Authors' response to Referee #1's comments for "Multilevel Monte Carlo methods for ensemble variational data assimilation "

Preprint egusphere-2024-3628

M. Destouches, P. Mycek, S. Gürol, A. T. Weaver,
S. Gratton and E. Simon

March 10, 2025

The authors discussed the background error covariance estimation using (weighted) multi-level Monte Carlo (wMLMC) method in variational data assimilation (DA). The authors discussed several practical considerations when MLMC is sued to estimate a covariance matrix: 1) the mean squared error and variance of a covariance MLMC estimator; 2) computational budget allocation; 3) localisation and positive definiteness of the estimated covariance matrix. The MLMC estimator and the performance of corresponding 3DEnVar is investigated by a two-dimensional two-layer quasi-geostrophic channel model after 12 hour forecast from an initial ensemble without data assimilation cycles. The paper is well-written and is worth publication.

We thank the referee for their comments and suggestions. In the following, we discuss the major and minor comments, and explain how we have updated or will update the manuscript to address them.

Other modifications, that were not directly asked by the referee, have also been applied to the manuscript. The reason for this is detailed at the end of this document.

Major comments:

1. *The Experimental setting section may be benefit from a figure to illustrate the model setup.*

Thank you for this comment, we have added such a figure.

2. Current results are all built on a single dynamical snapshot of the model. Optionally, is it possible to build a stronger case by running a long deterministic trajectory of the model, and a few select different time step with very different features of dynamics as initial condition to generate ensemble with 12 hour forecast such that the computational cost does not drastically increase?

The dynamics of the quasi-geostrophic model are not very complex, and the dynamical features observed are very similar across time. As such, we do not expect the conclusions of the paper would be altered by selecting another date to run the experiments.

On another, more practical level, it would indeed be costly to perform the full experiment again on several other situations. There are three type of results in our manuscript: a) the theoretical estimation of the MSE of a $\mathbf{B}$ estimator that can be reached with a multilevel approach (section 5.1), b) the empirical estimation of the MSE reduction that is reached on a column of the $\mathbf{B}$ matrix when actually building these estimators (section 5.2), and c) the impact on a single analysis (section 5.4).

The results of sections 5.2 and 5.4 would be especially cumbersome to reproduce, as they require to run more than 100, 000 forecasts for each date (200 realizations of multilevel estimators using about 600 different forecasts each). However, we can easily reproduce on other dates the experiment of section 5.1 (theoretical MSE reduction on $\mathbf{B}$), as it only requires to run 400 forecasts per dates.

We reproduced the results of section 5.1 for 4 other dates, selected 10, 20, 30 and 40 days later than the date studied in our manuscript. The seed of the random number generator was different for each date. The 4 background states are shown in Fig. 1 in this document. The key figures of section 5.1 are reproduced in Fig. 2 and Fig. 3. As summarized in Table1, the variance reductions are of similar order of magnitude than for the case studied in our paper. We do not plan to reproduce all these figures in the paper, but we will mention that the results of this section do not change significantly with the dynamical situation.

