[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 renamed 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 added the following text in the revision (new text is underlined here):

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

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 revised 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 deleted the problematic sentence and modified 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 1) that the rate of error growth between simulated and observed states must be similar between the twin framework and reality, and 2) 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: 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 revised 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 reduce the quantity of data used for assimilation when 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 modified 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 added a bit more explanatory text on DEnKF in the revision 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 added 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 added 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 $i=1,\ldots,N$ and N is the number of data pairs."

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 have thoroughly examined the manuscript and avoided 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 added 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 non-identical twin. We attribute this to the lack of intrinsic difference in the identical twin (e.g., physical model parameterizations, spatial resolution) between the 'truth' and forecast model runs making it easier to correct the subsurface model fields by assimilating SSH and SST alone. This close agreement of subsurface fields between the forecast model and 'truth' necessarily reduces the additional information content of subsurface observations during assimilation."

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 (now renamed as N3/I3). We clarified it in the revised version as below:

"The assimilation of SSH, SST and additional temperature and salinity profiles (spatial distance between profiles in experiment N3 is ~70km) in our non-identical twin experiments provides limited constraints on the small-scale circulation features in this region."

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 347-355: "The assimilation of SSH, SST and additional temperature and salinity profiles (spatial distance between profiles in experiment N3 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. <u>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.</u>"

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 now include all figures from the Supporting Information in the main text of revised manuscript.

Reviewer 2

The revision is good, and the paper will be ready for publication after minor revision.

I still have a major issue with terminology. The experiments where observations are sampled from the same model are correctly labeled as identical twin experiments. However, the experiments where observations are sampled from HYCOM and assimilated into ROMS are not fraternal twin experiments. The term 'fraternal twin' needs to be dropped in describing these experiments.

The category 'fraternal twin' is reserved for the case where observations are sampled from one model, and then assimilated into the same model that is set up with a substantially different configuration. This is possible with HYCOM, for example, be- cause this model contains multiple choices of numerical schemes and subgrid-scale parameterizations. These different choices can enable the version used to generate observations behave much differently from the version used to assimilate the observations. Even though the same model is used, the two different configurations can be set up to substantially behave like different ocean models.

R: We followed the suggestion and replaced all occurrences of 'fraternal twin' with 'non-identical twin' in the revised manuscript. The following explanatory text was also added: (new text is underlined here):

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

Evaluation of non-identical versus identical twin approaches for observation impact assessments: An EnKF-based ocean assimilation application for the Gulf of Mexico

Liuqian Yu[1,2], Katja Fennel[1], Bin Wang[1], Arnaud Laurent[1], Keith R. Thompson[1] and Lynn K. Shay[3]

[1]Department of Oceanography, Dalhousie University, Halifax, Nova Scotia, Canada

[2]Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong

[3]Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, Florida, USA

E-mail: liuqianyu@ust.hk

Deleted: fraternal

Abstract

Assessments of ocean data assimilation (DA) systems and observing system design experiments typically rely on identical or non-identical twin experiments. The identical twin approach has been recognized as yielding biased impact assessments in atmospheric predictions but these shortcomings are not sufficiently appreciated for oceanic DA applications. Here we present the first direct comparison of the non-identical and identical twin approach in an ocean DA application. We assess the assimilation impact for both approaches in a DA system for the Gulf of Mexico that uses the Ensemble Kalman Filter. Our comparisons show that, despite a reasonable error growth rate in both approaches, the identical twin produces a biased skill assessment overestimating the improvement from assimilating sea surface height and sea surface temperature observations while underestimating the value of assimilating temperature and salinity profiles. Such biases can lead to an undervaluation of some observing assets (in this case profilers) and thus misguided distribution of observing system investments.

Deleted: fraternal

Deleted: fraternal

1. Introduction

Ocean data assimilation (DA), i.e. the incorporation of observations into ocean models to obtain the best possible estimate of the ocean state, has become standard practice for improving the accuracy of model predictions and reanalyses. Benefiting from the rapid expansion of ocean observing platforms and advances in computing power, various ocean DA applications at both regional and global scales have been developed in support of ocean hindcasts, nowcasts and forecasts (e.g., see recent reviews in Moore et al. 2019 and Fennel et al. 2019). Necessarily the credibility of a DA system demands rigorous validation. It is straightforward to assess the assimilation impact (i.e. the differences between ocean state estimates from a model run with and without assimilation), where a better fit of the model state to observations following assimilation might be considered as positive. 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 reduce the quantity of data used for assimilation when reserving some for independent assessment.

