

Dear Dr. Barge and co-authors,

Thank you for submitting the revised version of your manuscript, entitled “Deploying Machine Learning components coupled to Earth System Models with OASIS3-MCT (v6) and Eophis (v1.1)”, and for providing detailed responses to the reviewers.

The relationship between Eophis and OASIS3-MCT is now more clearly explained, the ESM–OASIS3-MCT–Eophis workflow is presented more sequentially, and the descriptions of the Tunnel, Router, time-stepping mechanism, halo reconstruction, and asynchronous execution have been strengthened. The newly added scalability analysis also provides useful context for interpreting the computational-performance results.

I consider the manuscript scientifically relevant and suitable for publication in Geoscientific Model Development. However, several unclear issues should be resolved before the manuscript can be accepted. I therefore invite you to submit a revision addressing the following points.

We thank the Editor for the positive assessment of our manuscript and for the constructive comments and suggestions. Please find below our responses to each of the point raised.

1. Complete and verify the code and data availability statement

- The Code and Data Availability section currently contains an unresolved placeholder: “Same material for ORCA1 test case coupled with Storto et al. (2024) is shared at ... (???)” Please replace this placeholder with the complete and correct repository citation and reference. If the repository is not yet publicly available, please clearly state its expected availability and ensure that all material required to reproduce the reported experiments is accessible before final publication.

The placeholder comes from a missing bracket in the LaTeX file, we have corrected it.

Related modifications on revised manuscript: Line 508

- Please also reconcile the apparent software-version inconsistency. The manuscript title and availability statement refer to Eophis v1.1, whereas the log presented in Figure A1 displays Eophis 1.0.2. Please clarify which version was used for the benchmark experiments and ensure consistency among the title, methods, figures, archived code, and reproducibility material. If the simulations were performed using an earlier version, please explain whether the changes introduced in v1.1 affect any reported results.

The version numbers are now consistent in the text. Figure A1 showed a use case that was historically ran on a machine that had not been updated to Eophis v1.1 and was included for demonstration purpose only. We reproduced it with the right version and confirm that all the experiences presented in the article have been performed and can be reproduced with Eophis v1.1.

Related modifications on revised manuscript: Figure A1.

- Reconcile the duplicated and mutually inconsistent atmospheric-forcing statements: one sentence attributes the forcing to ERA5 (ECMWF/CDS), while the following sentence attributes it to JRA-55 via the Geoscience Data Exchange (GDEX). Please retain the correct source, remove the redundant sentence, and correct the typo “Atmopsheric.”

The text has been modified in order to clarify that eORCA025 uses JRA55, and ORCA1 use ERA5.

Related modifications on revised manuscript: Lines 515 and 517.

2. Distinguish wall-clock performance from total computational-resource cost

The benchmark results demonstrate that asynchronous execution and flexible resource allocation can substantially reduce—or hide—the additional wall-clock time associated with the Python component. However, this improvement is achieved by allocating additional CPU or GPU resources.

The manuscript should therefore distinguish more carefully between:

- wall-clock or time-to-solution overhead;
- additional CPU/GPU resources;
- total computational cost, such as core-hours, node-hours, or GPU-hours.

Statements such as “minimal overhead,” “without affecting time-to-solution,” “negligible extra cost,” and claims that the Eophis implementation performs better than wrapper-based methods should be qualified accordingly.

For example, the main conclusion could be expressed as follows:

“Asynchronous coupling can largely hide the Python-related wall-clock overhead when sufficient additional resources are allocated, although the total computational-resource requirement increases.”

Please revise the Abstract, Sections 4 and 6, and any other relevant passages to avoid equating a small wall-clock penalty with negligible total computational cost.

Section 4 has been modified to distinguish more clearly between the wall-clock time-to-solution overhead and the additional computational resources required to reduce this overhead, including the distinction between CPU and GPU resources. The abstract and main conclusion have been modified to clarify that the reduction of the time-to-solution overhead is permitted by the allocation of additional computational resources.

Related modifications on revised manuscript: Lines 14-16, 271, 328-334, 351-355 and 484-492.