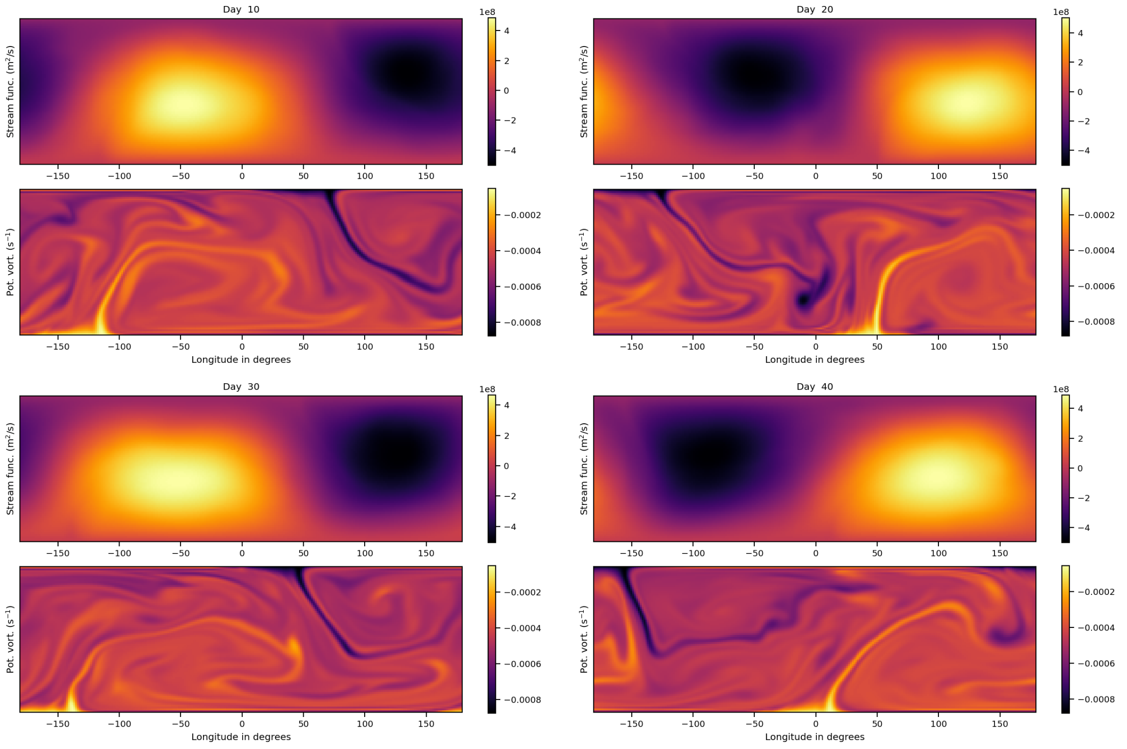

Figure 1: Background state at different times, 10 to 40 days later than the background state used in the paper.

simulation day	MLMLC	wMLMC
0	63%	66%
10	69%	72%
20	65%	69%
30	66%	69%
40	68%	71%

Table 1: Variance reduction achieved by MLMC and weigted MLMC with the same cost as a reference 20-sample MC estimator. First line (day 0) is the result shown in the paper.

3. When the ensemble member allocation is tuned based on Eq. (14) and (16), do we expect that the $a^{(k)}$ and $b^{(k)}$, or $C^{(k)}$ change significantly due to the flow-dependency of ensemble forecasting?