An alternative assessment approach is to conduct twin experiments (e.g., Anderson et al., 1996; Halliwell et al., 2014). The essential steps of a twin experiment are to 1) predefine a simulation as the "truth", 2) sample synthetic observations from this "truth", 3) assimilate these observations into a different simulation referred to as the forecast run, and 4) assess the skill of this assimilative run against a non-assimilative ("free") run using independent observations sampled from the "truth". If the chosen "truth" and forecast runs are from same model implementation but with perturbed initial, forcing or boundary conditions, the method is referred to as 'identical twin' approach. If two different model

Deleted: Conventionally,

Deleted: i

types are used, we refer to the method as the 'non-identical twin' approach. We note that the intermediate approach where the same model type is employed but with sufficiently different configurations (e.g., different physical parameterizations and/or spatial resolution) is conventionally termed fraternal twin (Halliwell et al., 2014). In addition to validating DA systems, twin experiments are used for Observing System Simulation Experiments (OSSEs) that evaluate the impact of different ocean observing system designs on predictive skill (e.g., Oke and O'Kane 2011; Halliwell et al. 2015, 2017). Ideally, the "truth" and forecast simulations in the twin system used for the OSSE should be from two different models, i.e. they should be non-identical twins.

The identical twin approach has been more commonly used in oceanic DA applications (e.g., Counillon and Bertino, 2009b; Simon and Bertino, 2009; Srinivasan et al., 2011; Song et al, 2016a; Yu et al., 2018) although it is well known from atmospheric OSSEs that this approach provides biased impact assessments when the error growth rate between the "truth" and forecast runs is insufficient (e.g., Arnold and Dey 1986; Atlas 1997; Hoffman and Atlas 2016). This fact is not yet sufficiently recognized in applications of ocean OSSEs and skill assessments of oceanic DA systems (Halliwell et al., 2014). To avoid the potential bias in impact assesments associated with identical twin experiments, Halliwell et al. (2014) proposed to apply a criterion that has long been used in realistic atmospheric OSSEs. 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

Deleted: or significantly different configurations of the same model type (e.g., using different physical parameterizations and/or spatial resolution)

Deleted: the method is

Deleted: red

Deleted: fraternal

Deleted: ;

Deleted: if the same model but with perturbed initial, forcing or boundary conditions is used, the method is referred to as 'identical twin' approach (Halliwell et al., 2014).

Deleted: Friedrichs 2001,

Deleted: In practice this

Deleted: is difficult to assess. ¶
Alternatively, Halliwell et al. (2014) suggested to

Deleted: e

Deleted: fraternal

and realistic configurations, the twin DA framework can be considered appropriate for assessing assimilation impact and conducting OSSEs. Fraternal OSSEs have proven instructive for evaluating the assimilation impact of different observing platforms in the Gulf of Mexico (Halliwell et al., 2015) and North Atlantic (Halliwell et al., 2017).

However, a direct comparison of fraternal or non-identical and identical twin approaches has not yet been conducted for an ocean application, to the best of our knowledge. Motivated by this, we use an ocean DA system for the Gulf of Mexico (GOM) to compare and contrast the non-identical and identical twin approaches in an assimilation impact assessment. The rationale for choosing the GOM as our testbed is that the non-deterministic aspects of the circulation in the GOM, including the northward penetration of Loop Current (LC) intrusions and the associated eddy shedding, require DA for accurately hindcasting/forecasting the circulation. The need for accurate nowcasts and predictions was particularily acute during the 2010 Deepwater Horizon (DwH) oil spill. Previous data assimilation applications in the GOM have focussed primarily on improvements of the surface current fields observeable from satellite or drifters but did not examine the assimilation impact on subsurface flow fields. As the DwH oil spill has shown, knowledge of model skill in simulating the subsurface circulation is also important. Utilizing twin experiments, we aim to examine the assimilation impact on the subsurface circulation.

Toward this objective we implement an advanced ensemble DA technique, the Ensemble Kalman Filter (EnKF), for a high-resolution (horizontal resolution of 5 km) model covering the entire GOM. The EnKF utilizes flow-dependent background error covariances in contrast to the time-invariant covariance in optimal interpolation (OI-) or variational-based DA systems that have previously been used in GOM (e.g., Counillon and

Deleted: fraternal

Deleted: (Halliwell et al. 2014)

Deleted: Such f

Deleted: fraternal

Deleted: We use 'Fraternal twin' to refer to the case where two different models are used.

Bertino 2009a, 2009b; Jacobs et al. 2014). By rigorously assessing the skill of the EnKF-based assimilative model (with an emphasis on the subsurface fields) through non-identical and identical twin experiments and OSSEs, we demonstrate how the identical twin approach yields misleading conclusions in this practical application. We also address whether an improved skill in reproducing the surface dynamics of the LC and associated eddies translates into improved skill in simulating the subsurface circulation.

