

Reviewer #1

This manuscript introduces a two-stage Rosenbrock solver (ROS2) into the E3SMv1 model to address the significant computational bottleneck associated with atmospheric chemistry integration. By evaluating the solver across two chemical mechanisms of varying complexity (chemUCI and trop_strat_mozart_mam4), the authors demonstrate that the ROS2 solver can reduce the computational cost by approximately 33% at a 180-second timestep, while maintaining a remarkably low global mean surface ozone bias compared to the default fully implicit solver.

Overall, this paper addresses a critical challenge in Earth System Modeling. The implications of improving the computational efficiency of atmospheric chemistry are broad and highly significant for the development of next-generation, high-resolution climate models. The motivation is strong, the methodology is novel, and the manuscript is generally well-written and clearly structured. I recommend this manuscript for publication after the authors address the following comments.

1. The core contribution of this paper is the enhancement of computational efficiency for atmospheric chemistry modeling within Earth System Models (ESMs), which has broad implications for the modeling community. However, the current Introduction and Conclusion seem to focus too heavily on ozone right from the beginning. I suggest revising the opening paragraphs of both sections to frame the narrative around the broader context, challenges, and historical development of atmospheric chemistry modules in ESMs. Ozone can then be introduced subsequently as a key metric for evaluation, rather than the primary motivation of the paper.

Response: We sincerely thank the reviewer for this important framing suggestion. We agree that the main contribution of the study is methodological: implementation and evaluation of an alternative time-integration scheme for stiff, multiscale atmospheric chemistry in an Earth System Model. Ozone is an important diagnostic species, but it should not appear to be the sole motivation for the numerical development. We therefore changed the title of this paper, not solely focusing on the surface ozone, and revised the Introduction to emphasize interactive atmospheric chemistry in GCMs/ESMs and the computational challenge created by chemical systems whose characteristic timescales span seconds to months. Ozone is introduced subsequently as a representative climate- and air-quality-relevant diagnostic. We also rewrote the Results and Conclusion to include NO, NO₂, and OH and to focus the principal conclusions on solver efficiency, numerical stability, and species-dependent numerical sensitivity rather than on ozone alone.

Revised title: “Using a two-stage Rosenbrock solver to increase computational efficiency in the DOE’s Exascale Earth System Model version 1 (E3SMv1)”

Revised manuscript text (lines 25-43): “The development of the Air Quality Model (AQM) has been closely intertwined with the knowledge of photochemical oxidants. The non-linear photochemical ozone formation process, which involves nitrogen oxides, hydrocarbons, and solar radiation, has been recognized as playing a major role in the Los Angeles smog event in the mid-twentieth century (Haagen-Smit, 1952). Reynolds et al. (1973) derived a sequence of mathematical programs to show the dynamical behaviors between chemical emissions and the response of air pollution contaminants. This model numerically represents atmospheric transport, chemical kinetic processes, complex photochemical reactions, and the impact of the emission inventory, and has been used to study urban ozone formation, as well as the precursors of these secondary pollutants. The development of the regional-scale model provides a framework for one-way (from the meteorological process to the chemical reaction) simulations and was eventually extended to a global scale. The chemistry-climate interaction process in the general circulation models (GCMs) has been used to study the global radiation budget, which is affected by tropospheric ozone (Mickley et al., 1999); while the global scale of the chemical transport models (CTMs) has been used to study the tropospheric ozone-NOX-hydrocarbon chemistry (Bey et al., 2001).

Tropospheric ozone is used as a key diagnostic species in interactive chemistry simulations, yet it is also one of the major air pollutants that has been monitored for decades due to its adverse impacts on human health, ecosystems, and agricultural productivity. Ground-level ozone can trigger respiratory disease, penetrate leaves, and reduce crop yields; therefore, it must be monitored and predicted as both an air pollutant and a climate-relevant species (Markandya et al., 2018; Vandyck et al., 2018). Accurate simulation of surface ozone remains a core capability of numerical models used for air-quality assessment and Earth-system prediction, which also requires consistency of the precursors, for instance the NOX emission inventory, and other species or photochemical reactions, which leads to an expensive cost in model computation.”

2. The authors note that at a 180-second timestep, the ROS2 solver achieves a ~33% efficiency improvement for both the MOZART and chemUCI mechanisms. Given the significant difference in complexity (number of species and reactions) between these two

mechanisms, is this identical 33% reduction a coincidence? Or does it imply that the relative efficiency gain of the ROS2 solver, compared to the default implicit solver, is independent of the chemical network's size? The authors should discuss the theoretical or practical computational reasons behind this consistent scaling.

Response: We thank the reviewer for raising this important scaling question. The similar ~33% reductions measured for chemUCI and trop_strat_mozart_mam4 at a 180s timestep should not be interpreted as evidence that the ROS2 speedup is independent of mechanism size. This reduction measured here should be interpreted as configuration-specific, since the mechanisms differ substantially in implicitly solved species, reactions, and non-zero Jacobian elements. The relative cost of the two solvers depends not only on matrix dimension, but also on how often the iterative implicit method requires additional Newton iterations, fails its convergence criterion, or subdivides and repeats a chemistry step. ROS2, in contrast, uses a fixed two-stage linearly implicit update for each accepted timestep. Therefore, the relative speedup depends on mechanism structure, local stiffness, chemical state, convergence behavior, sparse-matrix operations, and the fraction of chemistry cost spent in nonlinear iterations. In the one-day E3SM tests, these factors happened to yield very similar relative reductions for the two mechanisms. We do not regard this as a universal scaling relationship. This interpretation is also consistent with our box-model tests, for which the ROS2 improvement spans approximately 20-48%. We revised the manuscript to make this limitation explicit.