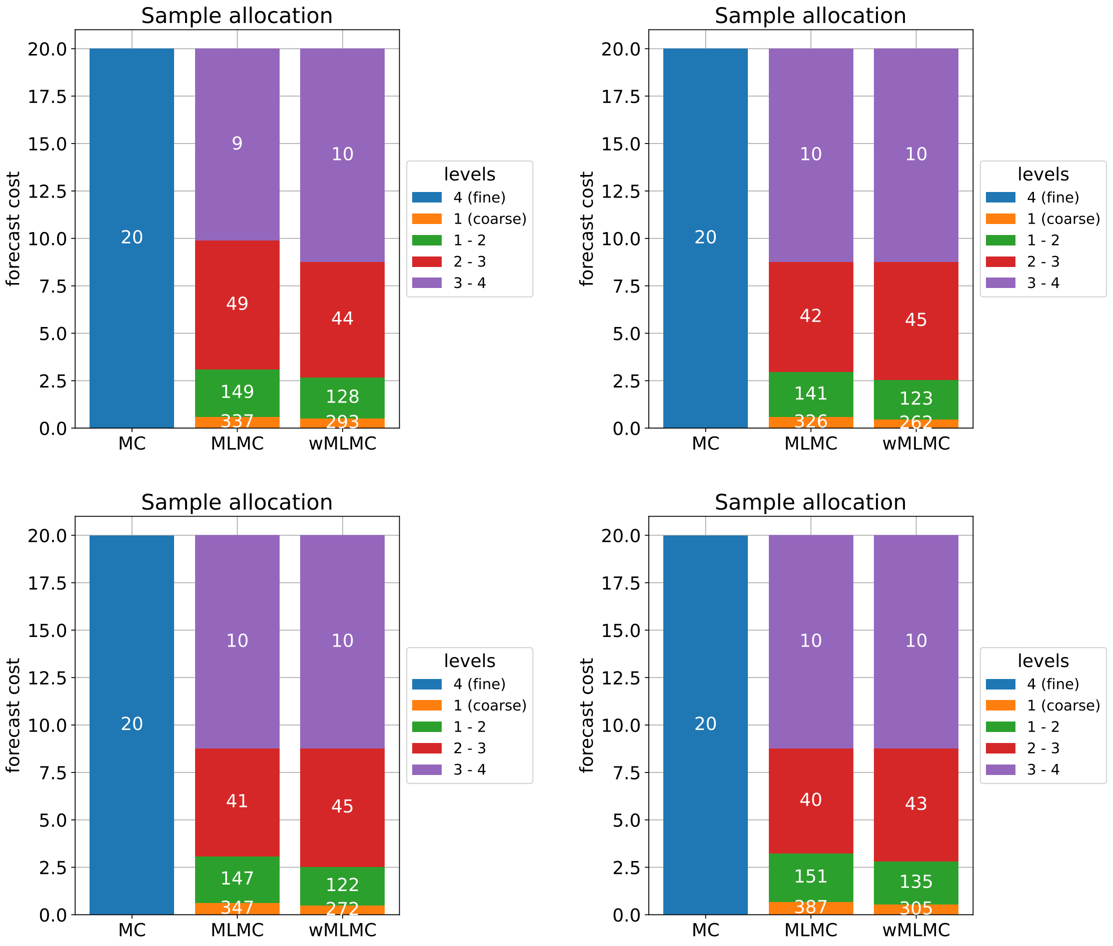

Figure 2: Optimal sample allocations associated to the background states in Fig. 1. Corresponds to manuscript Fig. 2.