Deleted: fraternal

2. Model description and experimental setup

2.1 The physical model

The model is configured using the Regional Ocean Modelling System (Haidvogel et al., 2008; ROMS, http://myroms.org) for the GOM (Fig. 1a). It has a horizontal resolution of 5 km and 36 terrain-following vertical layers with higher resolution near the surface and bottom. Vertical turbulent mixing is parameterized using the Mellor and Yamada (1982) Level 2.5 closure scheme, and bottom friction is specified using a quadratic drag formulation. The model utilizes a third-order accurate, non-oscillatory advection scheme for tracers (HSIMT, Wu and Zhu, 2010), which is mass-conservative and positive-definite with low dissipation and no overshooting, and is forced with the atmospheric forcing fields from the European Centre for Medium-Range Weather Forecasts (ECMWF) (http://apps.ecmwf.int/datasets/). River input is prescribed as in Xue et al. (2013), with daily runoff from US Geological Survey for rivers inside the US and long-term climatological estimates for rivers in Mexico and Cuba. The model is one-way nested inside the 1/12° data-assimilative global Hybrid Coordinate Ocean Model (HYCOM) (Chassignet et al., 2009). Tidal forcing is neglected because tides are small in the GOM.

Previous studies have highlighted two important aspects for model skill in the GOM, a sufficiently high horizontal resolution for representing the mesoscale dynamics (e.g., Chassignet et al., 2005) and an accurate representation of the LC inflow through the Yucatan Strait (e.g., Oey, 2003). Our model meets the two requirements. The 5-km horizontal resolution is sufficient to resolve mesoscale processes (the baroclinic Rossby radius is 30 to 40 km in the central GOM, see, Oey et al., 2005). And our ROMS model is nested in a data-assimilative HYCOM model which simulates an accurate structure of the LC and its eddies. 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).

Deleted: tendency to overestimating

2.2 Experimental framework

The deterministic formulation of the EnKF (DEnKF), first introduced by Sakov and Oke (2008), was implemented in the GOM model. The DEnKF has been successfully used in previous ocean assimilation applications (e.g., Simon et al., 2015; Jones et al., 2016; Yu et al., 2018). The algorithm consists of sequential forecast and analysis steps, where the model ensemble is propagated forward in time during the forecast step and updated with available observations using the Kalman Filter analysis equation during the analysis step. The analysis equation is given as:

$$\boldsymbol{x}^a = \boldsymbol{x}^f + \boldsymbol{K}(\boldsymbol{d} - \boldsymbol{H}\boldsymbol{x}^f), \quad (2)$$

where $\boldsymbol{x}$ is the $n \times 1$ model state estimate vector (n is the number of model state variables at all grid points), the superscripts a and f represent the analysis and the forecast estimates, respectively, $\boldsymbol{d}$ is the $m \times 1$ vector of observations (m is the number of available

observations), $\boldsymbol{H}$ is the linear $m \times n$ measurement operator mapping the model state onto the observations, and $\boldsymbol{K}$ is the $n \times m$ Kalman gain matrix, given as

$$\boldsymbol{K} = \boldsymbol{P}^f \boldsymbol{H}^T (\boldsymbol{H}\boldsymbol{P}^f \boldsymbol{H}^T + \boldsymbol{R})^{-1}, \quad (3)$$

where $\boldsymbol{P}^f$ is the $n \times n$ forecast error covariance matrix (approximated by the forecast ensemble), $\boldsymbol{R}$ is the $m \times m$ observation error covariance, and T denotes the matrix transpose. 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'. Details on the DEnKF derivation and implementation can be found in Sakov and Oke (2008).

2.2.1 Non-identical twin experiments

Deleted: *Fraternal*

In non-identical twin experiments, the "truth" is generated by interpolating the daily outputs of the 1/12° data-assimilative global HYCOM (Chassignet et al., 2009) onto the ROMS model grid. Synthetic observations are sampled from the "truth", including SSH, SST, and temperature and salinity profiles. Typical Gaussian observation errors of $N(0, 2$ cm) for SSH, $N(0, 0.3$ °C) for temperature (both SST and temperature profiles), and $N(0, 0.01)$ for salinity are added to the sampled data. 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). SSH in regions shallower than 300 m is not used for assimilation because dynamics in shelf areas where wind and buoyancy forcing

Deleted: fraternal

dominate could substantially deviate from the geostrophic state weakening the correlation between SSH and subsurface temperature and salinity fields. For SST, only those in regions shallower than 10 m are excluded. Importantly when preparing the synthetic SSH observations, the mean dynamic topography (MDT) of the HYCOM "truth" run had to be removed from the sampled SSH data and the MDT of the ROMS model had to be added. The MDTs of the HYCOM and ROMS models were obtained by averaging their respective daily SSH outputs from 2010 to 2016.

Temperature and salinity profiles were sampled with two different sampling schemes (see locations in Fig. 1a, b). The first scheme adopts the sampling dates and locations used in the survey described in Shay et al. (2011). The key features of this scheme are that the sampling is centered on the LC region, the majority (363 out of 472) of temperature profiles are limited to the upper 400 m, and very few (34) salinity profiles were collected. In the second scheme coverage was extended such that temperature and salinity profiles are sampled simultaneously over the entire central GOM down to 1000 m depth on 23 instead of 9 dates.

A non-assimilative run, subsequently referred to as the free run, is initialized on 1 April 2010 from the global HYCOM and compared with the data-assimilative runs to evaluate the impact of the assimilation.