3. Provide sufficient information for reproducing the performance benchmarks

Because computational scalability is a central component of the manuscript, the benchmark configuration should be documented more comprehensively. Please provide, either in the main text, a table, or the supplementary/repository documentation, the following information:

- HPC system and node architecture;
- CPU and GPU models;
- number of CPU cores and GPUs per node;
- MPI implementation and compiler versions;
- Python, TensorFlow, CUDA, and relevant library versions;
- OpenMP-thread configuration, if applicable;
- NEMO domain decomposition;
- mapping of MPI processes to physical cores and nodes;
- placement of the NEMO and Python processes;
- relationship between the number of “Python processes” and the number of GPUs in the GPU experiments;
- number of repeated benchmark runs;
- averaging procedure and run-to-run variability of SYPD.

We thank the editor for pointing out a relevant missing part in the manuscript. We have added a paragraph describing the software and HPC environment. The averaging method is also described.

Related modifications on revised manuscript: Lines 307-314 and 319.

Please also use the terms “processes,” “CPU cores,” “nodes,” “GPUs,” and “resources” consistently, as these are not interchangeable.

The manuscript has been revised accordingly. We now use *MPI/Python/NEMO process* when referring to a software instance, and *CPU core* when referring to a physical computational resource. The mapping between CPU cores and processes is also explicitly specified.

Related modifications on revised manuscript: Lines 294-297. Several locations in Section 4.

4. Add numerical verification of the coupling implementation

The manuscript evaluates computational performance but provides limited evidence that the various coupling implementations produce numerically consistent results.

Please provide a concise verification showing, where applicable, that:

- Eophis-coupled Python calculations agree with standalone Python calculations;
- synchronous and asynchronous configurations produce identical or acceptably close results;
- the ORCA1-E, ORCA1-F, and ORCA1-I implementations produce consistent neural-network outputs;
- halo reconstruction is correct at MPI-partition and global-domain boundaries.

This verification need not constitute a full scientific validation of the parameterizations. A concise report of maximum absolute or relative differences, numerical tolerances, or relevant unit/integration tests would be sufficient. Please also clarify explicitly whether the asynchronous method uses data from the same model time step and whether it introduces any temporal lag.

We thank the Editor for these suggestions. The consistency between Eophis-coupled Python calculations and standalone Python calculations, as well as the correctness of the halo reconstruction at MPI-partition and global-domain boundaries, were checked by regression tests included in the Eophis source code. These tests are documented and available in the Eophis GH repository. We explicitly mention these tests in the manuscript.

Related modifications on revised manuscript: Line 266-268.

The synchronous and asynchronous configurations were quantitatively compared during the development of the coupling implementation. The results were found to be equivalent, as expected from the fact that the asynchronous implementation changes only the ordering of operations, while the same fields are exchanged. The results of this comparison were not retained and therefore cannot be reported in the revised manuscript. We have nevertheless explicitly stated in the manuscript that the equivalence of the synchronous and asynchronous results has been verified.

Related modifications on revised manuscript: Line 305.

We have added a quantitative comparison of the ORCA1-E, ORCA1-F, and ORCA1-I implementations. Specifically, Table 1 reports the root-mean-square errors between the neural-network outputs obtained with the three implementations, and the results are discussed in the revised manuscript

Related modifications on revised manuscript: Lines 374-386. Table 1.

Finally, we have clarified the temporal behavior of the asynchronous coupling. The main text specifies that the coupling is asynchronous in term of execution but remains synchronous in terms of model time, and does not introduce temporal lag.

Related modifications on revised manuscript: Line 303-305.

5. Moderate claims concerning generality and comparative performance

The framework is designed to be applicable to other OASIS3-MCT-enabled components, but the present experimental evaluation is limited primarily to NEMO. Statements implying that the method has been demonstrated for any global or regional ESM component should therefore be moderated.

A more precise formulation would be:

“The framework is designed to be applicable to other global and regional model components equipped with OASIS3-MCT, although the present evaluation is limited to NEMO.”

Similarly, the comparison with Fortran-wrapper approaches is based on a specific ORCA1 benchmark and on different resource allocations. Please restrict comparative conclusions to the experiments actually conducted. For example:

“In the ORCA1 benchmark, Eophis achieved comparable or higher wall-clock performance than the tested wrapper-based implementations when additional Python resources were allocated.”

We have modified the text accordingly.

Related modifications on revised manuscript: Lines 48, 494-496.