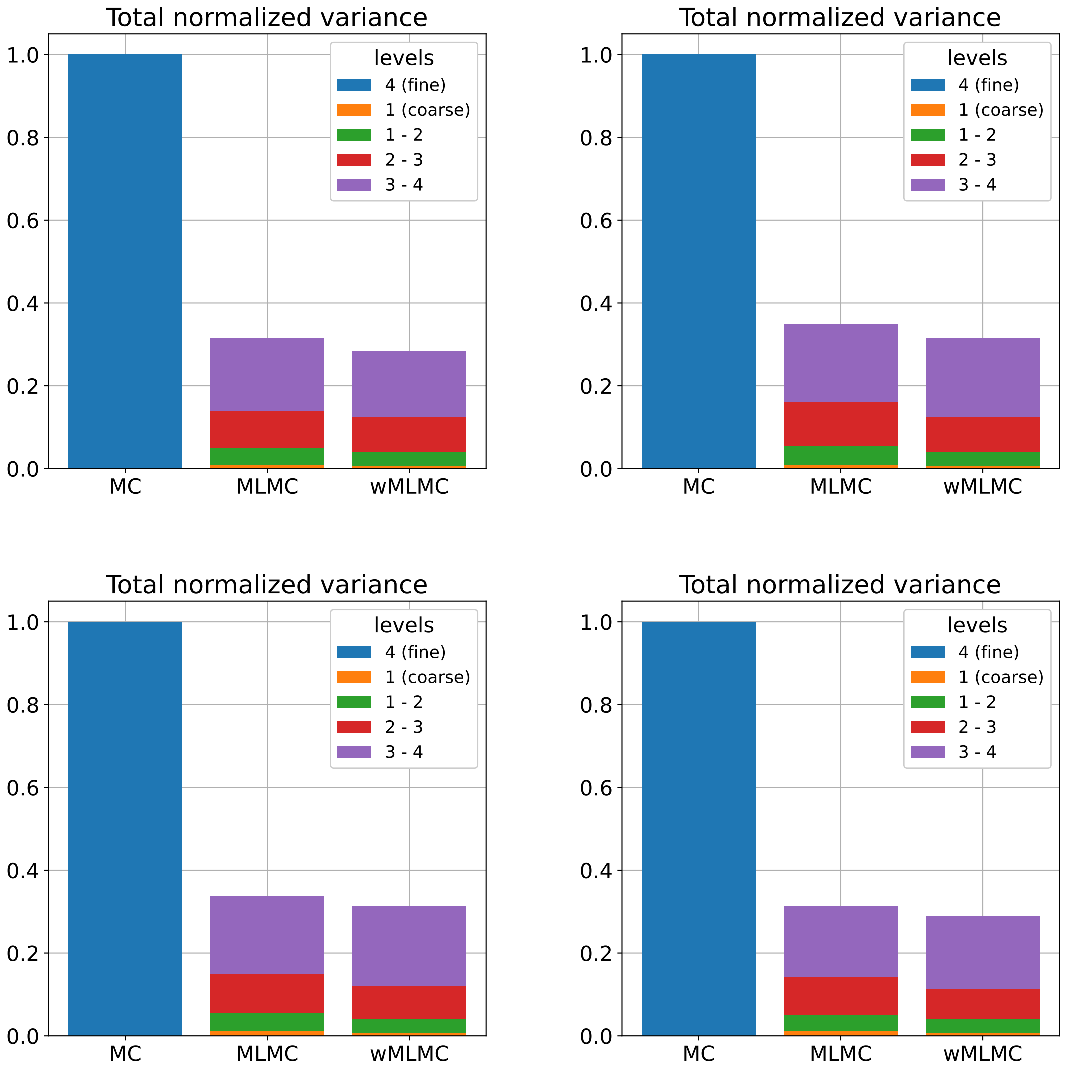

Figure 3: Theoretical variance of covariance matrix estimators associated to the background states in Fig. 1. Corresponds to manuscript Fig. 3.

The authors suggest to change the ensemble allocation less frequently (L573 - L575). Do we expect that MSE of estimated B matrix at least as accurate as high-resolution B matrix, i.e. the MC method?

This is an interesting question. The $a^{(k)}$, $b^{(k)}$ and $C^{(k)}$ coefficients are related to the spatially-averaged variance of the differences between estimators on different fidelity levels. How much they would change with time is not assessed in this manuscript, especially as it would strongly depend on the application. For instance, for a large domain where many small and independent dynamical features are represented, one snapshot of the model may already be representative of a full simulation run. In this context, the space average in these coefficients would imply that they should not vary much with time. Conversely, if the full system can switch from a global attractor to another one, and if the low-fidelity simulators fail to represent correctly one of these attractors, then the inter-level coefficients could change significantly with time. For numerical weather prediction applications for instance, we expect the time-variability of the coefficients to be larger in a limited-area model than in a global model.

If the ensemble member allocation is not updated at every cycle, the allocation becomes sub-optimal, and the MSE of the resulting multilevel estimator becomes larger than the MSE of the *optimal* multilevel estimator. There would be no guarantee in this context that the MSE of the multilevel **B** estimator would be smaller than the MSE of the high-resolution Monte Carlo **B**. More generally, we would like to stress that there is generally no such guarantee even with an optimal allocation. The variance of an optimal multilevel estimator with several fidelity levels can be larger than the variance of the same-cost MC estimator. This could be the case for instance if the coarse-fidelity levels were only weakly correlated to the high-fidelity level, and if their computational costs were only slightly cheaper. This is underlined in the paper as well, in Sect. 3.1.

4. Using the B matrix from MLMC and MC shows, the results show that the analysis has limited improvements. Could this be related to the smooth streamfunction in the QG model? Would we expect significant differences for other fields, e.g., PV?

We have not studied this, as the standard way to solve the data assimilation problem for this model is to use streamfunction as the control variable, due to its simpler background error structures and statistics. Performing data assimilation in the potential vorticity space would also be a challenge for positioning errors.

If we performed the experiments using potential vorticity as a control variable, we would expect the performance to be degraded, even on the MSE of the **B** matrix estimator. This can be understood with a scale-separation approach. The large length-scales of the fine-simulation can usually be described accurately by the low-fidelity models, which have a large inter-level correlation in this range of the spectrum. Conversely the coarsest fidelity models are not able to represent fine length-scales, with very small or zero correlations in the high-frequency end of the spectrum. For streamfunction, most of the signal is on the large scales, which ensures a satisfying representations on coarser models. Potential vorticity spectra are much less steep, with much more energy on the fine length-scales. This variable is not suited to a multilevel approach, at least not if the low-fidelity levels are based on coarser space discretizations.