In the DA experiments, 20-member ensembles are started from different initial conditions and forced by perturbed boundary conditions and wind fields. The initial conditions were created by using three-dimensional (3D) fields from daily HYCOM outputs within a 20-day window centered on the initialization date of 1 April 2010. The boundary conditions were generated by applying a time lag of up to +/-10 days to the boundary

condition (i.e., the first member's boundary conditions are 10 days ahead) following Counillon and Bertino (2009b). The perturbed wind fields were created by first conducting an empirical orthogonal function (EOF) decomposition of the wind field and then adding perturbations from the mixture of the first 4 EOF modes to the wind field, where the four perturbation modes were multiplied with zero-mean unit-variance random numbers and a scale factor of 0.5 similar to Thacker et al. (2012) and Li et al. (2016).

We used an ensemble of 20 as it was the largest size feasible given the computing resources available to us and found this to work well in our application. The same ensemble size has also been used in previous studies (e.g., Hu et al., 2012; Mattern et al., 2013). Distance-based localization with an influence radius of 50 km was applied as described in Evensen (2003) to prevent the potential negative effects of spurious correlations between distant grid points. An inflation factor of 1.05 was applied to the ensemble anomalies inflating the ensemble spread around its mean at every assimilation step as introduced by Anderson and Anderson (1999). This accounts for the potential underestimation of the forecast error covariance due to the small ensemble size. The choice of localization radius and inflation factor are based on initial tests and takes into account that the baroclinic Rossby radius in the central GOM is 30 to 40 km (Oey et al., 2005) to avoid choosing too small localization radius value.

Observations are assimilated weekly from 2 April to 3 September 2010 updating the 3D temperature and salinity fields. On each assimilation date, the observations (regardless of observation types) are assimilated simultaneously in one single step. After the last assimilation step on 3 September 2010, the ensemble is run without any data assimilation for 4 more weeks. Three assimilation experiments (referred to as N1, N2 and N3) are

Deleted: F1

Deleted: F2

Deleted: F3

conducted. N1 assimilates weekly SSH and SST, while N2 and N3 assimilate the temperature and salinity profiles following the two sampling schemes described earlier (Fig. 1a, b) in addition to SSH and SST. 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 i=1,…,$N$ and N is the number of data pairs. For ensemble assimilation runs, the forecast ensemble mean at assimilation steps is used for calculating the MAD.

Deleted: F1
Deleted: F2
Deleted: F3

2.2.2 Identical twin experiments

The identical twin experiments have a similar setup as the non-identical twin experiments except that the "truth" is not taken from HYCOM but generated from a ROMS simulation that differs from the free run only in its initial and boundary conditions and wind forcing. The "truth" run is started on 1 April 2010 from an initial state from an earlier ROMS simulation, and is forced with boundary conditions that are lagging behind those of the free run by 14 days and wind fields reconstructed from the first 10 EOFs of the realistic ECMWF wind. Since the same model architecture is used in free and reference runs for the identical twin, there is no need to correct MDT when sampling SSH observations.

Deleted: fraternal

Similar to the non-identical twin setup, three assimilation experiments are conducted in the identical twin framework (I1, I2 and I3) that assimilate the same combinations of observations as in N1, N2 and N3.

Deleted: fraternal
Deleted: F1
Deleted: F2
Deleted: F3

3. Results

3.1. Assessment of the non-identical and identical twin experiment setup

Deleted: fraternal

We first examine the credibility of the non-identical and identical twin setups by comparing the error growth rates in SSH between the free run and the "truth" for both twins (Fig. 2). The non-identical twin has a slightly higher error growth rate (0.048 cm/day) than the identical twin (0.040 cm/day), but both are of similar magnitude to that between the free run and real observations (0.042 cm/day). This meets the requirement suggested by Halliwell et al. (2014) that the errors between the free run and the "truth" should grow at a similar rate as errors that develop between state-of-the-art ocean models and the true ocean. The comparison in Fig. 3 also shows that differences between the "truth" and free runs in SSH and subsurface salinity fields are obvious and qualitatively comparable between the non-identical and identical twin experiments. This satisfies the other requirements suggested in Halliwell et al. (2014), namely that the free run is able to reproduce the main features of the simulated phenomenon (i.e. the LC intrusion) with some realism, and that there are sufficient differences between the free and "truth" runs for the assimilation method to correct.

Deleted: fraternal

Deleted: fraternal

Deleted: the

Deleted: fraternal

3.2. Impact of assimilation in non-identical twin experiments

Deleted: fraternal

Temporally and spatially averaged MADs between the non-identical twin assimilation runs and the free run are summarized in Table 1 (temporal evolution is shown in Fig. 4). Assimilating SSH and SST in N1 significantly reduces the MADs of SSH (by 51%), temperature (by 29%) and velocity fields (by 25%), and slightly reduces MADs in salinity (by 11%) (Table 1). After the last assimilation step, MADs remain low for at least 4 weeks (Fig. 4). Assimilating additional temperature and salinity profiles (in N2 and N3) further benefits temperature and especially salinity fields, in particular in N3, where the salinity MAD are reduced by 23%, but has almost no effect on SSH and velocity MAD.

