The authors thank the reviewer and editor for their helpful comments. In-line citations have been added to the two Zenodo archives in the “Code and data availability” section.

Reviewer #1

General Comments

The revised manuscript has addressed most of my previous comments in a satisfactory manner. In particular, the additional details on halo exchanges, communication volume, and memory layout considerations improve the clarity and reproducibility of the performance analysis. I appreciate the authors’ efforts to respond carefully to the review.

I have one remaining issue that should be clarified before publication.

Specific Comments

1. Clarification of CPU–GPU utilization statement (around line 479)

   The manuscript currently states:

   “In the ‘GPU’ simulations, all CPUs and GPUs on each node are fully utilized for the timing test.”

   Based on the authors’ response to my previous comment, in GPU builds the compute-intensive kernels are executed exclusively on the GPUs, while CPUs are primarily responsible for flow control, kernel launches, communication, and I/O. In other words, there is no actual numerical workload sharing between CPUs and GPUs in the reported GPU runs.

   As written, the phrase “all CPUs and GPUs are fully utilized” strongly suggests that both CPUs and GPUs are actively used for numerical computation in a coordinated or balanced fashion. This is potentially misleading and could lead readers to misunderstand the execution model used in the performance benchmarks.

   I recommend revising this sentence to clearly distinguish between:

   * resource allocation (e.g., one MPI task per CPU–GPU pair), and

   * execution of compute-intensive kernels exclusively on the GPU.

   A more precise formulation would improve clarity and prevent misinterpretation of the performance results.

   Author response: We have added a sentence to specify that for the “GPU” tests the majority of computational work is handled by the GPU, while a small number of CPUs are used for flow control, kernel launches, synchronization, and I/O. We refer the reader to Table 5, which specifies the number of CPUs per node used for each of these performance tests.