More grounded reasoning around the spectral analysis of multilevel Monte Carlo can be found in the recent preprint by Briant et al. (2023).

Minor comments:

L5: "...affordable We investigate..." → "...affordable. We investigate..."

L57: "...the Ensemble Kalman Filter(..." → "the ensemble Kalman filters ("

Thanks for spotting these typos. We have corrected them.

L87: "the composition operator" → "a composition operator"

We rather modified the sentence into "the function composition operator".

L159: Perhaps a set should be represented with curly brackets?

Yes, indeed. This has been modified.

L160: "...stochastic inputs are all independent..." → "... number of stochastic inputs are all independent..."

We think the suggested modification would change the meaning of the sentence. We have rephrased this sentence to make it clearer:

> The total $\sum_{k=1}^{L} N^{(k)}$ stochastic inputs are all independent and identically distributed.

has become:

> There are thus $\sum_{k=1}^{L} N^{(k)}$ stochastic inputs in total, all independent and identically distributed.

L221: The author states that "...related to small fourth-order moments of the correction terms, and so to strong correlations between stochastically-coupled simulations...". Does this mean that adjacent level of model must yield similar outcome? How close should be these levels? Does this also justify the 0 value for level 0?

Yes, models from adjacent fidelity levels should yield similar outcomes for the multilevel approach to be effective. In practice, in our experiments, the inter-level correlations derived from the space average of inter-level covariances ranged from 0.77 (between the two levels of highest fidelity) to 0.94 (between the two levels of lowest fidelity).

In general, there is no easy rule on how close these fidelity levels should be. One has to go through the process of defining the inter-level correlation coefficients, defining the cost model and finding the optimal member allocation to know what variance reduction can be expected. In the very simple case of the weighted 2-level MLMC estimator of a scalar mean, it can be shown from the MLBLUE formalism (Schaden and Ullmann, 2020) that the variance of the multilevel estimator is:

$$\mathrm{Var}[\widehat{\mu}] = \frac{N^{(\mathrm{low})}(1-\rho^2) + N^{(\mathrm{high})}}{N^{(\mathrm{high})}\left(N^{(\mathrm{low})} + N^{(\mathrm{high})}\right)}, \tag{1}$$

where $\widehat{\mu}$ is the multilevel estimator with optimal weights, $N^{(\mathrm{low})}$ is the number of samples on the low-fidelity level, $N^{(\mathrm{high})}$ is the number of samples from the high-fidelity level, ρ is the inter-level correlation, and the variance of the random variable is assumed to be 1 without loss of generality. In the limit of infinitely large cost ratio between high and low simulations, i.e., in the limit of infinitely large $N^{(\mathrm{low})}$, we find a variance of $(1-\rho^2)/N^{(\mathrm{high})}$. Compared to the variance $1/N^{(\mathrm{high})}$ of the same-cost MC estimator, this gives a relative variance reduction of ρ^2. So a 0.5 correlation for instance would give 25% variance reduction. Such simple considerations are of little use in practice, given that the cost-ratio is never infinite, more than 2-levels can be used, smaller reduction should be expected for the estimation of larger-order moments that the mean, and uniform scalar weights are suboptimal for a non-scalar problem.

L270: "corrections term" → "correction term"

Done.

L318: "1:3 ratio..." → "1:3 ratio between the width and length of the domain..."

Done.

L443: "apply them to a Dirac impulse" reads as if the estimator is used to estimate the covariance of a Dirac impulse, which is not the case, I think. This also means that Fig. 4 and 5 are covariance of one grid point with the entire domain.

This is what happens, unless I am missing something in the question. Fig. 4 and Fig. 5 are indeed covariances of the entire domain with respect to a single grid point. The general variance reduction estimated in section 5.1 applies to the full covariance matrix estimator, but we can only show the impact on three-dimensional columns of this estimator. Computing the actual impact on a full covariance matrix would have required about 4×10^4 more memory usage (the number of columns in the covariance matrix), which is prohibitive.