Deleted: fraternal

Deleted: S1

Deleted: in the supplement

Deleted: F1

Deleted: S1

Deleted: F2

Deleted: F3

Deleted: F3

In N1 the MAD in SSH, temperature, and velocity components is reduced for almost the entire domain, with the most significant reductions in the LC region (Fig. 5). The reduction in salinity MAD is relatively small in N1 but larger in N3 where additional temperature and salinity profiles are assimilated (Fig. 6). In contrast to SSH, temperature, and velocity, the biggest impact of assimilation on the salinity field is on the shelf where salinity is more variable than in the open Gulf because of river inputs.

Vertically, the reductions of spatially and temporally averaged MAD extend to nearly 900 m depth for temperature and velocity, and 500 m for salinity (Fig. 7). The maximum reductions in MAD amount to 0.6 ℃ for temperature at 200 m, 0.12 for surface salinity, and 0.07 m/s for surface velocity (Fig. 7). Assimilating temperature and salinity profiles in N3 leads to greater reductions of temperature and salinity MAD primarily in the upper 300 m compared to N1.

Next, we assess the impact of assimilation on subsurface temperature and salinity fields (Fig. 8). The "true" spatial distribution of mean temperature and salinity at 400 m depth in August shows only a weak northward intrusion of warm and salty LC water and a detached anticyclonic eddy. Compared to the "truth", the free run overestimates the northward extension of the LC (depicted by the 12 ℃ isotherm and 35.5 isohaline), and the detached eddy is misaligned. Assimilation corrects the extension and angle of the LC and the position of the eddy, significantly reducing the averaged MAD error by 47% and 31% for temperature and salinity, respectively in the N1 run, and 52% and 46% for those in the N3 run.

Lastly, we examine the assimilation impact on subsurface circulation in a comparison of August mean circulation at 400 m depth of the non-identical twin runs (Fig.

Deleted: F1
Deleted: 4
Deleted: F1
Deleted: F3
Deleted: S2 in the supplement
Deleted: 5
Deleted: 5
Deleted: F3
Deleted: F1
Deleted: (Fig. S3 in the supplement)
Deleted: 6
Deleted: (52%)
Deleted: (46%)
Deleted: F1
Deleted: (F3
Deleted:)
Deleted: fraternal

9). The "truth" shows a limited northeastward extension of the LC with two eddies shedding (Fig. 9d). As mentioned already above, the free run overestimates the northward extension and simulates a more energetic detached anticyclonic eddy that has propagated further west (Fig. 9e). Assimilation in N1 brings the simulated shape, strength and location of the LC and LC eddies closer to the "truth" with an overall MAD reduction of ~45% compared to the free run (Fig. 9f). A closer look at the LC intrusion region (Fig. 9g, h, i) and the western (Fig. 9a, b, c) and northern shelf breaks (Fig. 9j, k, l) shows that the greatest improvement in subsurface circulation is in the open Gulf and LC region where mesoscale processes dominate (MAD reduction of ~57%), whereas the improvement in circulation is weaker along the shelf regions where submesoscale processes are important and influences of the open ocean, bathymetry and local wind and river forcing coexist (MAD reductions of ~25% and ~42% on the western and northern shelf, respectively). Specifically, the small-scale currents surrounding the spill site observed in the "truth" (i.e., the strong anticyclonic eddy to the east of the spill site and cyclonic eddy to its southwest) are not satisfactorily represented in either the free run or N1. The results of N2 and N3 are very similar to N1.

3.3. Assimilation impact in identical versus non-identical twins

Assimilating SSH and SST in identical twin I1 leads to even larger error reductions than in the non-identical twin N1 with domain-averaged MAD reductions in temperature of 45%, salinity of 21% and velocity fields of 46%, relative to 29%, 11%, and 25%, respectively, in the non-identical twin N1 (Table 1). However, the benefit of assimilating additional temperature and salinity profiles in I2 and I3 on temperature and salinity fields in the identical twin framework is much smaller than in the non-identical twin (Table 1).

Deleted: 7
Deleted: 7d
Deleted: 7e
Deleted: F1
Deleted: 7f
Deleted: 7g
Deleted: 7a
Deleted: 7j
Deleted: F1
Deleted: F2
Deleted: F3
Deleted: F1
Deleted: fraternal
Deleted: fraternal
Deleted: F1
Deleted: (versus 29% in the fraternal twin)
Deleted: (versus 11%)
Deleted: (versus 25%)
Deleted: fraternal

With respect to the simulated subsurface circulation, the improvement by assimilating SSH and SST is also much greater in identical twin I1 (Fig. 10) than in non-identical twin N1 with a MAD reduction of ~67% versus ~45%. In addition, a remarkable improvement in subsurface circulation following assimilation in I1 is observed not only in the LC intrusion region (MAD reduction of ~69%) but also on the shelves (~55% and ~63%, respectively, on the western and northern shelves), including the region near the DwH spill site (Fig. 10).

