An Offline Framework for High-dimensional Ensemble Kalman Filters to Reduce the Time-to-solution

Yongjun Zheng[1], Clément Albergel[1], Simon Munier[1], Bertrand Bonan[1], and Jean-Christophe Calvet[1]

[1]CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

December 5, 2019

Dear reviewer,

Thank you very much for your comments. The followings in blue are our responses. The texts in red are those we revised or added in our new manuscript.

Best regards,

Yongjun ZHENG on behalf of the authors

Reviewer #1:

The paper "An offline Framework for High-dimensional Ensemble Kalman Filters to Teduce the Time-to-solution" by Zheng et al. is, as far as the reviwer is aware, the first empirical study of the maximized wall-clock efficiency of both online and offline approaches to ensemble Kalman filterting as used in an operational context.

Thank you for your time and effort to read and comment our paper carefully.

1 General Comments

The paper utilizes the LESTKF, yet the only mathematical description is of a general ETKF-like filter. A subsection on how the ESTKF, and a section on localization (and implementation challanges with localization therein) would greatly help the reader.

The main purpose of this study is to compare the time-to-solution of offline and online ensemble Kalman filters. Even though the results presented in this paper are obtained by using the LESTKF, the conclusions can be generalized to other EnKF schemes as well. For example, the conclusions are still held when using LETKF scheme (results are not shown in the paper). The LESTKF is adopted in this paper because the PDAF documents recommend it and we thought the LESTKF may be well tested in PDAF. The following sentence is added near line 18 of page 20 in the new manuscript for referring readers to those papers on the full description (Neger et al, 2012a) of the ESTKF and its localizations (Nerger et al, 2006): We refer readers to the

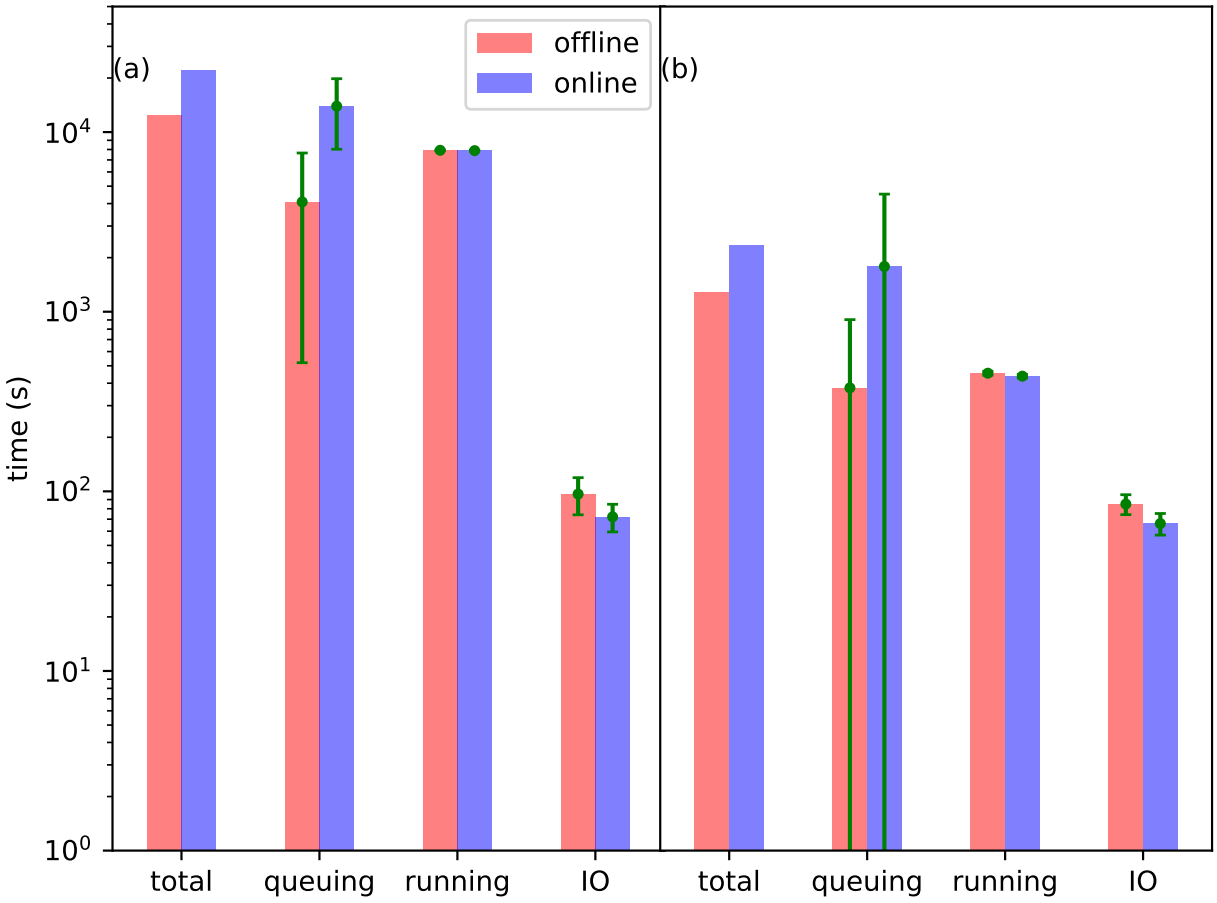

Figure 1: The average total time, queuing time, running time, and IO time of the offline (red bars) and online (blue bars) EnKFs, The panel (a) and (b) are for the medium and large scale problems, respectively. The green line indicates the corresponding standard deviation.

paper of Nerger et al. (2012a) for a full description of the ESTKF, and the paper of Nerger et al. (2006) for the domain and observation localizations using in LESTKF. This might be a proper way for the readers who are interested in how the ESTKF and its localizations work but also for the paper being succinctly around its subject.

