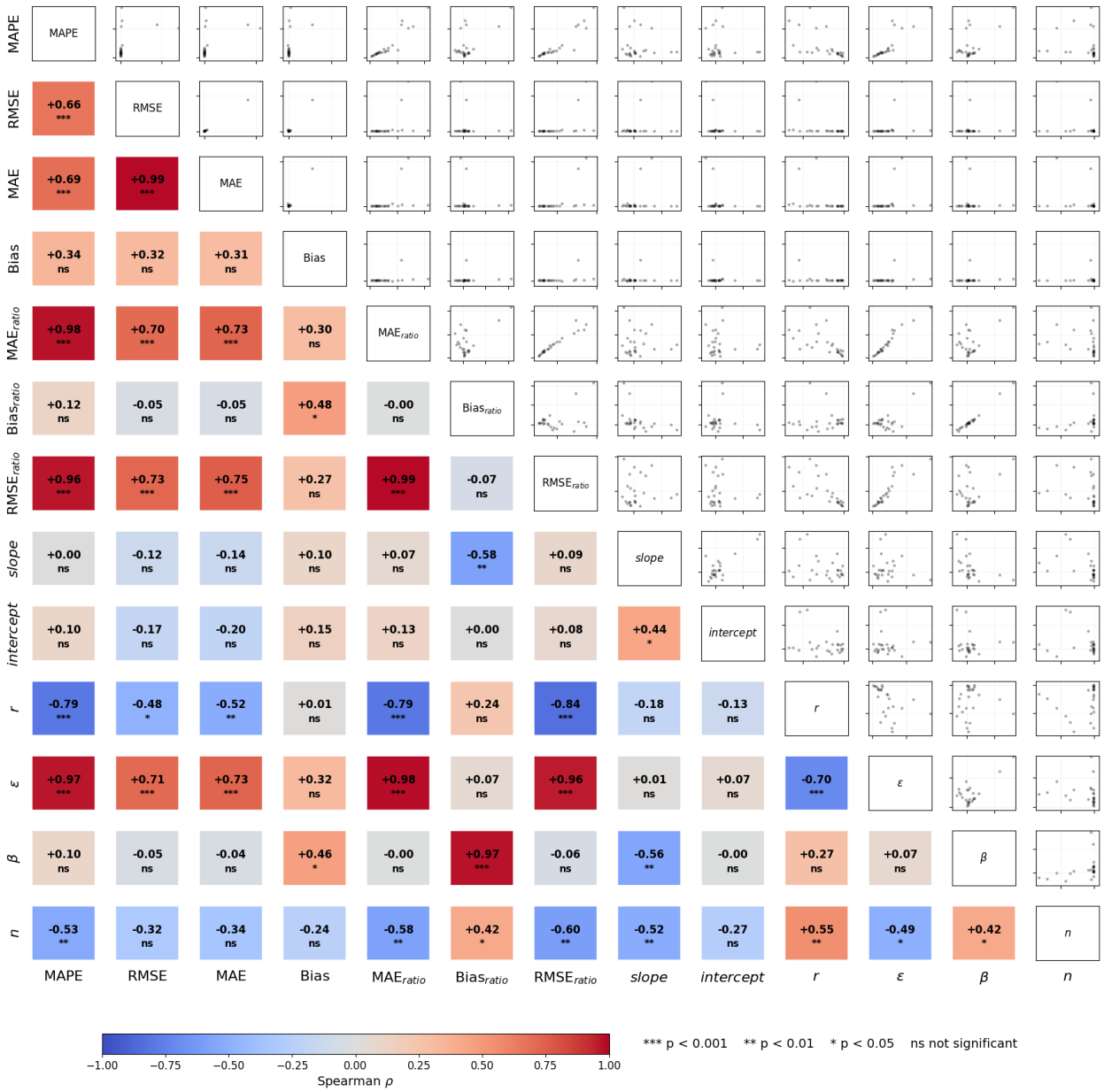

Figure 4. Pairwise Spearman rank correlation coefficients (ρ) between candidate metrics evaluated across 25 model–dataset–variable instances. The lower triangular matrix shows correlation coefficients with statistical significance indicated by asterisks, while the upper triangle displays scatterplots of the corresponding metric pairs for illustrative purposes.

E. Behaviour in the Example Cases

We thank the reviewer for this comment. We agree that the examples highlighted in the manuscript revealed limitations of the original EDS formulation, particularly the disproportionate influence of regression-based metrics on the final score.

In the b_{bp} retrieval case, the originally low EDS value was indeed driven primarily by a very high regression slope (which had high uncertainty in its estimation due to small range), despite low bias ($\beta \approx -4\%$) and acceptable median symmetric accuracy ($\epsilon \approx 24\%$). In the revised formulation, regression slope and correlation are no longer included in the distance calculation. As a result, the revised EDS for this case increases to 0.75, which is consistent with the overall assessment indicated by the error, bias, and robustness metrics.

In the oligotrophic k_d example, the original formulation yielded a moderate EDS despite low error and bias, due to the inclusion of correlation-based diagnostics. With the revised EDS, the score increases to 0.86. This value is consistent with the corresponding EDS obtained when considering all trophic states (0.82) and when analysing trophic regimes separately (mesotrophic: 0.84; eutrophic: 0.78), indicating coherent behaviour with stratifications and the change of dynamic range. Because agreement-based metrics remain comparable across these stratifications, consistency among the performance is expected. Such consistency was not observed in the original formulation, where association-based metrics introduced sensitivity to changes in data range.

In the revised manuscript, these updated EDS values will replace the original scores in the example cases, and we will additionally discuss the behaviour of commonly used metrics, including regression slope, highlighting situations in which they may yield misleading conclusions.

F. Treatment of Variables

We thank the reviewer for this comment. The EDS aims to provide a unified, dimensionless framework for summarizing retrieval performance. A key design feature of the revised EDS is that all retained components are expressed in relative or fractional terms that are range-independent: error magnitude and bias are quantified a percentage deviations, and retrieval robustness is expressed as a fraction. The regression slope and Pearson correlation coefficient, present in the original formulation, are indeed more sensitive to factors such as data dynamic range, which can compromise cross-variable comparability. As such, with the revised version we believe cross-variable comparisons are plausible.