Deleted: S4
Deleted: in the supplement
Deleted: fraternal
Deleted: F1
Deleted: S4

4. Discussion

We implemented the EnKF technique in a high-resolution regional model for the GOM. The skill of this data-assimilative system was assessed through a series of non-identical and identical twin experiments assimilating data from different observing system configurations. The differences between the two approaches have important implications for observing system design studies.

Deleted: fraternal

Consistent with previous assimilation studies in the GOM (e.g., Wang et al., 2003; Counillon and Bertino 2009b; Hoteit et al., 2013), our non-identical and identical twin experiments both show that assimilating altimetry data can constrain a range of large-scale to mesoscale features such as the LC and associated eddies. The warmer and more saline LC and its eddies have a temperature and salinity signature that is distinct from the so-called Gulf Common Water and have a clear signal of elevated SSH. Assimilation of SSH using the multivariate EnKF therefore can adjust temperature and salinity profiles based on the SSH information. Assimilation of SSH and SST substantially corrects the subsurface

Deleted: fraternal

temperature, salinity and velocity fields from the surface to depths of up to 900 m, with clear improvements in location and intensity of the LC and LC eddies.

The non-identical twin experiments show that salinity is less constrained than temperature when assimilating only SSH and SST. Assimilation of additional temperature profiles (experiment N2) only slightly improves salinity; inclusion of salinity profiles (experiment N3) is more effective in improving salinity. This highlights the value of assimilating salinity profiles to constrain model salinity fields. The importance of salinity measurements has also been reported in the realistic DA configuration by Halliwell et al. (2015). However, such additional benefits of assimilating temperature and salinity profiles on model-simulated temperature and salinity fields are not observed in the identical twin experiments, which already yield much greater improvements when assimilating SSH and SST alone. 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 non-identical twin. We attribute this to the lack of intrinsic difference in the identical twin (e.g., physical model parameterizations, spatial resolution) between the 'truth' and forecast model runs making it easier to correct the subsurface model fields by assimilating SSH and SST alone. This close agreement of subsurface fields between the forecast model and 'truth' necessarily reduces the additional information content of subsurface observations during assimilation.

Another major difference between the non-identical and identical twin approaches lies in the assimilation impact on subsurface circulation. In the non-identical twin experiments, assimilating satellite altimetry effectively constrains the large to mesoscale structures on the order of 100 km that dominate the deep GOM. The improved circulation

Deleted: fraternal

Deleted: F2

Deleted: F3

Deleted: fraternal

Deleted: fraternal

Deleted: fraternal

in deep GOM has a positive but relatively limited impact on the circulation near the DwH spill site, which is located in the transition zone between the open Gulf (where the circulation is dominated by the mesoscale LC and its eddies) and the shelf (where currents are largely driven by wind and density forcing). The assimilation of SSH, SST and additional temperature and salinity profiles (spatial distance between profiles in experiment N3 is ~70km) in our non-identical 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. In contrast to the non-identical 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.

These results provide two examples of how the identical twin approach yields misleading impact assessments: 1) the improvement in subsurface fields resulting from assimilating SSH and SST is overestimated, and 2) the value of additional profiles is

Deleted: at a resolution of ~70 km

Deleted: fraternal

Deleted: fraternal

underestimated. Undervaluing the information provided by a class of observational assets is particularly troublesome in the context of OSSEs. While this issue is well known in the context of atmospheric OSSEs (e.g., Arnold and Dey 1986; Atlas 1997; Hoffman and Atlas 2016), it is not yet sufficiently recognized for ocean OSSEs and skill assessments of oceanic DA systems. 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 1) that the rate of error growth between simulated and observed states must be similar between the twin framework and reality, and 2) 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.

Deleted: criterion

Deleted: is

Deleted: However,

Deleted: w

Deleted: yet the identical twin proved problematic. Thus, assessing error growth in just one ocean property appears to have been insufficient. In all, our results clearly support the use of the fraternal over the identical twin approach, but they also hint that other criteria in addition to assessing the rate of error growth between the forecast run and "truth" are needed to obtain more credible impact assessment from fraternal twin. …

5. Conclusions

We presented a direct comparison of non-identical and identical twin approaches for assessing data assimilation impact in an EnKF-based ocean DA system for Gulf of Mexico. To the best of our knowledge, this is the first direct comparison of non-identical

Deleted: fraternal

Deleted: fraternal

and identical twin approaches for an oceanic DA system and first demonstration of how the identical twin approach can yield misleading assessments in practice. Our comparisons show that the identical twin approach overestimates the improvement in model skill resulting from assimilating SSH and SST, including for the subsurface circulation, while underestimating the value of additional information from temperature and salinity profiles. In the context of observing system design, such biased assessments are problematic and can lead to misguided decisions on balancing investments between different observing assets. We conclude that skill assessments and OSSEs from identical twin experiments should be avoided or, at least, regarded with caution. While the non-identical twin approach is more robust, questions remain about how to best choose a credible framework. In our case, the rate of error growth in SSH alone appears to have been an insufficient criterion.