Revised manuscript text (lines 268-277): “The computational cost of each numerical configuration is evaluated using one-day simulations on 693 and 1350 CPU cores (Fig. 3). For trop_strat_mozart_mam4, ROS2_180 requires 67.64% of the IMP_180 computational time using 693 cores and 67.20% using 1350 cores. For chemUCI, the corresponding values are 67.38% and 67.66%. Thus, at the same 180s chemistry timestep, ROS2 reduces the measured integration cost by approximately one third relative to the original iterative implicit solver on this platform. The similar percentage reduction for the two mechanisms should not be interpreted as a mechanism-independent scaling law. The relative cost of IMP varies with nonlinear convergence behavior, including additional iterations and timestep subdivision/repetition, whereas ROS2 uses a fixed two-stage update. Consistent with this interpretation, the standalone box-model tests show a broader improvement range of approximately 20-48%. IMP_1800 remains the least expensive of the three tested configurations because it takes one chemistry step for every ten steps at 180s; therefore, the demonstrated ROS2 performance advantage applies specifically to the same-timestep IMP_180 versus ROS2_180 comparison.”

3. While the global mean surface ozone bias is small, Figure 5 reveals substantial regional discrepancies. The current manuscript lacks a detailed discussion on what drives these specific regional differences (e.g., are they related to specific emission hotspots or distinct meteorological regimes?). Furthermore, focusing almost exclusively on ozone provides a somewhat limited perspective. The spatial differences of other critical atmospheric components, such as NO₂, OH and aerosols should also be presented and discussed. I highly recommend strengthening this section to provide a more comprehensive spatial evaluation of the solver's performance.

Response: We thank the reviewer for requesting a more comprehensive spatial evaluation. We agree that global-mean ozone alone is insufficient. We therefore added global surface-difference maps for NO, NO₂, and OH in Figs. S2-S4. These species provide complementary diagnostics because they include precursor species and a highly reactive radical with much shorter chemical timescales than ozone. The added maps show that the sign and magnitude of the ROS2_180-IMP_1800 differences vary by species and chemistry mechanism rather than exhibiting a single global offset: NO generally shows small decreases, NO₂ changes sign between the two mechanisms, and OH differences remain small. We now interpret this as species- and mechanism-dependent numerical sensitivity rather than a systematic ROS2 bias.

We also made the interpretation of regional ozone differences more cautious. The present experiments do not isolate whether a particular regional pattern arises from emissions, meteorology, photolysis, transport, or local nonlinear chemistry. A targeted process-budget analysis would be needed for that attribution. Regarding aerosols, we agree that they are important; however, aerosol concentrations in the MAM-coupled configuration depend on gas-phase chemistry together with condensation/evaporation, phase partitioning, nucleation, coagulation, and other aerosol microphysical processes. An aerosol difference therefore would not isolate the gas-phase time integrator without additional process separation. We now state this limitation explicitly and use NO, NO₂, and OH as the direct additional gas-phase solver diagnostics.

Revised manuscript discussion (lines 278-289): “To assess solver behavior over a one-year integration, we compare the spatial distribution of several surface species. Figure 4 shows the surface-ozone difference between ROS2_180 and IMP_1800. For chemUCI, the differences are generally small, with modest positive differences over parts of the Northern Hemisphere and

localized regional features. For trop_strat_mozart_mam4, ROS2_180 produces lower surface ozone over much of the globe, with larger negative differences over North Africa. Supplementary Figs. S2-S4 provide corresponding differences for NO, NO₂, and OH. The sign and magnitude of these differences vary by species and mechanism: NO generally shows small decreases, NO₂ changes sign between the mechanisms, and OH differences remain small. No common-sign spatial response is evident across the evaluated gas-phase species. Because these regional patterns can reflect interactions among precursor abundance, radical chemistry, photolysis, transport, and meteorological conditions, the present experiments do not uniquely attribute individual regional differences to emissions or meteorology. A process-budget analysis would be required for that attribution. Aerosol fields are not used as a primary solver diagnostic here because they also depend on aerosol microphysics and phase partitioning, which are not isolated by the present gas-phase time-integration experiment.”

4. Although the baseline simulation shows small bias for ozone, the sensitivity of the chemical system to non-linear responses might show larger difference when using the two solvers. To ensure the ROS2 solver behaves consistently with the implicit solver under perturbed conditions, I suggest conducting a simple sensitivity test. For instance, the authors could run a short test where global NO_x emissions are reduced by 20%. Comparing the ozone response between the two solvers under this perturbed scenario would robustly demonstrate whether the acceptable bias holds for non-linear chemical regimes.

