[Reviewer's comments are inserted in regular font and responses are in blue.]

Reviewer 1

General comments:

In this manuscript, the authors analyze the differences between the identical and fraternal twin approaches for Observing System Simulation Experiments, using the ROMS model with an Ensemble Kalman Filter (EnKF) in the Gulf of Mexico. They find that the impact assessment differs in both cases, as the identical twin approach tends to over- estimate the error reduction from satellite observations and underestimate the error reduction from vertical profiles of temperature and salinity, compared to the fraternal twin approach.

This manuscript is concise and well written. It provides convincing results to illustrate, for the first time, the differences between identical and fraternal twin experiments using an ocean model. I thus recommend publication, however with minor revision, in order for the authors to better present their results with respect to the reference work by Halliwell et al. (2014, 2015, 2017). Please see my specific comments below.

We thank the reviewer for the positive evaluation and constructive comments, which we respond to in more detail below.

Specific comments

Comment 1 (C1): I find that the approach followed by the authors has a methodological limitation in that, in their fraternal twin approach, they use a data-assimilative model, and that this model has a coarser resolution for the "Truth" than the assimilative simulation. This contradicts the recommendations by Atlas et al. (1997) and Halliwell et al. (2014, 2015, 2017), the work of whom is the reference for the present study. Halliwell et al. (2014), based on Atlas et al. (1997), stated (p. 106): "The established procedures to design and per- form OSSEs documented in the atmospheric OSSE literature are summarized by Atlas (1997). The [Nature Run (i.e. the Truth)] is a long unconstrained simulation performed at high resolution using a state-of-the-art general circulation model." Here, the Truth simulation is not unconstrained, as it is derived from a global operational model that assimilates observations, and it cannot be considered to be at high resolution, since its resolution is ~9 km (1/12°), larger than the data-assimilative ROMS model resolution (5 km). It is not clear why the authors did not follow the recommendations from Halliwell et al. (2014) and Atlas et al. (1997). They should mention this limitation in their approach in the manuscript, in the methodological section, and include it in their discussion.

Response (R): The reviewer is right that the "truth" in our fraternal twin did not follow exactly the definition from Halliwell et al. (2014), but it follows one of the alternative approaches Halliwell et al. described. Following this and the other reviewer's comment, we have decided to rename the 'fraternal twin' as 'non-identical twin' in the revised version.

The 'non-identical twin' in our definition is specifically mentioned as a viable approach by Halliwell et al. (2014, p. 107 first paragraph) when they state:

"…These requirements can be substantially realized **by using two different model types** and running the forecast model at lower resolution to introduce additional truncation errors. Alternatively, the chosen forecast model can be a different configuration of the same model type used for the NR as long as different physical parameterizations, truncation errors, and boundary condition errors are appropriately introduced. This latter method is referred to as the ''fraternal twin'' approach, and it is used for the ocean OSSE system presented herein."

In Halliwell's interpretation the fraternal twin definition is limited to cases where the "truth" and forecast runs are significantly different configurations of the same model type. We had interpreted this more broadly in the previous version of the manuscript to include twin experiments like ours where the "truth" and forecast runs are from two different model types (i.e., 'truth' from HYCOM while 'forecasts runs' from ROMS). We intend to add the following text in the revision (new text is underlined here):

"If the chosen "truth" and forecast runs are from same model but with perturbed initial, forcing or boundary conditions, the method is referred to as 'identical twin' approach; if two different model types or significantly different configurations of the same model type (e.g., using different physical parameterizations and/or spatial resolution) are used, the method is referred to as the 'non-identical twin' approach. We note that the approach where the same model type is employed but with sufficiently different configurations is conventionally termed fraternal twin (Halliwell et al., 2014), but here a different model type is used in the 'non-identical twin'."

We would like to point out that the primary objective of this study is to illustrate how the commonly used identical twin, which employs the same model but with perturbed initial, forcing or boundary conditions for the 'truth' and forecast runs, could lead to biased assessment for the DA system, in contrast to a 'non-identical' twin approach. The essence of the latter approach is to obtain sufficiently different configurations for the 'truth' and forecast runs, and doesn't necessarily have to be "from the same model type and with 'truth' unconstrained and at higher resolution" but can also come from two different model types as in this study (see Halliwell et al. 2014).

