Response to Reviewer 1 comments

Quantifying the scaling penalty of sequential coupling: The sequential coupling cost (Cseq) metric for Earth System Model components

by

Oriol Tintó Prims, Sergi Palomas, Simone Vacondio, and Mario C. Acosta

Reviewer (R)
Authors (A)

We sincerely thank the reviewer for their positive evaluation of our manuscript, for recognizing the value and timeliness of the sequential coupling cost (C_{seq}) metric, and for verifying our central calculations. We are particularly grateful for the constructive suggestions provided, which have significantly strengthened the paper's positioning, methodological rigor, and practical usability.

We have carefully revised the manuscript in line with your suggestions. Our detailed responses to each of your specific comments, along with the exact text updates, are provided below:

- **Positioning against your own prior work.**
  Two of you (Palomas and Acosta) are authors of Palomas et al. (2025, doi:10.5194/gmd-18-3661-2025), which already introduces a “partial coupling cost” and builds the optimal concurrent allocation N_conc = N_A(S) + N_B(S) from per-component scalability curves, which is the exact construction Cseq rests on; the motivating “13% on average” sentence at L24-L25 is in fact that paper’s phrasing. Yet Palomas et al. (2025) is not cited. As I read it, the novelty of Cseq is that it turns this same concurrent-optimization machinery around to diagnose the sequential case, which CPMIP scores as zero. The paper would benefit greatly from citing Palomas et al. (2025), relating it to the already-cited Acosta et al. (2023), and stating in one short paragraph exactly what Cseq adds. As it stands, a reader familiar with your group’s work will be unsure where the new contribution begins.

  The introduction has been updated to reference prior work from the same authors properly, and to better position how this work builds on top of that.

  Specifically, in lines 24-27:

> Performance studies involving multiple Coupled Model Intercomparison Project Phase 6 (CMIP6) experiments Acosta et al. (2024) showed that the coupling cost adds a computational overhead (in core-hours) between 5%-15% (13% on average). Previous work done to reduce this overhead by manually fine-tuning Acosta et al. (2023) or using dedicated heuristic-based tools Palomas et al. (2025) has proven to consistently cut this overhead down to ~7%.

Also in lines 40-44:

> While state-of-the-art couplers—such as OASIS (Craig et al., 2017) or YAC (Hanke et al., 2016)— provide diagnostic tools to quantify these synchronization costs in concurrent systems, and dedicated load-balancing methods and tools such as auto-lb (Palomas et al., 2025, Acosta et al. 2023) were developed to optimize concurrent resource allocations to minimize this idle time, existing frameworks offer no mechanism to evaluate the computational overheads inherent to sequential couplings.

- **OpenIFS-M7 target definition.**
  A second point, of similar weight, is that this use case defines its target throughput as the maximum the sequential setup itself can reach at 32 nodes (L231-L232, L246-L247). Because Cseq grows with node count, this is the choice that maximizes it, so the 46% and 51% headline figures are upper bounds rather than representative costs. I would ask you to say this plainly, and also to report Cseq for a fixed, science-relevant throughput so that the three use cases can be compared on a common basis.

  We agree with the reviewer that selecting the maximum throughput reachable at 32 nodes evaluates the model near its strong-scaling limit, making the 46.2% and 50.9% values upper bounds. We have updated Section 4.3 and the figure captions to state this explicitly. Furthermore, we clarify that the 32-node limit was chosen because it corresponds to the operational ceiling for high-priority queue submissions on the target HPC platform (MareNostrum 5). Rather than standardising the target definition across all three use cases (which are already significantly different scientifically and computationally), each was intentionally chosen to illustrate a different real-world decision-making scenario: a fixed scientific throughput milestone (4.1 IFS–NEMO), a parallel efficiency threshold (4.2 NEMO–SI3), and a max throughput under a resource-bounded limit (OpenIFS–M7).

  Specifically, lines 253-256:

> In this use case, rather than prescribing a fixed target throughput (as done for IFS-NEMO in Section 4.1) or an efficiency threshold (as done for NEMO-SI3 in Section 4.2), we evaluate an operational scenario common in HPC centres: maximising throughput under a fixed resource constraint of 32 compute nodes, which corresponds to the allocation threshold for the high-priority queue on our production platform (MareNostrum 5).

Furthermore, the captions of Figures 6 and 7 now state that the measured Cseq under this scenario is an upper bound:

> Notably, evaluating at this 32-node limit captures the highest achievable throughput, but also represents an upper bound on Cseq for this setup.