Response: We appreciate this suggestion and performed a 20% NO_x-emission perturbation in the standalone box-model framework for both chemistry mechanisms. We also revised the interpretation to distinguish clearly between chemical sensitivity and solver consistency. Figure S1 now explicitly identifies ozone on the x-axis/caption and compares the empirical ozone distributions produced by IMP and ROS2 under the baseline external NO_x emissions (100%) and the perturbed case (80%). The chemically meaningful question is not whether the 100% and 80% cases are identical—they should not be—but whether the two numerical solvers remain mutually consistent within each emissions scenario. In both mechanisms, the IMP and ROS2 ozone distributions remain closely aligned after the NO_x perturbation, indicating that solver-to-solver agreement does not degrade when the chemical system is moved away from the baseline state.

We agree that a fully coupled global 20% NO_x experiment would be an even stronger test because it would include transport, meteorology, and spatially varying emissions. We therefore describe the present experiment as a targeted numerical sensitivity test rather than as a complete substitute for the suggested global perturbation. Its purpose is to test whether ROS2 develops an evident instability or qualitatively different numerical response under a nonlinear NO_x perturbation. This scope and limitation are now stated explicitly in both the manuscript and supplement.

Revised manuscript text (lines 252-258): “To test solver consistency under a perturbed nonlinear chemical regime, the standalone box-model calculations were repeated after reducing external NO_x emissions by 20%. Figure S1 compares the empirical cumulative distributions of ozone produced by IMP and ROS2 under the baseline (100% NO_x) and perturbed (80% NO_x) cases for both mechanisms. The NO_x perturbation changes the chemical state, as expected; however, within each emissions scenario the IMP and ROS2 ozone distributions remain closely aligned. Thus, the solver-to-solver consistency observed under the baseline is retained under the tested NO_x perturbation. This is a targeted numerical sensitivity test; a fully coupled global emissions-perturbation experiment would be required to quantify the complete spatial atmospheric response.”

Technical Comments: The x and y axes are currently missing Fig.1 labels. I believe they represent the number or index of the chemical species. Please add appropriate axis labels to clarify what is being plotted.

Response: Thank you for pointing this out. We revised both the axis description and the mathematical explanation. Figure 1 shows the sparsity pattern of the chemistry Jacobian. If the chemical system is written as $dC/dt = f(C)$, then $J_{ij} = \partial f_i / \partial C_j$. The x-axis corresponds to the permuted index j of the concentration/state species C_j , and the y-axis corresponds to the permuted index i of the chemical tendency/equation f_i . Each blue point indicates a non-zero derivative. The complete permuted species order is given in Tables S1 and S2. This is more precise than the original wording that described the axes simply as reactants and products.

Revised Fig. 1 caption: “Figure 1. Sparsity pattern of the chemistry Jacobian for (a) chemUCI (53 implicitly solved species) and (b) trop_strat_mozart_mam4 (130 implicitly solved species). For $J_{ij} = \partial f_i / \partial C_j$, the x-axis gives the permuted concentration/state-species index j and the y-axis gives the permuted tendency/equation-species index i . Each blue point denotes a non-zero Jacobian element. The corresponding species order is provided in Tables S1 and S2.”

Revised explanatory text (lines 175-178): “For the chemical system $dC/dt = f(C)$, the Jacobian element $J_{ij} = \partial f_i / \partial C_j$ represents the sensitivity of the chemical tendency of species i to the concentration of species j . Because most species do not interact directly

through every reaction pathway, the Jacobian is sparse. The fixed species permutation used by the solver changes only the ordering of the unknowns, not the chemistry.”

Reviewer #2

Goal: evaluate the ROS2 solver in E3SMv1 with two different chemistry packages.

ROS2 was previously shown to work well in the CESM.

1. Introduction: References describing the status of atmospheric chemistry in modern Earth system models seems dated. It mentions CMIP3 models (from 2007) and all the papers regarding chemistry treatment in global models only go up to 2013. Has the situation changed in the CMIP6 models?

Response: We thank the reviewer for noting that the original background relied too heavily on older CMIP3-era examples. We updated the Introduction with CMIP6-era information and now distinguish between model configurations that include interactive or partially interactive atmospheric chemistry and those that use specified or prescribed oxidants. Based on the model descriptions summarized in IPCC AR6, 32 of the 60 CMIP6 coupled climate/Earth system models describe atmospheric chemistry components, and 11 of those use specified or prescribed oxidants. We use this count only to establish that interactive chemistry is still not universal in CMIP6, not to imply that every model has only one fixed chemistry configuration. The revised Introduction therefore uses the CMIP6 information as current context for why reducing the cost of interactive chemistry remains relevant.

Revised manuscript text (lines 55-60): "Of the 60 coupled climate and Earth system models participating in CMIP6, 32 describe atmospheric chemistry components, while 11 of those use specified or prescribed oxidants in the configurations summarized by IPCC AR6 (IPCC, 2021). Thus, interactive or partially interactive atmospheric chemistry remains present in only a subset of CMIP6 models. This provides a current motivation for numerical methods that reduce the computational burden of solving stiff chemical systems without changing the underlying reaction mechanism."

2. Line 166: "In this paper (Chien and Fu, 2026b), ..."

It is odd to self-site the actual paper I'm reading. Seems like an AI thing.

Response: We agree that the original wording was confusing. Chien and Fu (2026b) is not a citation to the present manuscript; it is the archived source-code record associated with the ROS2 implementation/box-model code and is cited for reproducibility and GMD code-availability requirements. We removed the phrase "In this paper" so that the citation cannot be mistaken for self-citation to the manuscript itself.