L477: Here, can the authors briefly explain why the cost is proportial to the ensemble member instead of grid size?

The cost scales with both the number of ensemble members and the number of grid point in each ensemble member. We will add a citation to a paper explaining why the localization cost scales with the ensemble size (appendix B of Buehner, 2005).

L534: "...is to remain..." → "...remains..."

Here, "is to remain" was supposed to convey the idea of a constraint provided by the user and that must be met. The text has been rephrased to make it clearer:

> We are thus led to conclude that while randomization approaches may be of interest for offline diagnostics, they are not a viable solution if the cost of applying **B** is to remain comparable to the cost of applying a standard localized ensemble **B**.

has been rephrased into:

> We are thus led to conclude that while randomization approaches may be of interest for offline diagnostics, they are not a viable solution to the negative eigenvalue problem, unless we allow the cost of applying **B** to increase significantly compared to the cost of applying a standard localized ensemble **B**.

L553: What is the B norm?

For an SPD matrix $\mathbf{B}$, the $\mathbf{B}$-norm of a vector $\mathbf{u}$ is the norm defined by $\|\mathbf{u}\|^2 = \mathbf{u}^\intercal\mathbf{B}\mathbf{u}$. The definition is missing in our manuscript, thank you for spotting this. We will correct it.

L641: alpha should be bold?

Yes, indeed. This is now corrected.

Other modifications

In addition to the changes suggested by the referee, a few modifications have been applied to the manuscript (visible in the latest version of the manuscript when it will be submitted). When replying to the referee's comments, we realized that Figure 9 had not been obtained with the early-stopping method described in the article. It was a remnant of another solution that we had explored, where we relied on backtracking on the residual norm rather than early-stopping criteria. We had moved away from this backtracking approach as there was little ground for this, but we accidentally kept all figures and data in the paper. Although this has no impact on the conclusions, it did alter some data and figures. This will be corrected in the next submission of the manuscript. We reproduced all the results to be consistent with the early-stopping approach presented in the text:

- There is no clearly visible difference in Figures 1 to 6.
- The conclusions of the paper are not affected.
- The weights of the multilevel estimator given in the paper were not the correct ones, which has been corrected (this is unrelated to the choice of backtracking or early-stopping approach)
- The wMLMC localization parameters tuned for the early-stopping approach are different than those tuned for the backtracking approach.
- As a result, the spectrum of the localized wMLMC covariance matrix estimate in Fig. 7 is different, with less negative eigenvalues than in the previous backtracking approach.
- As the experiments have been fully reproduced in a different computing environment compared to the first submission, the random samples used to build the MLMLC estimator are different. This explains minor differences in the spectra of the unlocalized wMLMC covariance matrix estimate (first negative eigenvalue at index 8 rather than 11, with amplitude 11% rather than 9%, but with no change in the global proportion of negative eigenvalues).
- In Fig. 9, the relative error reduction compared to the best achievable reduction is increased from 2% on average (and 10% on median) to 11% on average (and 13% on median). The message of the figure is not affected.

References

Briant, Jérémy et al. (2023). *A filtered multilevel Monte Carlo method for estimating the expectation of discretized random fields.* DOI: 10.48550/arXiv.2311.06069. arXiv: 2311.06069 [math.NA].

Buehner, Mark (2005). "Ensemble-derived stationary and flow-dependent background-error covariances: Evaluation in a quasi-operational NWP setting". In: *Quarterly Journal of the Royal Meteorological Society* 131.607, pp. 1013–1043. DOI: 10.1256/qj.04.15.

Schaden, Daniel and Elisabeth Ullmann (Jan. 2020). "On Multilevel Best Linear Unbiased Estimators". In: *SIAM/ASA Journal on Uncertainty Quantification* 8.2, pp. 601–635. ISSN: 2166-2525. DOI: 10.1137/19M1263534. URL: https://epubs.siam.org/doi/10.1137/19M1263534.