Deleted: fraternal

Code and data availability. The ROMS model code can be accessed at http://www.myroms.com (last access: 16 June 2016). ROMS data assimilation model outputs are publicly available through the Gulf of Mexico Research Initiative Information & Data Cooperative (GRIIDC) at https://data.gulfresearchinitiative.org/data/R5.x275.000:0009. HYCOM data can be downloaded at http://tds.hycom.org/thredds/catalog.html (last access: 9 July 2019).

Author contributions. LY and KF conceived the study. LY carried out the model simulations and analysis. BW assisted in preparing the HYCOM data and validating the free model run. AL, KT and LS provided inputs to the model setup and data assimilation techniques. LY and KF discussed the results and wrote the manuscript with contributions from all co-authors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This research was made possible in part by a grant from The Gulf of Mexico Research Initiative (GoMRI-V-487). Data are publicly available through the Gulf of Mexico Research Initiative Information & Data Cooperative (GRIIDC) at https://data.gulfresearchinitiative.org/ data/R5.x275.000:0009. This work used the Extreme Science and Engineering Discovery Environment (XSede) Comet at the San Diego Supercomputer Center through allocation (TG-OCE170001). LY also acknowledges support from the Nova Scotia Graduate Fellowship program.

Deleted: and thanks Thyng et al. 2016 for making the "cmocean" color schemes available

References:

Anderson, D. L. T., Sheinbaum, J., and Haines, K.: Data assimilation in ocean models, Rep. Prog. Phys., 59, 1209–1266, 1996.

Anderson, J. L., and Anderson, S. L.: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts, Mon. Weather Rev., 127(12), 2741–2758, https://doi.org/10.1175/1520-0493(1999)127<2741:AMCIOT>2.0.CO;2, 1999.

Arnold, C.P., Dey, C.H.: Observing-Systems Simulation Experiments : Past, Present, and Future, Bull. Am. Meteorol. Soc., 67, 687–695, 1986.

Atlas, R.: Atmospheric observations and experiments to assess their usefulness in data assimilation, J. Meteorol. Soc. Japan, 75, 111–130, 1997.

Burgers, G., Jan van Leeuwen, P., Evensen, G.: Analysis scheme in the Ensemble Kalman Filter. Mon. Weather Rev. 126, 1719–1724 doi:10.1175/1520-0493(1998) 126<1719:ASITEK>2.0.CO;2. 1998.

Berta, M., Griffa, A., Magaldi, M. G., Ozgokmen, T. M., Poje, A. C., Haza, A. C., and Josefina Olascoaga, M.: Improved surface velocity and trajectory estimates in the Gulf of Mexico from blended satellite altimetry and drifter data, J. Atmos. Ocean Tech., 32(10), 1880–1901. https://doi.org/10.1175/JTECH-D-14-00226.1, 2015.

Carrier, M. J., Ngodock, H., Smith, S., and Jacobs, G.: Impact of assimilating ocean velocity observations inferred from Lagrangian drifter data using the NCOM-4DVAR, Mon. Weather Rev., 142(4), 1509–1524. https://doi.org/10.1175/MWR-D-13-00236.1, 2014.

Chassignet, E. P., Hurlburt, H. E., Metzger, E. J., Smedstad, O., Cummings, J., Halliwell,

G., Bleck, R., Baraille, R., Wallcraft, A. J., Lozano, C., Tolman, H.L., Srinivasan, A., Hankin, S., Cornillon, P., Weisberg, R., Barth, A., He, R., Werner, F., Wilkin, J.: US GODAE: Global ocean prediction with the HYbrid Coordinate Ocean Model (HYCOM), Oceanography, 22(2), 64–75. https://doi.org/10.5670/oceanog.2009.39, 2009.

Chassignet, E. P., Hurlburt, H. E., Smedstad, O. M., Barron, C. N., Ko, D. S., Rhodes, R. C., Shriver, J. F., Wallcraft, A. J., Arnone, R.: Assessment of data assimilative ocean models in the Gulf of Mexico using ocean color, in: Circulation in the Gulf of Mexico: Observations and Models, edited by: Sturges, W., and Lugo-Fernández, A., Geophysical Monograph Series (Vol. 161, pp. 87–100). Washington, DC: American Geophysical Union, 2005.

Counillon, F., and Bertino, L.: Ensemble Optimal Interpolation: Multivariate properties in the Gulf of Mexico, Tellus, Series A: Dynamic Meteorology and Oceanography, 61(2), 296–308. https://doi.org/10.1111/j.1600-0870.2008.00383.x, 2009a.

Counillon, F., and Bertino, L.: High-resolution ensemble forecasting for the Gulf of Mexico eddies and fronts, Ocean Dynam., 59(1), 83–95. https://doi.org/10.1007/s10236-008-0167-0, 2009b.