C2: Second, I find that the introduction and the discussion sections give a misleading account of the results and recommendations exposed in Halliwell et al. (2014). In the introduction, at the bottom of p. 4, the authors suggest that Halliwell et al. (2014) recommend investigating the error growth between the various models and observations (l. 61-66), or ("alternatively" l. 67) performing a set of comparable OSSEs and Observing System Experiments (OSEs) using the same data-assimilative model and actual observations. This is misleading, as both steps are recommended by Halliwell et al. (2014).

R: Yes, we do acknowledge that both steps are recommended by Halliwell et al. (2014), while unfortunately we didn't make this clear enough in the previous version of the manuscript. We intend to revise the relevant text as below (changed text is underlined here):

"They suggested that the model for the forecast run should be configured differently enough from that for the "truth" run so that the rate of error growth between them has the same magnitude as that between state-of-the-art ocean models and the true ocean. They also suggested comparing the assimilation impact in the twin framework with that in a realistic configuration; if a similar impact is obtained in both twin and realistic configurations, the twin DA framework can be considered appropriate for assessing assimilation impact and conducting OSSEs."

C3: I find that the account of the recommendations from Halliwell et al. (2014) is more problematic in the discussion. The authors write (l. 355-359): "[Halliwell et al. (2014)] main criterion is that the rate of error growth between simulated and observed states must be similar between the twin framework and reality. However, we found a similar rate of error growth in Sea Surface Height (SSH) in both twin experiments and in reality, yet the identical twin proved problematic. Thus, assessing error growth in just one ocean property appears to have been insufficient." Not only is the comparison of the error growth one of two main criterions exposed by Halliwell et al. (2014) (see previous paragraph), but Halliwell et al. (2014) never suggested to compare the error budget in only one ocean property, which is what the authors suggest here. Indeed, Halliwell et al. (2014) compared the error budget in SSH, Sea Surface Temperature (SST) and Sea Surface Salinity. The authors' account of Halliwell et al. (2014) is mis- leading and they should re-write that part of their discussion to avoid confusion. I also suggest that the authors present the error growth in SST, in addition to SSH, so that their evaluation of the error budget is not performed in one ocean property only. The last sentence of that paragraph (l. 359-362) should also be modified, as Halliwell et al. (2014) recommended a comparison between comparable OSSEs and OSEs as part of the evaluation of the OSSE system, in addition to (and not alternatively to) a comparison of the error rate.

R: Taken. We intend to delete the problematic sentence and modify the relevant text as below (changed text is underlined here):

"Halliwell et al. (2014)'s set of design criteria and evaluation procedures for ocean OSSEs serves as guidance for designing twin experiments for a data-assimilative system. Their main criteria include both that the rate of error growth between simulated and observed states must be similar between the twin framework and reality and that the assimilation impact in the twin framework should be comparable to that of a realistic configuration assimilating actual observations. We found a similar rate of error growth in SSH in both twin experiments and in reality, and the impact of assimilation in the non-identical twin experiment is found to be very similar to that in a realistic assimilation configuration presented in Yu (2018). Thus our direct comparisons of identical versus non-identical twin not only lend support to the recommendation of using the non-identical over the identical twin approach, but also hint that assessing error growth in just one ocean property is insufficient. Additional criteria, such as a comparative assessment of skill between twin and realistic assimilation configurations as described in Halliwell et al (2014), are needed to obtain a more credible impact assessment from the twin framework."

Yu, L.: Improved prediction of the effects of anthropogenic stressors in the Gulf of Mexico through regional-scale numerical modelling and data assimilation, Ph.D. thesis, Dalhousie University, Canada, http://hdl.handle.net/10222/75005, 2018.

Regarding the error growth for other ocean property like SST, we find it not as useful as that of SSH in assessing the two twin setups because the model SST is largely dependent on the imposed surface air temperature. This has also been pointed out in Halliwell et al. (2014): "Consideration was given to comparing satellite-derived SST maps, but SST is dominated by the annual cycle and model SST variability tends to follow the imposed surface air temperature, limiting the usefulness of this comparison."