6. Clarification of Figure 2

Reviewer 2 suggested distinguishing the configuration phase from the execution phase and, if possible, presenting a timeline showing where Eophis operates in the coupling workflow. Although the revised caption and Section 3.1 provide a clearer textual explanation, Figure 2 still combines the structural components and runtime workflow in a single schematic. Please consider revising the figure to distinguish explicitly between (i) the configuration stage, including Tunnel, grid, exchange, and namelist setup, and (ii) the execution stage, including field reception, routing, Python-function execution, and return of the results. Alternatively, please explain clearly why the current representation is preferable.

The figure has been revised accordingly. The configuration and execution phases are now clearly distinct and the different steps of the workflow described in the main text are represented.

Related modifications on revised manuscript: Figure 2.

7. Clarification of the scope of the cost analysis

Reviewer 2 specifically suggested that the manuscript briefly address the relative price and energy consumption of CPU- and GPU-based execution. In your response, you state that Durand et al. (2027) provides a dedicated analysis of these topics and that the new Appendix B partly addresses cost considerations.

Please revise this response and the corresponding discussion in the manuscript. Durand et al. identifies GPU accessibility, computational cost, and carbon footprint as important considerations for the adoption of learning-based ocean models. However, it does not provide a quantitative comparison of CPU and GPU monetary costs, power consumption, or energy-to-solution. Similarly, Appendix B primarily presents an analytical interpretation of runtime and scalability rather than an assessment of financial cost, energy consumption, total resource consumption, or carbon emissions. Please consider renaming or describing it as a runtime and scalability analysis, unless an explicit resource-cost metric is added.

We thank the Editor for this clarification. We agree that Durand et al. (2026) does not provide a quantitative comparison of CPU and GPU costs or energy consumption, and that Appendix B should not be interpreted as such. We have revised the corresponding discussion accordingly and clarified that the cost/benefit trade-offs between CPU- and GPU-based neural-network inference are

outside the scope of the present study. Appendix B has also been described more explicitly as a runtime and scalability analysis.

Related modifications on revised manuscript: Appendix B, lines 340-347.

A new hardware-energy benchmark is not required for the present revision. However, please:

1. avoid characterizing Durand et al. as providing a dedicated quantitative CPU–GPU cost or energy analysis;
2. add a concise statement in the manuscript acknowledging that the present benchmarks evaluate wall-clock performance and resource allocation, but not monetary or energy efficiency;
3. distinguish clearly among wall-clock overhead, total computational-resource use, financial cost, and energy consumption;
4. consider renaming Appendix B as “Runtime and scalability analysis”, unless an explicit resource- or energy-cost metric is added; and
5. verify and correct the bibliographic details of Durand et al. The cited work is a non-peer reviewed EarthArXiv preprint (<https://doi.org/10.31223/X5HB6S>) dated 2026, whereas your response cites it as 2027; please correct the year, indicate its preprint status, and update the citation if a peer-reviewed version becomes available before publication.

8. Final technical and editorial corrections

Please carefully proofread the manuscript. In particular:

- Figure 8 refers to eORCA075.L75, whereas the experiment appears to use eORCA025.L75.
- “ORCA-I” should be standardized as “ORCA1-I.”
- Correct the incomplete reference marked “(???)”.
- Correct duplicated DOI prefixes such as [https://doi.org/https://doi.org/....](https://doi.org/https://doi.org/...)
- Correct “Journal of Open Source article” to the proper journal title.
- Correct “Nature Reviews Earth amp; Environment.”
- Check the spelling and bibliographic details of all software and reference authors.
- Correct grammatical expressions including “allows to deploy,” “If the Fortran code send,” “fields sendings and receptions,” “outcoming fields,” “from these information,” “the 20 first levels,” “take into account of,” and “implmentations.”
- Arrange the institutional affiliations in a consistent numerical order.
- Please use a single, consistent designation for each component throughout the manuscript. In particular, terms such as “ESM,” “geoscientific model,” “geophysical model,” “Fortran model,” and “physical model” appear to refer to the same component. Similar inconsistencies occur for “Python script,” “ML Python script,” “Python component,” and “ML component,” as well as for “coupling fields,” “exchanged fields,” and “input fields.”

We thank the Editor for carefully proofreading the manuscript. All the technical, grammatical, bibliographic, and typographical corrections raised in this comment have been addressed, and the terminology has been standardized throughout the manuscript.

Please submit a revised manuscript together with a detailed point-by-point response explaining how each of the above issues has been addressed. Please also provide a marked-up version highlighting the changes.

Thank you again for your careful revision and for this valuable contribution to the development of reproducible hybrid Earth system modelling.

Sincerely,
Jinkyu Hong