Revised sentence (lines 171-172): "The source code for the two-stage Rosenbrock implementation used in this study is archived in Chien and Fu (2026e)."

3. Line 198 "All simulations are conducted on the Cray XC-based Cori supercomputer..." This computer was decommissioned in 2023 and the architecture did not prove to be successful, so it's not representative of leadership-class architectures in 2026. No objections to using this old data to draw computational speed conclusions, but the authors should drop the last sentence in this paragraph and the previous claims about exascale modeling.

Response: We agree with the reviewer. Cori/KNL is a decommissioned CPU architecture and should not be presented as representative of 2026 leadership-class or exascale systems. We therefore removed language that uses the Cori timing results to claim performance on current leadership-class, GPU-accelerated, or exascale platforms. The revised manuscript describes Cori explicitly as the historical CPU platform on which the experiments were performed and treats the measured speedup as a controlled, hardware-specific benchmark for CPU-based E3SMv1. Statements about higher-resolution future models are retained only as motivation for why efficient short-timestep chemistry integration is useful; performance portability to contemporary CPU/GPU architectures is now identified as future work.

The corresponding exascale-performance claims were also removed from the Results and Conclusion. The revised Conclusion now states that additional benchmarking on contemporary CPU/GPU platforms is required to quantify architecture-dependent speedup and performance portability.

Revised manuscript text (lines 207-208): "All simulations were conducted on the now-decommissioned Cray XC-40 Cori system at NERSC using Intel Knights Landing processors in stand-alone CPU mode (Doerfler et al., 2018)."

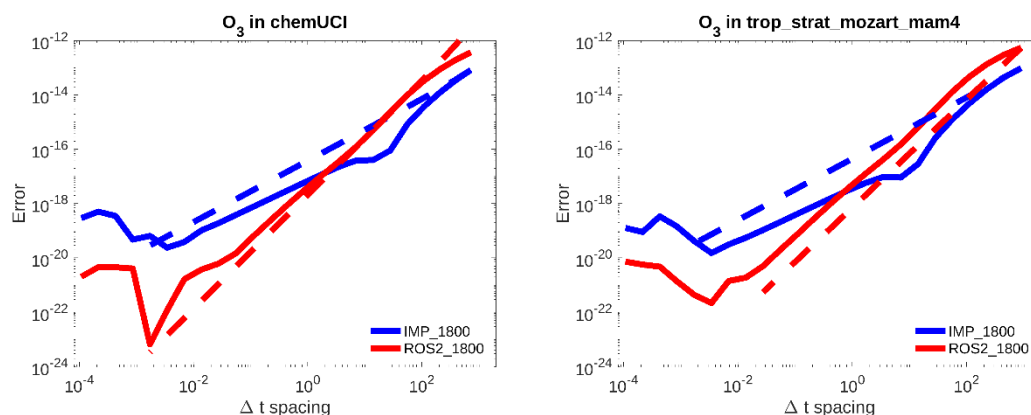
Revised manuscript text (lines 210-214): “The same hardware and model configuration were used for all solver comparisons. These timings therefore provide a controlled benchmark for the CPU-based E3SMv1 implementation and should not be interpreted as a direct measure of performance on contemporary GPU-accelerated or exascale architectures.”

4. Figure 2: Confusing. x-axis label "number of halven". Part of my confusion was my own ignorance - I was not familiar with the word "halven". However, for this types of plots I think it is more conventional to choose an x-axis so the slope is k . I think with the choice here, the slope is $-1/k$? If the authors prefer their current plot, at least mention in the text how the slope is related to k .

Response: We thank the reviewer for this correction. We replaced the non-standard “number of halven” representation with timestep spacing Δt and use conventional log-log axes. We also clarified the interpretation of the slope. If an error estimate behaves asymptotically as $E(\Delta t) \propto (\Delta t)^k$, the slope of $\log E$ versus $\log \Delta t$ is an empirical estimate of k . In the present full stiff chemical system, however, an analytical reference solution is unavailable and the sampled states are not guaranteed to lie in the asymptotic regime. We therefore no longer describe the fitted slopes as direct measurements of the formal first- or second-order accuracy of the solvers. They are reported as empirical timestep-refinement rates. ROS2 shows the steeper empirical trend in both mechanisms, with mean slopes of approximately 0.68 and 0.50 versus 0.34 and 0.30 for IMP.

Revised manuscript text (lines 240-242): “The log-log slope of the timestep-refinement difference is used here as an empirical convergence-rate diagnostic. Because the analytical solution of the full stiff chemistry system is unavailable and the sampled states are not guaranteed to lie in the asymptotic regime, the fitted slope is not interpreted as a direct measurement of the formal order of either solver.”

Revised result statement (lines 248-250): “For chemUCI, the mean empirical slopes are 0.34 for IMP and 0.68 for ROS2; for trop_strat_mozart_mam4, they are 0.30 and 0.50, respectively. ROS2 therefore exhibits a steeper timestep-refinement trend in both mechanisms. These values should not be equated directly with the ideal first- and second-order limits,”



5. Figure 6: "In both cases, the amplitude, phase, and variability of the diurnal cycle are nearly identical across solvers, indicating that ROS2 preserves key temporal characteristics while improving numerical efficiency."