And lines 261-264 also acknowledge the results representing an upper bound:

> Results indicate that Cseq grows with node count as the parallel efficiency of both the atmospheric and aerosol components degrades. Because evaluating performance at the maximum throughput reachable under the 32-node budget pushes the sequential model to its strong-scaling limit within this partition, the resulting values—46.2% for the SIMC configuration and 50.9% for FULC—represent operational upper bounds for this setup.

- **A consistent overhead assumption.**
Related to this, the headline Cseq values assume a concurrent coupler with zero overhead, whereas your own OpenIFS-M7 node estimates add a 10% overhead (L242-L249); the two are then not on the same footing. I would report Cseq under a single, clearly stated overhead assumption. In the same paragraph, the claim at L267-L269 that concurrent frameworks "typically incur only 1-5% overhead (Balaji et al., 2017)" is not supported by that reference: Balaji et al. (2017) report costs rising with resolution to about 25%, and your own Acosta et al. (2024) give a 5-15% range. Please correct both the figure and its attribution.

We acknowledge that the sources of coupling costs and overhead assumptions were mixed in the manuscript.

We chose to retain the values reported by Acosta et al. (2024) (5–15%, with an average of 13%) rather than those from Balaji et al. (2017), as the former are more up-to-date and based exclusively on CMIP6 production runs.

Lines 24-27:

> Performance studies involving multiple Coupled Model Intercomparison Project Phase 6 (CMIP6) experiments (Acosta et al., 2024) showed that the coupling cost adds a computational overhead (in core-hours) between 5%-15% (13% on average).
> Previous work done to reduce this overhead by manually fine-tuning (Acosta et al., 2023) or using dedicated heuristic-based tools (Palomas et al., 2025) has proven to consistently cut this overhead down to ~7%

Lines 288-291 now read:

> CMIP6 performance evaluations show that well-balanced concurrent coupling frameworks typically incur only 5-15% overhead (averaging ~13%) in production (Acosta et al., 2024), and can be reduced to ~ 7% through dedicated load balancing (Acosta et al., 2023; Palomas et al., 2025). These values are substantially lower than the 13-50% inefficiencies identified in this work for sequential configurations.

Regarding the 10% overhead (L242–249): We've decided not to use this estimate in the comparison, as it only made the paragraph harder to read without adding any value.

- **Cost and accuracy as two axes.**
  I would encourage you to frame Cseq as one axis of a two-axis evaluation of coupling. Marti et al. (2021, doi:10.5194/gmd-14-2959-2021) and the more recent Schüller et al. (2025, doi:10.5194/gmd-18-9167-2025) quantify the physical inaccuracy of coupling schemes with a Schwarz iterative method run to convergence, whereas Cseq and CPMIP quantify the cost. The strict concurrent-versus-sequential dichotomy at L26-L28 would also read better if it acknowledged iteratively converged coupling as a distinct third category.

  We agree that coupling among model components is a source of error, and that, as highlighted by the two works referenced by the Reviewer, the choice of the coupling strategy ultimately impacts the accuracy of a model. Therefore, said choice should not just maximize performance, but also allow for scientifically acceptable accuracy.

  In this sense, within the context of a "two-axis evaluation" proposed by the Reviewer, our work is to be considered as solely focusing on the "performance" axis. We believe this is already made clear in our text. Lines 59-63 read

  > Furthermore, the choice of coupling strategy can dictate the underlying numerical algorithms. For instance, enabling concurrency may necessitate different schemes or lagged exchanges that influence model accuracy. Having a standardized way to quantify the computational cost is therefore essential for evaluating the inherent trade-offs between performance and physical fidelity.

  acknowledging possible trade-offs between accuracy and performance that the choice of a coupling strategy entails. The paragraphs that follow make it clear that our focus is solely performance.

  We take the opportunity to thank the reviewer for pointing us to the two works by Marti et al. and Schüller et al. respectively, which present some practical scenarios where the Schwarz iterative method (SIM) was applied to quantify the error introduced by common coupling schemes. The two works are now referenced in the passage quoted above.

  However, we would refrain from abandoning the concurrent-vs-sequential dichotomy mentioned by the Reviewer at lines 26-28. Both works mentioned above present the SIM as a means to quantify the coupling error, rather than a new coupling scheme itself. They propose practical solutions to ameliorate the coupling error that range from simply tuning the coupling period to adjusting implementations of specific methods within individual model components; meanwhile, they explicitly rule out the SIM as a solution due to its high computational cost.

