

We would like to thank the referees for their careful reading of our manuscript. We are most grateful for the comments, constructive criticism and very useful suggestions received on how to improve the paper. Please find attached our detailed answers to the questions and a new version of the paper, which we hope satisfactorily to address the points raised during the discussion. The referee's comments are presented in black, and our responses in blue.

REFeree 2 COMMENTS:

- The primary limitation I see is the restriction to NVIDIA devices. I would recommend including a brief discussion of why NVIDIA devices were chosen and any thoughts the authors might have about alternative device support and implications for the accelerated model's portability.

We agree that portability across different heterogeneous architectures is an emerging requirement for the ESM community. NVIDIA devices were targeted to leverage the existing MEDINA Legacy code and because of the availability and support on the JUWELS Booster supercomputer that was used for our simulations runs.

The following paragraph has been added to the manuscript (line 362):

“Medina v2.0 has exhibited great performance gains. However, the implementation currently only generates CUDA kernels and hence, only NVIDIA GPUs are supported. These devices were specifically targeted for this development due to the MEDINA legacy code that allowed for quick refactoring and development, and the availability of NVIDIA devices on the JUWELS booster supercomputer. As the HPC and Earth System Modelling community adapts different heterogeneous architectures featuring AMD and Intel accelerators, portability is becoming a critical requirement. In the future, MEDINA, as an automated code-to-code compiler, can potentially address this limitation, by extending the framework to other programming models and open standards. This would ensure that the accelerated chemical kinetics simulations are vendor-agnostic and portable across different supercomputing environments and heterogeneous architectures.”

- Additionally, complementing the strong scaling analysis with a weak scaling analysis could support the authors' scalability assertions. Figure 8's two panels together act as a sort of informal weak scaling analysis, but a more formal weak scaling, e.g., co-varying resources with VL_GLO, would strengthen this argument.

We thank the reviewer for this suggestion, which strengthens our scalability assertions. We agree that a comprehensive weak scaling analysis will demonstrate whether the implementation suffers from communication overhead at scale and to address this, we conducted a weak scaling assessment with up to 64 GPUs (16 nodes), maintaining a constant workload per GPU while proportionally increasing the offloaded total grid size. Figure 4 was added to the manuscript, which demonstrates the processing rate and weak scaling efficiency.

The following discussion has been added in line 227:

“Figure 4 shows the weak scaling behaviour of the two representations, single- and double-precision, for the GPU implementation of the box model. The processing rate (P_N) increases with the number of GPUs, as shown in the left panel. The FP32 single precision implementation constantly sustains a higher processing rate compared to FP64 across all node configurations. Fig. 4 (right panel) presents the weak scaling efficiency (E_{weak}), with high scalability for both representations, as both manage an efficiency of over 99% for up to 64 GPUs. These results indicate that the execution time remains almost constant regardless of the total number of nodes employed, as the embarrassingly parallel nature of the problem means there is no communication overhead between GPUs.”

- Figs. 2, 6: Please include a measure of deviation from the 1:1 line, e.g. RMSE or NME. Visual inspection for deviations on a log scale spanning ~50 orders of magnitude obscures any actual deviations. Alongside RMSE or NME, something like an integrated difference over the box or domain could provide an idea of how much mass is over- or underpredicted compared to the baseline.

This is an insightful suggestion, and we agree that relying on log-log representation can obscure minor deviations across a vast range. We have introduced Normalized Mean Error (NME) and the integrated mass bias to address this comment. Both Figure 2 and Figure 7 have been updated to show the metrics, and we have expanded the manuscript text to discuss these statistical validations.

The following text has been added in Line 191:

“As these logarithmic scale scatter plots span 50 orders of magnitude, minor deviations might be obscured. Thus, two additional metrics are introduced to assess the solver’s fidelity: the Normalized Mean Error (NME) (Eq. 1) and the integrated bias (Eq. 2). The NME measures the absolute magnitude of the deviations relative to the reference, defined as:

$$\text{NME}[\%] = \left[\sum |x^{\text{SP}}_i - x^{\text{DP}}_i| / \sum x^{\text{DP}}_i \right] \times 100$$

where x^{SP}_i and x^{DP}_i represent the final concentrations (or masses) of species i for the single-precision and double-precision implementations, respectively.

The bias is used to identify any systematic over- or under-prediction, which indicates an artificially created or destroyed mass. Bias is expressed as:

$$\text{Bias}[\%] = [\sum (x^{SP}_i - x^{DP}_i) / \sum x^{DP}_i] \times 100$$

As displayed in the two plots, the NME is exceptionally low at $1.16 \times 10^{-5} \%$, which is also complemented by a negligible integrated bias of $-3.7 \times 10^{-6} \%$, confirming that single precision manages to strictly conserve mass and does not systematically mispredict concentrations.”

and in line 280:

“The visual agreement across the full domain is also confirmed mathematically with the statistical deviations NME (Eq. 1) and Bias (Eq. 2). As shown within the plot, these are exceptionally small, measuring at $1.19 \times 10^{-8} \%$ for the former and at $1.12 \times 10^{-8} \%$ for the latter, demonstrating that no systematic mass drift or numerical divergence occurs with the employment of single precision solver for global-scale simulations over extended periods.”