How do the authors conclude this from Figure 6? ROS2_180 seems significantly different than IMP_1800 and IMP_180. How do we know it is more accurate? Maybe if the timestep was reduced, ROS2 would quickly converge to the IMP_180 result - which would mean that it is actually less efficient than IMP_180.

Response: We agree with the reviewer that Fig. 6 does not establish that ROS2_180 is more accurate than IMP_180 or IMP_1800. The original wording overinterpreted the curves. We have removed the claim that the cycles are “nearly identical” and no longer use Fig. 6 as an accuracy benchmark. The revised text states only what the figure supports: the broad phase/day-night structure is retained, while ROS2_180 exhibits a mechanism-dependent mean offset relative to the IMP configurations. Without an exact fully coupled reference solution, proximity of one curve to another cannot establish an accuracy ranking. We therefore separate the accuracy question from the timing result. The demonstrated computational advantage is the controlled same-timestep comparison: at $\Delta t = 180s$, ROS2 is less expensive than IMP on the tested E3SMv1/Cori configuration.

Revised manuscript text (lines 297-303): “Figure 6 compares the global diurnal cycles of surface ozone. In chemUCI, ROS2_180 has a slightly higher mean ozone level than IMP_1800, whereas in trop_strat_mozart_mam4 it has a slightly lower mean. The

timing and broad day-night structure remain similar, but the curves are not treated as an accuracy benchmark. In particular, the separation between ROS2_180 and the two IMP configurations shows that solver formulation can influence the mean numerical trajectory. Because no exact reference solution is available for the fully coupled one-year integrations, proximity of one curve to another does not establish an accuracy ranking. The computational result is therefore kept separate: ROS2 is less expensive than IMP when both are operated at the same 180-s timestep on the tested platform.”

6. For figure 7 - same question. For plot (a) (c) and (d), we see that ROS2_180 matches the IMP_180 result, which we can assume is more accurate than IMP_1800, and thus ROS2_180 is more efficient (since it is faster than IMP_180. But in panel (b), IMP_1800 and IMP_180 agree- suggesting the IMP method has converged, and leaves open the possibility that ROS2_180 needs a smaller timestep to converge as well, making it less efficient.

I actually would expect ROS2_180 to be more accurate than IMP_180, and thus I think the authors conclusions are probably correct, but I don't see how the evidence presented justifies the conclusion. If the authors can better justify these claims, the results would be a nice contribution.

Response: We appreciate this careful analysis and agree that panel (b) prevents a simple claim that the configuration closest to another curve is necessarily the more accurate one. Agreement between IMP_1800 and IMP_180 for OH in trop_strat_mozart_mam4 may indicate relatively weak timestep sensitivity of IMP for that diagnostic, but it does not provide an exact reference against which ROS2_180 can be ranked. Conversely, the closer agreement of ROS2_180 and IMP_180 in panels (a), (c), and (d) does not prove that ROS2 is more accurate. We therefore removed all accuracy-ranking language from this discussion, and we also revised the title to focus only on the computational efficiency.

The revised interpretation is species- and mechanism-dependent numerical sensitivity. For some diagnostics, changing the timestep within IMP produces the larger shift; for others, changing the solver at the 180-s timestep produces the larger shift. A rigorous coupled-model convergence assessment would require additional timestep refinements for both solvers and a sufficiently converged reference. Our supported conclusion is narrower: ROS2 is less expensive than IMP at the same 180-s timestep in the tested configuration, and the one-year integrations show no evidence of numerical instability. Figure 7 is now used to diagnose sensitivity, not to establish superiority in absolute accuracy.

Revised title: “Using a two-stage Rosenbrock solver to increase computational efficiency in the DOE’s Exascale Earth System Model version 1 (E3SMv1)”

Revised manuscript text (lines 304-310): “The sensitivity to timestep and solver formulation is species- and mechanism-dependent. For chemUCI OH and for NO in both mechanisms, IMP_180 and ROS2_180 are relatively close, whereas for trop_strat_mozart_mam4 OH and for NO₂ the solver-dependent separation is more apparent. None of the three fully coupled configurations is an exact reference solution; therefore, Fig. 7 is used to diagnose numerical sensitivity rather than to rank absolute accuracy. Establishing coupled-model convergence for individual species would require additional timestep-refinement simulations for both solvers.”

7. The conclusions section is not concise and is somewhat rambling, seems like AI.

Response: We agree and have substantially rewritten and shortened the Conclusion. Repeated background material and broad forward-looking statements were removed. The revised Conclusion separates three evidence-based findings: (1) at the same 180-s timestep, ROS2 reduces integration cost by about one third relative to IMP on the tested Cori/KNL platform; (2) the one-year integrations remain numerically stable, with small global-mean ozone differences and species/mechanism-dependent spatial differences; and (3) both timestep and solver formulation can affect individual species, so the present simulations do not provide an absolute accuracy ranking. We also explicitly state the two principal limitations: IMP_1800 remains the cheapest configuration because of its larger timestep, and the Cori results do not demonstrate performance on contemporary exascale hardware.

Revised Conclusion excerpt (lines 324-330): “This study implements the two-stage Rosenbrock (ROS2) chemistry integrator in E3SMv1 and evaluates it with two chemically and numerically distinct mechanisms. At a 180-s chemistry timestep, ROS2 requires approximately one third less integration time than the original iterative implicit solver on the tested Cori/KNL CPU configuration. The magnitude of this reduction is configuration dependent and should not be interpreted as a universal scaling factor. IMP_1800 is the least expensive of the three tested configurations because of its tenfold larger timestep; thus, the benefit of ROS2 demonstrated here is specifically relevant when a short chemistry timestep is required.”