Below are minor specific comments:

C4: - l. 31: Moore et al. (2019) is not in the reference list.

R: Will be added in the revised version.

C5: - l. 32-37: It is also possible to keep some of the observations from the pool of observations to be assimilated, to be used for independent assessment of the performance of a data-assimilative simulation. However, this leads to a reduction in the quantity of data that are assimilated. The authors might mention that approach here.

R: Taken. We will revise the sentence as below (added text is underlined here):

"But in practice, the value of such an assessment is limited because it either does not consider independent observations (i.e., observations that have not been assimilated into the system) or has to sacrifice the quantity of data for assimilation while reserving some for independent assessment."

C6: - l. 124-126: Can the authors be more precise about how the model has a tendency to overestimate the Loop Current northward penetration?

R: We found the free run overestimated the Loop Current northward penetration during our specific simulation period (April-September 2010) based on the comparisons with the satellite observed Sea level anomaly and Argo profiles of temperature and salinity files (Yu 2018). However, we didn't find the model has a persistent tendency of overestimating the Loop Current intrusion in different years of simulation. We will modify the sentence and refer to Yu (2018) in the revised version:

"Initial model-data comparisons showed that the model has skill in statistically simulating the main features of the LC intrusion with a slight overestimation of its northward penetration during the simulation period (Yu, 2018)."

Yu, L.: Improved prediction of the effects of anthropogenic stressors in the Gulf of Mexico

through regional-scale numerical modelling and data assimilation, Ph.D. thesis, Dalhousie University, Canada, http://hdl.handle.net/10222/75005, 2018.

C7: - l. 128-145: That part describes the EnKF. Can the authors briefly mention what the specificities of the DEnKF are?

R: Taken. We will add a bit more explanatory text on DEnKF such as below:

"Different from the traditional EnKF (Burgers et al., 1998) which requires perturbing observations to obtain an analysis error covariance consistent with that given by the Kalman Filter, the DEnKF updates the ensemble mean using the analysis equation (2) and ensemble anomalies with the same equation but half the Kalman gain $\boldsymbol{K}$ without perturbing observations, and is hence termed 'deterministic'."

C8: - l. 152-153: Altimetry data are available daily along satellite tracks with a repetitive period of ~10 days for the reference altimetry missions, and the SST data are available daily with higher resolution than ¼°, in the absence of cloud coverage. The assimilation of weekly maps of SSH and SST at ¼° resolution is thus a choice of the authors for their experiments, which they should make clear and explain the reason for.

R: Yes, we acknowledge that there are various satellite products with varying spatial and temporal resolution, and different DA applications have adopted different products. We will add some explanatory text for our choice as below:

"SSH and SST are sampled weekly at every fifth horizontal grid point to yield a spatial resolution of ~1/4° as such assimilation time window or spatial resolution has been adopted in previous realistic DA applications (e.g., weekly gridded product of SSH used in Moore et al., 2011, Song et al., 2016b, and weekly gridded product of SST in Hoteit et al. 2013)."

C9: - l. 204: Is the MAD equal to the RMS Error? If yes, I suggest the authors to mention it. If not, I recommend that the authors provide the equation to estimate the MAD.

R: The MAD does not equal the RMS error. We will add following explanatory text for MAD:

"Model-data misfit is quantified by computing the Mean Absolute Deviations (MAD), i.e. the average of the absolute deviations, of model simulations from the "truth" for the open Gulf (defined as regions deeper than 300 m). That is, MAD = $\frac{1}{N}\sum_{i=1}^{N}|model_i - truth_i|$, where N is the number of data pairs for comparison and i denotes the i-th element."

C10: - l. 248-250: How do the authors explain such a difference in salinity MAD difference in the northeastern shelf of the Gulf, whereas there are no observations assimilated in the area?

R: The salinity MAD difference on the northeast shelf is due to the assimilation of SST observations which cover the region deeper than 10 m.