  For all the above, since our analysis targets practically feasible coupling methods, we would eschew the introduction of a third coupling category.

- **The EC-Earth3 to EC-Earth4 comparison (Table 3).**
  This comparison should be framed as illustrative only. TM5 (about 300 km, 34 levels) and M7 (about 80 km, 91 levels) differ by more than an order of magnitude in degrees of

freedom, on top of differing in model version and HPC platform, so the implied 8.25% versus 46-51% contrast is not like-for-like. You are already candid about this (L227-L229, L254-L256); I would simply make it explicit at the table.

The caption of Table 3 now reads

> [...] A comparison **for illustrative purposes** is shown with the TM5 aerosol scheme [...]

- **The EC-Earth3 to EC-Earth4 comparison (Table 3).**
The 1 SYPD target is described as a milestone “for DestinE” (L155) but is actually a nextGEMS cornerstone, and its published reference is Segura et al. (2025, doi:10.5194/gmd-18-7735- 2025). For OpenIFS, L128 cites the OpenIFS@home paper (Sparrow et al., 2021), where Huijnen et al. (2022, doi:10.5194/gmd-15-6221-2022) would fit better. The M7 module at L127 is attributed to Huijnen et al. (2022), but its canonical reference is Vignati et al. (2004), which you already cite elsewhere.

We thank the reviewer for making us notice the mismatch between the project and the reference. We included both projects, DestinE and nextGEMS as both have 1 Simulated Years Per Day as objectives. We adjusted the references accordingly:

> The IFS-NEMO configuration used by ECMWF and within the nextGEMS (Segura et al., 2025) and DestinE (Wedi et al.,
>
> 2022; Doblas-Reyes et al., 2025) projects employs a sequential strategy where the NEMO ocean model is integrated via a function call inside the IFS atmospheric executable, …

and

> As can be seen in Figure 4, the computational requirements were evaluated using these analytical functions for a target throughput of S* = 1 Simulated Year Per Day (SYPD), a key performance milestone for nextGEMS (Segura et al., 2025) and DestinE (Doblas-Reyes et al., 2025).

The OpenIFS and M7 citations were corrected as follows:

> Finally, aerosol interactions are represented using the M7 microphysics model (Vignati et al., 2004) coupled to OpenIFS (Huijnen et al., 2022).

- **A worked recipe in the paper.** You already archive the Escalador library and the reproduction scripts openly on Zenodo, which is exactly right and lets others reproduce the three cases presented here. My suggestion is complementary: a reader should also be able to apply Cseq to a new model from the paper alone, without having to work through the code. A short appendix giving the quantities you take from the logs, the USL fit and its functional form, the N(S) inversion and the N_conc optimization, and one fully worked example would achieve that and would, I think, most increase uptake of the metric.

  We thank the reviewer for this excellent suggestion to improve the paper's practical utility and adoption. We completely agree that a reader should be able to apply the metric directly from the manuscript alone. To achieve this, we added Appendix A, which serves as a self-contained recipe to compute the sequential coupling cost. It is a standalone python implementation using only numpy as scipy (with no dependence on specialized profiling software). Additionally, we updated section 3 to clarify that escalador is just a helper library that we created to standardise and automate the analyses, but that the metric can be computed without any special tool. Lines 149-515 now read:

  > While the calculation can be performed using basic scientific software (a complete 20-line Python implementation using only `numpy` and `scipy` is provided in Appendix A), we used the *Escalador* library (Tintó Prims and Palomas, 2026) to standardize and automate this workflow across all test scenarios.

- **Extrapolation uncertainty.** Finally, several results rely on evaluating the USL fit outside the measured node range: the IFS-NEMO ocean optimum of 30.6 nodes lies below the 100-500 node range, and the NEMO-SI3 sea-ice optimum of 3.7 nodes below the 4-node reference. Since Cseq depends directly on these extrapolated node counts, a sensitivity estimate or confidence interval per use case would strengthen the headline numbers.