Revised Conclusion excerpt (lines 330-334): “Across one-year integrations, ROS2_180 remains numerically stable. Global-mean surface-ozone differences relative to IMP_1800 are small, and the spatial differences in ozone, NO, NO₂, and OH are species- and mechanism-dependent rather than showing a common systematic sign. (The diurnal comparisons further demonstrate that both timestep and solver formulation can affect individual species). Because the fully coupled simulations do not provide an exact reference solution, these comparisons are not used to rank the absolute accuracy of IMP_1800, IMP_180, and ROS2_180.”

Revised Conclusion excerpt (lines 336-336): “The present Cori-based timing results provide a controlled CPU benchmark, not a performance demonstration for current exascale hardware. Additional timestep-refinement experiments and benchmarking on contemporary CPU/GPU platforms will be needed to quantify coupled-model convergence, architecture-dependent speedup, and performance portability for the future higher-resolution E3SM configuration.”

Chien, R.-Y. and Fu, J. S.: pshchien/ROS2-E3SM: ROS2-E3SM v1.1 for GMD submission (v1.1), Zenodo [code], 10.5281/zenodo.19124522, 2026.

Doerfler, D., Austin, B., Cook, B., Deslippe, J., Kandalla, K., and Mendygral, P.: Evaluating the networking characteristics of the Cray XC-40 Intel Knights Landing-based Cori supercomputer at NERSC, *Concurrency and Computation: Practice and Experience*, 30, e4297, 10.1002/cpe.4297, 2018.

IPCC: Annex II: Models, in: *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by: Gutiérrez, J. M., Tréguier, A. M., Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2087–2138, 10.1017/9781009157896.016, 2021.

Reviewer #3

This manuscript reports the first implementation of ROS2, a numerically accurate and computationally efficient integration scheme widely used in regional-scale atmospheric chemistry models, into the E3SMv1 Earth system model. The motivation is to prepare for future high-resolution global Earth system simulations in which atmospheric chemistry is fully coupled with dynamical and radiative processes. Although the same scheme has previously been implemented in another global Earth system model (CESM), reporting its implementation in a different ESM could still be considered within the scope of GMD. However, the manuscript suffers from numerous serious deficiencies in both the presentation and interpretation of the results, as detailed below. In its current form, I do not believe that the manuscript is suitable for publication in GMD. Rejection, or alternatively resubmission after major revision, would be appropriate.

-The manuscript is excessively verbose. Similar statements and phrases are repeated unnecessarily throughout the text.

Response: We thank the reviewer for the detailed and critical assessment. We agree that the earlier manuscript was too repetitive and that several interpretations went beyond what the numerical evidence could support. We therefore revised the manuscript as a whole rather than only editing individual sentences. Repeated motivation and forward-looking statements were removed; the EAM and chemistry configurations were described more explicitly; the definitions and captions for Figs. 1-2 were expanded; and the Results and Conclusion were rewritten to distinguish demonstrated findings from limitations. In particular, the revised manuscript now (i) restricts the performance result to the tested CPU-based E3SMv1/Cori configuration, (ii) acknowledges that IMP_1800 is the cheapest of the three tested configurations because of its larger timestep, (iii) does not infer an absolute accuracy ranking from the diurnal curves, and (iv) describes timestep-versus-solver sensitivity as species dependent. The supplement also contains additional NO, NO₂, and OH spatial diagnostics and the NO_x perturbation experiment. These changes substantially reduce repetition and align the interpretation more closely with the evidence.

-The configurations of the atmospheric model (EAM) and the chemistry schemes (chemUC and MOZART4) are not described clearly enough. As a result, readers cannot properly understand the differences between these two configurations, making it difficult to follow the subsequent discussions.

Response: We agree that the original description did not provide enough information to understand why the two chemistry mechanisms are useful as distinct numerical tests. We expanded the model/method section to state the number of chemical species and reactions, the number treated implicitly, and the resulting Jacobian sparsity for each mechanism. chemUCI contains 60 gas-phase and aerosol species, 53 solved implicitly, 90 reactions, and 262 non-zero Jacobian elements after permutation. trop_strat_mozart_mam4 is substantially larger, with 155 gas-phase and aerosol species, 130 solved implicitly, 365 reactions, and 1232 non-zero Jacobian elements. Tables S1 and S2 provide the complete permuted species lists. We also state explicitly that the mechanisms are not being compared as competing chemical representations; they are used as numerically distinct testbeds for ROS2 across different system sizes and sparsity structures.

Revised manuscript text (lines 121-128): “The first, chemUCI, is based on the UC Irvine chemistry-transport formulation (Prather et al., 2017; Tang and Prather, 2012) and contains 60 gas-phase and aerosol species, of which 53 are solved implicitly. It includes 90 chemical reactions and 262 non-zero Jacobian elements after the solver permutation. The second, trop_strat_mozart_mam4, represents more comprehensive tropospheric and stratospheric chemistry and contains 155 gas-phase and aerosol species, of which 130 are solved implicitly. It includes 365 reactions and 1232 non-zero Jacobian elements. The complete permuted species lists are provided in Tables S1 and S2. These mechanisms are used as numerically distinct testbeds; the objective is not to compare their chemical fidelity or predicted concentrations, but to evaluate ROS2 across different chemical-system sizes and structures.”

