**General Comments:**

This paper proposes the innovative application of Multi-Level Monte Carlo (MLMC) and Multi-Level Data Assimilation (MLDA) techniques within the realm of simplified ocean model based on the shallow-water equations, aiming to enhance computational efficiency while maintaining or improving forecast accuracy. The authors effectively highlight the theoretical foundations of MLMC and MLDA, compare these methods to traditional single-level approaches, and discuss the integration of GPU-accelerated frameworks to address the computational demands of high-resolution simulations. Despite its contributions, the manuscript suffers from structural and clarity issues that need addressing. Consequently, my recommendation is for publication following major revisions.

**Major Comments:**

1. The paper lacks necessary explanations for several terms. For instance, there is hardly any description of MLMC, MLDA, and their single-level counterparts. The differences between MLDA and MLMC are not clearly stated. Discussions on GPUs and CPUs appear suddenly without prior introduction or context. Additionally, the structure of sections and subsections is complex, making some parts difficult to read, especially the introduction section. In the equations, the use of superscripts and subscripts is prevalent, but their application seems cluttered in places, indicating room for improvement.

2. One of the primary objectives of this paper is the proposal of a new method to improve computational efficiency. However, there is insufficient comparison of computation time and computational costs between this new method and traditional methods. It would be beneficial to identify which aspects of the traditional methods incur significant computational costs and how much time the proposed method takes in comparison. Comparing computational costs and error scores between single-level and multi-level approaches could verify whether this method is truly effective and practical. Furthermore, discussing potential issues and differences when applying this method to actual ocean models, in addition to idealized experiments, would be beneficial.

**Detailed Comments:**

1. Section 1: What is MLMC? Please provide a brief explanation.

2. L16: 'By harnessing the cost-effectiveness of low-fidelity simulations within the ensemble'—please elaborate on this statement with specific details.

3. Section 1: Can you explain the benefits of performing MLDA, compared to traditional methods?

4. L26: Since 'search-and-rescue (SAR)' hardly appears in the text, defining the abbreviation 'SAR' may not be necessary unless it is used more extensively.

5. Section 1 is divided into three parts, making the content quite difficult to understand. Rather than dividing it, reconsidering and organizing the content in a more orderly manner into a single section would likely make it easier for readers to comprehend.

6. Section 1.1: A brief mention of the relationship between data assimilation and MLMC could improve the connection between the first and second paragraphs.

7. L74: What is MLEnKF, and how does it differ from the traditional EnKF?

8. L84: What does 'robust data assimilation' mean?

9. L87: What is a GPU? What is meant by a GPU-accelerated framework, and how do data assimilation and MLMC relate to GPUs?

10. L88: What are 'sparse observations'?

11. L117: 'In sequential data assimilation, the state is ...'— this statement seems specific to 3D data assimilation and not applicable to 4D methods like smoothers. The same applies to line 152.

12. L123: You are using **H** (in bold font), indicating a linear observation operator. Is it possible to use *H* for nonlinear observations in practice?

13. L120: What is 'The model for the observation'?

14. L130: Please explain the meaning of 'single-level'.

15. L166: What ensemble update method do you use? Perturbed observation method? Square root filter? Or something else?

16. Section 2.3: Before entering this section, could you briefly state which parts of the computation cost are problematic in normal Single-level Monte Carlo EnKF, and how using Multi-level Monte Carlo aims to avoid these issues?

17. L242: Considering the frequent use of superscript throughout, it might be better to avoid the expression K^(ML).

18. L282: 'Assessment scores may also be used to evaluate the quality of the ensemble-based representation, and these tasks require various kinds of functions.' I'm having trouble understanding this sentence.

19. L290: In the long sentences in Section 3.1, the most crucial statement appears to be 'we choose the ensemble size tailored for the Kalman gain estimation in the MLEnKF,' yet no concrete method is discussed. Please provide a detailed explanation.

20. Section 3.1.1: I don't understand the necessity of this section. What is the relation between Multi-level data assimilation and GPUs or CPUs? How does this section's content relate to the equations used in Multi-level data assimilation? Could you explain which specific parts of the equations in this paper contribute to the difference in computational costs?

21. The structure of Section 3.1 (and Section 3) is very confusing. What role does Section 3.1.2 play within it?

22. Would it be better to add the content of Section 3.1.3 at the end of Section 3.1?

23. L359: Please explain "level-local formulations."

24. Section 3.2: Do you use vertical localization? If not, why is it unnecessary?

25. Section 3.3: Please discuss further the impact of negative eigenvalues in real applications.

26. Section 3: Is inflation used?

27. L459: What is "a relaxation factor"? Why is inflation or relaxation necessary?

28. L465: Rather than solely relying on visual comparisons, would converting the truth shown for 10d in Fig3 to the appropriate grid size and comparing them with the results in Fig4, including calculating scores, provide a more objective evaluation?

29. L487-489: I'm having trouble understanding this passage.

30. Figure 6: In hu and hv, the results of single-level experiments and multi-level experiments are similar. Can you provide theoretical insights into what this means and what it implies?

31. Figure 6: Why does the difference between ERR and STD become significant after around day 8 for all variables?

32. Figure 6: The scores are not stable during this experimental period. Is the experimental period too short?

33. L532: "It is reasonable to do so for the coarsest level of case (A) with 275 members." Really? Why do you think that?

34. L537: "Once the dynamics becomes more turbulent around day 6 to 7," Why is that? Is there a reason for this setting? What's the reasoning behind this setting? Sorry if it was mentioned somewhere and I missed it.

35. L542: "As discussed before, MLDA differs from MLMC" - which part are you referring to? Also, a brief explanation here again might make it easier for readers to understand.

36. L544: Could you briefly explain what the value 'V^l/v^l t' represents and its significance in the context of your study?

37. Figure 8: Why does the relative variance of the variable eta in the experiment with the darkest blue line gradually decrease after day 10?

38. L564: What does "For the true observation value y" mean?

39. Figure 9: By plotting the results of conventional methods over the MLEnKF results, it might be possible to compare the two methods and further highlight the effectiveness of MLEnKF.

40. L624: What does 'outlier' mean?

41. Figure 10: The results of the three methods are very similar, making it hard to see the differences in this figure. You have Figure 11, so is Figure 10 necessary? As you mention in L626, especially in this experiment with a short spin-up and experimental period, it heavily depends on the initial values. So, I think the results of Figure 10 highly depends on initial ensembles.

42. L631: "We notice that the calibration curves for all methods are very close, but the spread using multi-level methods is a little bit bigger." This expression is ambiguous.

43. Section 5: One of the main reasons for using a new method in this study was thought to be improving computational efficiency. How about comparing computation time and computational cost in this setting and experiment? Discussing whether this method is truly practical by comparing computational load and error scores between SL and ML could be helpful.

44. L654: "we have discussed various practical challenges that naturally arise when MLDA is implemented." Could you please explain this part in detail?

45. L660: "the devil was in the details" Could you please explain this part in detail?

46. Section 6: Please describe the potential issues when applying your method to actual models and observations.