  We thank the reviewer for this crucial suggestion. To quantify the extrapolation uncertainty associated with evaluating USL fits outside the measured range (e.g., N_b = 30.6 nodes for NEMO), we conducted a residual bootstrapping analysis (B = 2,000 resamples) alongside an explicit sensitivity evaluation of C_{seq}. Propagating parameter uncertainty via bootstrapping yields a 95% confidence interval of [27.0, 35.5] nodes for N_b and [11.8%, 14.3%] for C_{seq} (median 13.1%). However, because our empirical dataset is small (n = 5), bootstrapping alone might underestimate out-of-sample extrapolation risk. We therefore complemented this with a direct sensitivity analysis of C_{seq} to account for potential underestimation when predicting node requirements at smaller scales.In this specific coupled configuration, the total node footprint (N_{seq} \approx 458.6) is overwhelmingly dominated by the atmosphere component (N_a \approx 368.6), making C_{seq} substantially insensitive to errors in the smaller ocean component. Analytically, the partial derivative of $\frac{\partial Cseq}{\partial Nb} = -\frac{100}{Nseq} = -0.22$ per node demonstrates that each additional node allocated to NEMO shifts C_{seq} by less than 0.22 percentage points. Furthermore, stress-testing shows that even an aggressive +30% underestimation of NEMO's node requirement (N_b = 39.8 nodes) only shifts C_{seq} from 12.94% to 10.94%. This dual analysis confirms that while N_b carries extrapolation uncertainty, the

headline sequential coupling cost remains substantial and robust due to the strong scale asymmetry between components. We have removed the previous paragraph where we were suggesting that the uncertainty would be small and updated it to include both the bootstrap confidence intervals and this sensitivity analysis. Lines 194-202:

> Calculating the optimal node count for the ocean component (Nb ≈ 30.6 nodes) requires extrapolating USL fits below the empirical range of 100–500 nodes. To estimate parameter uncertainty, we performed a residual bootstrapping analysis (B = 2, 000 resamples),yielding a 95% confidence interval of [27.0, 35.5] nodes for Nb and [11.8%, 14.3%] for Cseq (median 13.1%). However, because the empirical dataset is small (n = 5), residual bootstrapping may underestimate out-of-sample extrapolation risk. We therefore complement this with a direct sensitivity analysis of Cseq . In this specific coupled setup, where the atmosphere component heavily dominates the node footprint (Na ≫ Nb), Cseq is substantially insensitive to errors in Nb: partial derivative analysis shows ∂Cseq/∂Nb = − 100 / Nseq ≈ −0.22% per node. Even an aggressive +30% underestimation of the ocean node requirement (Nb = 39.8) only shifts Cseq from 12.94% to 10.94%, confirming that Cseq is overwhelmingly governed by Na ≈ 368.6 and robust against extrapolation variance in the smaller component.

We did not extend uncertainty quantification to NEMO-SI3 because the same logic applies and in that case the number of nodes allocated to sea-ice is much closer to the empirical range.

- **Technical corrections**

• Figure 5: the caption calls the ocean “the ocean component (green)”, but NEMO is drawn in purple and green is the concurrent model.
• Figures 6 and 7: the in-figure values (45.9%, 50.6%) differ from the text (46.2%, 50.9%), and the IFS-NEMO case is 12.9% in Figure 4 but 13% in the text.
• line 15: “leading inefficiencies” should be “leading to inefficiencies”.
• line 40: stray parenthesis in “EC-Earth (IFS() coupled to NEMO”.
• line 67: “we provide model developers and with a standardized formulation”, remove “and”.
• line 94: missing closing parenthesis after “(Cseq”.
• line 95: “Being T_A(N_A) the time it takes…” reads better as “Let T_A(N_A) be the time it takes…”.
• line 202: “with at N_NEMO = 15.6 nodes”, delete the stray “at” (or “with”).
• line 217: component naming is inconsistent (OpenIFS, OIFS and IFS-M7 all appear); please standardize.
• line 222: “the aerosols component” should be “the aerosol component”.
• line 282: “one-eight” should be “one-eighth”.
• line 283: “it can be seen than the ocean…” should be “that”.
• line 284: “drops by a 8%” should be “by 8%”.
• line 285: “of a similar magnitude than the ocean one” should be “to”.
• line 291: “than the previous two cases” should be “than in the previous two cases”.
• line 292: “resulted between 40% and 50%” needs a word (“ranged between…”); also hyphenate “32-node allocation”.

We thank the reviewer for their careful reading and thorough list of technical corrections. All suggested edits, including updating the Figure 5 caption colors, aligning in-figure values with the main text, standardizing component names, fixing missing or stray punctuation, and correcting all noted typographical and grammatical errors, have been fully incorporated into the revised manuscript.