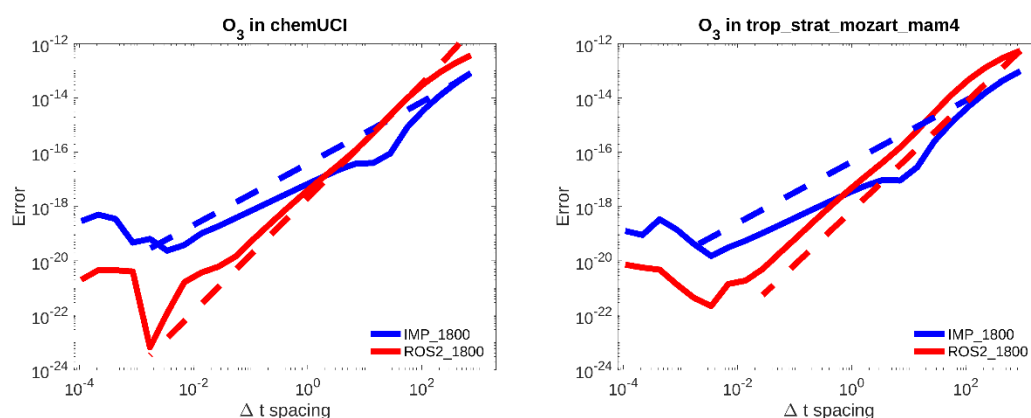
-Figures 1 and 2 are inadequately documented. For example, the axes are not properly labeled, and the figure captions provide insufficient information. Consequently, the reader cannot accurately interpret the contents of these figures.

Response: We agree and have revised both figures and their supporting text. For Fig. 1, we now define the axes mathematically rather than referring loosely to “reactants” and “products.” For $dC/dt = f(C)$, the Jacobian element $J_{ij} = \partial f_i / \partial_j$. The x-axis therefore gives the permuted concentration/state-species index j , and the y-axis gives the permuted tendency/equation-species index i ; each blue point is a non-zero derivative. Tables S1 and S2 map those indices to species. We also clarify that the permutation changes only the ordering of unknowns for sparse matrix solution and does not alter the chemistry.

Revised manuscript text (lines 175-179): “For the chemical system $dc/dt = f(C)$, the Jacobian element $J_{ij} = \partial f_i / \partial C_j$ represents the sensitivity of the chemical tendency of species i to the concentration of species j . Because most species do not interact directly through every reaction pathway, the Jacobian is sparse. The fixed species permutation used by the solver changes only the ordering of the unknowns, not the chemistry. Figure 1 shows the resulting sparsity patterns, and Tables S1 and S2 provide the corresponding permuted species indices.”

Revised Fig. 1 caption: “Figure 1. Sparsity pattern of the chemistry Jacobian for (a) chemUCI (53 implicitly solved species) and (b) trop_strat_mozart_mam4 (130 implicitly solved species). For $J_{ij} = \partial f_i / \partial C_j$, the x-axis gives the permuted concentration/state-species index j and the y-axis gives the permuted tendency/equation-species index i . Each blue point denotes a non-zero Jacobian element. The corresponding species order is provided in Tables S1 and S2.”

Response: Figure 2 was also revised to use conventional log-log axes with timestep spacing Δt on the x-axis. The caption and text now define the plotted quantity as a timestep-refinement error estimate. We explicitly state that if $E(\Delta t) \propto (\Delta t)^p$ over an asymptotic range, the fitted log-log slope is an empirical convergence-rate diagnostic p . Because an analytical reference solution is unavailable for the full stiff chemistry system and the sampled states are not guaranteed to be asymptotic, we no longer present the slopes as direct measurements of the formal order. ROS2 has the steeper empirical timestep-refinement trend in both mechanisms, but the measured slopes are interpreted cautiously.



-Based on the results shown in Figure 3, does it not follow that IMP_1800 is ultimately the best-performing configuration? The figure does not appear to provide a compelling justification for selecting ROS2_180 instead.

Response: This is an important distinction, and we agree that “best-performing” depends on the metric. IMP_1800 is indeed the least expensive of the three tested configurations because it takes one chemistry step for every ten steps taken at 180 s. Figure 3 therefore does not justify claiming that ROS2_180 is universally preferable to IMP_1800, and we have removed that implication. The demonstrated performance advantage of ROS2 is the controlled same-timestep comparison: when both methods use $\Delta t = 180$ s, ROS2 requires approximately 33% less integration time than IMP on the tested platform.

We also removed the implication that ROS2_180 is necessarily more accurate than IMP_1800. The box-model timestep-refinement analysis shows a steeper empirical trend for ROS2, but the fully coupled simulations do not contain an exact reference solution that would permit an absolute accuracy ranking. The revised manuscript therefore states that IMP_1800 is the cheapest tested configuration, while ROS2_180 offers a computational advantage specifically for applications in which a short chemistry timestep is required. The optimal timestep-solver combination for any particular application requires a separate accuracy and coupling analysis. The principal performance result of this study is therefore not that ROS2_180 is faster than IMP_1800, but that ROS2 reduces cost relative to the original implicit solver when both are operated at the same 180s timestep. This makes ROS2 relevant to E3SM configurations in which a short chemistry timestep is required. Determining the optimal timestep-solver combination for a particular application requires a separate accuracy and coupling analysis.