In the experimental design section (4.2.1 in this draft), the choice of randomly selecting the observation points is concerning, a more uniform, or atleast reproducible approach would instill more confidence in the methodology.

Thank you very much for pointing out the reproducibility. Following your suggection, we did an experiment by having one observation of every ten grid points regularly instead of selecting 10% of the grid points randomly, the results show no big differences in both the analysis performance and the time-to-solution (see Figure 1). Actually, we believe the randomly selecting method does give us a higher confidence since the observations are usually irregularly scattered over the domain.

Again in the experimental design, if the aim is to reproduce operational conditions, why is more realistic data, say generated from some model like SPEEDY or WRF not used?

This is a really good question. In fact, the initial purpose of this study is to test the possibility

of substituting an online EnKF for an offline one. First, building an online EnKF demands more substantial time and effort than building an offline one, we did not know whether it is worth investigating time to do so. Second, an online EnKF does reduce the IO time, but it was not clear whether an online EnKF is faster than an offline EnKF in terms of time-to-solution. Therefore, it is more reasonable to address these questions with a very simple and idealized model before seriously investigating time to build a new system. This is the main reason this study did not use a complex model.

Again in the experimental design section, operationally we consider the amount of observations as being three orders of magnitude lower than the state space, yet the choice made in this paper is only one order of magnitude lower. This might bias the results in favor of the offline approach.

We agree with the reviewer that the observations used in our paper are denser than those in nowaday operational practice. The large amount of observations certainly increases the running time of EnKF analysis, but the IO time is barely affected. As shown by Figure 10 in the paper, the dominant time is the queuing time, especially for a large scale problem. The queuing time can be stable in a fixed loaded condition but the running time decreases if less observations are used, that is, the ratio of queuing time to running time will become biger . This evidently further convinces that an offline EnKF wins the favor over its online counterpart.

Equation 17 on page 22 should have 14 in the denominator as the mean is estimated, and not exactly known. This fact is used earlier in the EnKF description.

Thank you very much for pointing out this typo. It has been corrected as $\sigma_x = \sqrt{\frac{\sum_{j=1}^{j=15}(t_{j,x}-\bar{t}_x)^2}{14}}$ in the new manuscript.

In the conclusions (section 5 in this draft) maybe don't use bullet points, and try to more fluidly outline the main results? Though this is not that much of an issue.

Thank you for the suggestion. This paragraph has been re-organized slightly without using bullet points in page 25 of the new manuscript as follows:

In summary, the proposed parallel IO algorithm can drastically reduce the IO time for reading or writing multiple files with an identical structure. The tuning parameters of a stripe count and a stripe size should be consistent, and high values of these two parameters usually allow high concurrent IO operations and low competitions which significantly reduce the IO time. Using the proposed parallel IO algorithm, the running times of both offline and online EnKFs for high-dimensional problems are almost the same since the IO time only accounts for a small fraction which further decreases as the increase of the scale of the problem. This implies that the proposed parallel IO algorithm is very scalable. On the contrary, in a low-loaded supercomputer, the queuing time might be equal to or less than the running time, thus the offline EnKF is at least as fast as, if not faster than, the online EnKF in terms of the time-to-solution because the offline mode requires less simultaneously available nodes and more easily and quickly obtains the requested nodes to reduce the queuing time than the online mode. But in a high-loaded supercomputer, the queuing time is usually several times larger than the

running time, thus the offline EnKF is substantially faster than the online EnKF in terms of time-to-solution because the queuing time is dominant in such a circumstance. Therefore, The loaded condition of a supercomputer varies greatly which justifies the dynamically running job scheme of an offline EnKF.

2 Technical Corrections

- p1l17, 'for intermittent'.
- p3l32, 'demands substantial'
- Figure 7 is of a particularly low DPI, and looks jarring compared to the other figures. Perhaps a flat 2D figure could convey the same information more clearly?
- p22l11, maybe use 'longer' instead of 'larger' in reference to time?

Thank you very much for the very careful reading. In the new manuscript, all are corrected ("for the intermittent data assimilation" $\longrightarrow$ "for intermittent data assimilation", "demands a substantial time and effort" $\longrightarrow$ "demands a substantial time and effort", and "larger" $\longrightarrow$ "longer") as suggested except for the Figure 7. Figure 7 seems look good in our side. But we will try to improve it if the quality is not enough for publication.