C11: - l. 264: Although it is very common, in the scientific literature, to use parentheses to present results from two different datasets or experiments, this way of presenting results is generally confusing and should be avoided, as there is really no reason to use parentheses that way. I recommend the authors to read Robock (2010, https://eos.org/opinions/parentheses-are-are-not-for-references- and-clarification-saving-space).

R: Thanks for recommending the reference. We will thoroughly examine the manuscript and avoid the improper use of parentheses.

C12: - l. 324-325: Do the authors have an idea as to why "the additional information content in the subsurface observations within the identical twin system is much smaller than that for the fraternal twin"? I suggest that the authors discuss this and offer some possible explanation.

R: We will add some explanation as below (added text is underlined here):

"It follows that, the additional information content in the subsurface observations (i.e., profiles) within the identical twin system is much smaller than that for the fraternal twin. We attribute it to the lack of intrinsic difference (e.g., physical model parameterizations, spatial resolution) in the identical twin 'truth' and forecast model runs that makes it easier to correct the subsurface model fields with assimilating SSH and SST alone in identical twin; a closer agreement of the model and 'truth' subsurface fields subsequently reduces the additional information content of subsurface observations sampled from 'truth'."

C13: - l. 335: Where does this 70 km resolution come from? Is it from the spatial distance between vertical profiles in experiment F3/I3? This should be clarified in the text.

R: Yes, this is from the spatial distance between vertical profiles used in experiment F3/I3. We will clarify it in the revised version.

C14: - l. 342-345: Do the authors have an idea as to why the experiment I1 leads to improvement in resolving the small scale processes on the shelf break, in addition to the large scale in the deep Gulf, whereas such improvement on the shelf break was not seen in experiment F1? I suggest that the authors discuss this and offer some possible explanation.

R: The lack of clear improvement on small scale process on the shelf break in non-identical experiment F1 was due to the not sufficiently fine resolution of observations and model in resolving those processes. This was explained in Lines 334-342: "The assimilation of SSH, SST and additional temperature and salinity profiles (spatial distance between profiles in experiment F3/I3 is ~70km) in our fraternal twin experiments provides limited constraints on the small-scale circulation features in this region. This is consistent with Wang et al. (2003) who found that assimilating SSH and SST could not accurately resolve smaller-scale eddies in the DeSoto Canyon region near the DwH site. It has been suggested previously that higher-resolution localized observations (Lin et al., 2007; Jacobs et al., 2014; Carrier et al., 2014; Berta et al., 2015; Muscarella et al., 2015) and even finer model resolution (< 5 km, Ledwell et al., 2016) are needed to better constrain these submesoscale features."

By contrast, the identical twin I1 leads to substantial improvement on small-scale process on the shelf break. This is because in the identical twin setup, the differences between the "truth" and forecast model runs are purely from external forcing (i.e., wind forcing, initial and boundary conditions) while the intrinsic model structures (e.g., subgrid-scale parameterizations, horizontal and vertical resolution) of the two are identical. Therefore, the identical experiment I1 can well reproduce the subgrid-scale processes in the 'truth' once the large-scale processes are corrected through the assimilation of SSH and SST. We will add some explanations as below (added text is underlined here):

"In contrast to the fraternal twin, the identical twin I1, which assimilates only SSH and SST, yields remarkable improvements not only in the mesoscale circulation dominating the open GOM but also the smaller-scale processes prevailing along the shelf breaks, including the DeSoto Canyon region where the spill site is located. This is largely because in the identical twin setup, the intrinsic model structures (e.g., subgrid-scale parameterizations, horizontal and vertical resolution) for the "truth" and forecast model runs are identical so that an improvement in large-scale processes due to assimilation of SSH and SST can readily translate to an improvement in the simulated subgrid-scale processes."

C15: - Supporting Information and Figures: I do not understand why some figures are in a Supporting Information section and others are in the main manuscript. All the figures from the Supporting Information have a comparable role as the figures in the main manuscript. I strongly recommend including all the figures in the main manuscript and get rid of the Supporting Information section.

R: Taken. We will include all figures from the Supporting Information in the main manuscript.