Revised manuscript text (lines 270-277): “Thus, at the same 180s chemistry timestep, ROS2 reduces the measured integration cost by approximately one third relative to the original iterative implicit solver on this platform. The similar percentage reduction for the two mechanisms should not be interpreted as a mechanism-independent scaling law. The relative cost of IMP varies with nonlinear convergence behavior, including additional iterations and timestep subdivision/repetition, whereas ROS2 uses a fixed two-stage update. Consistent with this interpretation, the standalone box-model tests show a broader improvement range of

approximately 20-48%. IMP_1800 remains the least expensive of the three tested configurations because it takes one chemistry step for every ten steps at 180s; therefore, the demonstrated ROS2 performance advantage applies specifically to the same-timestep IMP_180 versus ROS2_180 comparison.”

Revised Conclusion text (lines 325-330): “Thus, at the same 180s chemistry timestep, ROS2 reduces the measured integration cost by approximately one third relative to the original iterative implicit solver on this platform. The similar percentage reduction for the two mechanisms should not be interpreted as a mechanism-independent scaling law. The relative cost of IMP varies with nonlinear convergence behavior, including additional iterations and timestep subdivision/repetition, whereas ROS2 uses a fixed two-stage update. Consistent with this interpretation, the standalone box-model tests show a broader improvement range of approximately 20-48%. IMP_1800 remains the least expensive of the three tested configurations because it takes one chemistry step for every ten steps at 180s; therefore, the demonstrated ROS2 performance advantage applies specifically to the same-timestep IMP_180 versus ROS2_180 comparison.”

-Figures 4 and 5 are described as showing deviations relative to IMP_1800. If so, it is unclear why the blue line representing IMP_1800 itself does not remain at zero throughout. This inconsistency requires clarification.

Response: Thank you for identifying the ambiguity. The plotted quantity is not the instantaneous difference “simulation minus IMP_1800 at the same time,” which would force the IMP_1800 line to zero. Instead, each time series is referenced to the single annual-mean IMP_1800 value. Thus, the same annual mean is subtracted from IMP_1800, IMP_180, and ROS2_180; the IMP_1800 curve itself retains its day-to-day variability around zero. We revised the Results text and figure caption to use the term “anomaly referenced to the annual mean of IMP_1800” consistently, which resolves the apparent inconsistency.

Revised manuscript text (lines 290-292): “The temporal behavior of surface ozone is summarized in Fig. 5 as daily global-mean anomaly referenced to the annual mean of IMP_1800. Specifically, the same annual-mean IMP_1800 value is subtracted from all three time series; consequently, the IMP_1800 curve retains its temporal variability about zero and is not expected to remain identically zero.”

Revised Fig. 5 caption: “Figure 5. Daily global-mean surface-ozone anomaly (ppb), calculated by subtracting the annual-mean IMP_1800 surface-ozone value from each configuration, for (a) chemUCI and (b) trop_strat_mozart_mam4.”

-Figures 6 and 7 indicate that the difference between the blue line (IMP_1800) and the red line (IMP_180), i.e., the effect of changing the timestep, is substantially smaller than the difference between the red line and the yellow line (ROS2_180), i.e., the effect of changing the numerical solver. Nevertheless, the conclusion states that “differences in diurnal behavior are shown to arise primarily from timestep selection rather than from the numerical solver itself.” This statement appears inconsistent with the results presented in the figures. Is the interpretation of these figures correct?

Response: We agree with the reviewer that the original conclusion was inconsistent with Figs. 6 and 7. The statement that diurnal differences arise “primarily from timestep selection” was an overgeneralization and has been removed. The figures instead show species- and mechanism-dependent sensitivity. For ozone, the separation between ROS2_180 and the IMP configurations indicates that solver formulation contributes appreciably to the numerical trajectory. For chemUCI OH and for NO in both mechanisms, IMP_180 and ROS2_180 are relatively close, whereas trop_strat_mozart_mam4 OH and NO₂ show clearer solver-dependent separation. Because none of the three fully coupled simulations is an exact reference solution, these differences cannot be converted into an accuracy ranking.

The revised manuscript therefore makes a narrower statement: both timestep and solver formulation can affect the simulated chemical trajectory, and their relative influence varies by species and mechanism. The timing conclusion is kept separate and refers only to the same-timestep IMP_180 versus ROS2_180 comparison.

Revised manuscript text (lines 304-310): “The sensitivity to timestep and solver formulation is species- and mechanism-dependent. For chemUCI OH and for NO in both mechanisms, IMP_180 and ROS2_180 are relatively close, whereas for trop_strat_mozart_mam4 OH and for NO₂ the solver-dependent separation is more apparent. These differences are expected in a stiff coupled chemical system because individual species sample different chemical timescales and nonlinear pathways. None of the three fully coupled configurations is an exact reference solution; therefore, Fig. 7 is used to diagnose numerical sensitivity rather than to rank absolute accuracy. Establishing coupled-model convergence for individual species would require additional timestep-refinement simulations for both solvers.”

