§2 Response to Reviewer #2

(Nota: Referee comments in black, reply in bold italics)

This work developed an ocean model called swNEMO_v4.0 based on a new-generation Sunway supercomputer and obtained significant modeling performance by sophisticated tuning methods that fully exploited the computing recourses of the new machine. Optimizing methods proposed are based on the architectural features, and thus achieves promising modeling performance. Thread-level communication and mixed-precision arithmetic are very attractive approach today, and this work demonstrates the possibility of applying them into resolving the most complicated scientific project such as ocean model. Firstly, in order to scale the ocean model onto the large-scale and extremely complicated supercomputer, four-level parallel framework are proposed. Sophisticated tuning techniques such as customizable domain decomposition according to the grid feature, are included as well. This enables the capability of fully utilizing the rich computing resources of the new system. The new feature of the system, thread-level RMA communication mechanism, is also wisely used for algorithms such as composite blocking, to further optimize the bandwidth performance. Moreover, mixed-precision optimization is proposed and performed on certain part of the algorithms. With sufficient material and proof to support its feasibility. Significant performance speedup is obtained thanks to these innovations. About 20 million cores are used for the large-scale test, and sustained performance of nearly 2 Petaflops. These innovations are solid, and can be very interesting to domain experts that expect to perform similar work by using the new Sunway supercomputer or other supercomputers with alike architecture. Besides, the work is also very useful for computer scientists like me, to rethink the architecture design for better supporting numerical scientific applications. I have no further comments, but some minor suggestions.

Reply: Thank you very much for your recognition of our work, which is important for us. In fact, this work took more than one year, and we re-wrote almost all the code to port and then to improve the parallel efficiency. Fortunately, we achieved up to 99.29% parallel efficiency with a resolution of 500 m using 27,988,480 cores, which should be the largest parallel scale on the ocean simulation up to now. We are happy it is beneficial to your work and the community.
Minor issues:

1. What is the portability of proposed methods of this work? Eg, to other models, or other applications from different domain.

*Reply: Thank you. Several new optimization approaches proposed, such as a four-level parallel framework with longitude-latitude-depth decomposition, a multi-level mixed-precision optimization method that uses half-, single-, and double-precision, are the methods of general applicability. We test these optimization approaches in the NEMO, but these can be incorporated into other global/regional ocean general circulation models (e.g., MOM, POP, ROMS, etc.). Moreover, the optimizations on the stencil computation can be applied to any model with stencil computations. The above description was added in the Conclusion and Discussion section in the revision.*

2. What is the lesson learned of this work, in terms of architecture design for future supercomputing systems.

*Reply: Thank you. From the view of future ocean simulations, we propose the following aspects that should be paid more attention to.*

*The first is the memory bandwidth. The architecture of the new generation of Sunway processors (SW26010 Pro) adopts a more advanced DDR4 compared with the original SW26010. It not only expands the capacity but also greatly improves the DMA bandwidth of the processor. In this work, we resolved the memory bandwidth problem through fine-grained data reuse technology, thus improving the memory bandwidth utilization rate to approximately 88.7% for DDR4 and paving the way for the ultrahigh scalability of NEMO. However, we noted that the efficiency increases using single precision instead of double precision. As the peak performance of SW26010 Pro are the same between double and single precision, the increased efficiency is mainly from the reduced memory access with changing double precision to single precision. It indicates that the memory bandwidth is still a bottleneck.*

*The second is the half-precision. The finer resolution and more complex processes are the main directions of OGCMs development. Therefore, computational efficiency becomes more and more important. Reduced precision is an effective method for improving efficiency. In the past decade, ECMWF successfully implemented the single-precision in the weather forecast system, which achieves about 40% greater computational efficiency almost without degrading forecast quality. The savings in computational cost mainly come from reduced memory access. The half-precision can not only reduce memory access but also improve the floating-point computing power. Our results also prove that implementing the half-precision in the model can increase the computational efficiency, although we only revised several subroutines of NEMO. We noted that the new architectures of HPC become to support the half-precision, but the support is still incomplete, e.g., the transcendental function cannot be calculated with half-precision in the new generation Sunway.*

*The third is the I/O efficiency. The output data volume becomes larger with
finer resolution. In our work, we tried to store the results with 1 km resolution, but the data volume is more than 65 TB per output, which took more than one-day of clock time. Therefore, the I/O efficiency is still a limitation for finer resolution models.

3. What are the major obstacles that caused the performance loss. What can be done in future to further improve the performance of HPC ocean modeling, from perspectives of both model development and computer design.

Reply: Thank you. We think the major performance losses are from the communications and bandwidth. From the view of software and hardware co-design, the following should be focused on in the future.

The first is the decomposing and load-balance. For the model design, we should find the proper decomposing scheme to fully utilize the computer architecture. Besides the time dimension, solving an ocean general circulation model is a 3-dimension problem, with longitude, latitude, and depth. Usually, only the longitude-latitude domain is decomposed. In our work, driven by the RMA technology, we achieved the longitude-latitude-depth domain decomposed, which enables the better largescale scalability. Meanwhile, keeping a good load-balance is also important for scalability. For the computer design, the RMA technology is a good example, which enables the longitude-latitude-depth domain decomposing. In other words, the high communication bandwidth between different cores or nodes will help to largescale scalability.

The second is communications. With the increasing processes used for model simulation, the ratio of communications time to computational time will become higher. For the model design, the first thing is to avoid the global operator, such as ALLREDUCE, and BCAST, which will take more time with increasing the processes. Otherwise, it will be the crucial bottleneck. Meanwhile, we also should pack the exchanged data between different processes as much as possible. For the computer design, the low latency will help in saving the communications time.

The third is reduced-precision. The results of our work demonstrate that there is a great potential to save computational time by incorporating the mixed double-, single-, and half-precision into the model. For the model design, we should understand the minimum computational precision requirements essential for successful ocean simulations, and then revise or develop arithmetic. For the computer design, the support for half-precision should be considered in future.

Overall, the above are only several examples for further improving the performance of ocean modeling from perspectives of model development and computer design. Furthermore, other aspects such as I/O efficiency, and the trade-off between precision and energy consumption should also be considered. And it should be noted that these suggestions are from different aspects of the model and computer development and need to be considered based on the software and hardware co-design ideology.

The above description was added in the Conclusion and Discussion section in the revision.