Evensen, G.: The Ensemble Kalman Filter: Theoretical formulation and practical implementation, Ocean Dynam., 53(4), 343–367. https://doi.org/10.1007/s10236-003-0036-9, 2003.

Fennel, K., Gehlen, M., Brasseur, P., Brown, C.W., Ciavatta, S., Cossarini, G., Crise, A., Edwards, C.A., Ford, D., Friedrichs, M.A.M., Gregoire, M., Jones, E., Kim, H.-C., Lamouroux, J., Murtugudde, R., Perruche, C.: Advancing Marine Biogeochemical and

Ecosystem Reanalyses and Forecasts as Tools for Monitoring and Managing Ecosystem Health, Front. Mar. Sci., 6, 1–9. https://doi.org/10.3389/fmars.2019.00089, 2019

Haidvogel, D. B., Arango, H., Budgell, W. P., Cornuelle, B. D., Cur- chitser, E., Di Lorenzo, E., Fennel, K., Geyer, W. R., Hermann, A. J., Lanerolle, L., Levin, J., McWilliams, J. C., Miller, A. J., Moore, A. M., Powell, T. M., Shchepetkin, A. F., Sherwood, C. R., Signell, R. P., Warner, J. C., and Wilkin, J.: Ocean forecasting in terrain-following coordinates: formulation and skill assessment of the regional ocean modeling system, J. Comput. Phys., 227, 3595–3624, 2008.

Halliwell, G. R., Kourafalou, V., Le Hénaff, M., Shay, L. K., and Atlas, R.: OSSE impact analysis of airborne ocean surveys for improving upper-ocean dynamical and thermodynamical forecasts in the Gulf of Mexico, Prog. Oceanogr., 130, 32–46. https://doi.org/10.1016/j.pocean.2014.09.004, 2015.

Halliwell, G. R., Srinivasan, A., Kourafalou, V., Yang, H., Willey, D., Le Hénaff, M., and Atlas, R.: Rigorous evaluation of a fraternal twin ocean OSSE system for the open Gulf of Mexico, J. Atmos. Ocean Tech., 31(1), 105–130. https://doi.org/10.1175/JTECH-D-13-00011.1, 2014.

Halliwell, G.R., Mehari, M.F., Le Hénaff, M., Kourafalou, V.H., Androulidakis, I.S., Kang, H.S., Atlas, R.: North Atlantic Ocean OSSE system: Evaluation of operational ocean observing system components and supplemental seasonal observations for potentially improving tropical cyclone prediction in coupled systems, J. Oper. Oceanogr., 10, 154–175. https://doi.org/10.1080/1755876X.2017.1322770, 2017.

Hoffman, R.N., Atlas, R.: Future observing system simulation experiments. Bull. Am.

Deleted: Friedrichs, M.A.M.: A data assimilative marine ecosystem model of the central equatorial Pacific: Numerical twin experiments, J. Mar. Res. 59, 859–894. https://doi.org/10.1357/00222400160497544, 2001.¶

621 Meteorol. Soc. 97, 1601–1616. https://doi.org/10.1175/BAMS-D-15-00200.1, 2016.
622 Hoteit, I., Hoar, T., Gopalakrishnan, G., Collins, N., Anderson, J., Cornuelle, B., Kohl, A.,
623 Heimbach, P.: A MITgcm/DART ensemble analysis and prediction system with
624 application to the Gulf of Mexico, Dynam. Atmos. Oceans, 63, 1–23.
625 https://doi.org/10.1016/j.dynatmoce.2013.03.002, 2013.
626 Hu, J., Fennel, K., Mattern, J. P., and Wilkin, J.: Data assimilation with a local Ensemble
627 Kalman Filter applied to a three-dimensional biological model of the Middle Atlantic
628 Bight. J. Marine Syst., 94, 145–156. https://doi.org/10.1016/j.jmarsys.2011.11.016,
629 2012.
630 Jacobs, G. A., Bartels, B. P., Bogucki, D. J., Beron-Vera, F. J., Chen, S. S., Coelho, E. F.,
631 Curcic, M., Griffa, A., Gough, M., Haus, B. K., Haza, A. C., Helber, R. W., Hogan, P.
632 J., Huntley, H. S., Iskandarani, M., Judt, F., Kirwan, A. D., Laxague, N., Valle-
633 Levinson, A., Lipphardt, B. L., J. Mariano, A., Ngodock, H. E., Novelli, G., Olascoaga,
634 M. J., Özgökmen, T. M., Poje, A. C., Reniers, A. J.H.M., Rowley, C. D., Ryan, E. H.,
635 Smith, S. R., Spence, P. L., Thoppil, P. G., Wei, M.: Data assimilation considerations
636 for improved ocean predictability during the Gulf of Mexico Grand Lagrangian
637 Deployment (GLAD), Ocean Model., 83, 98–117.
638 https://doi.org/10.1016/j.ocemod.2014.09.003, 2014.
639 Jones, E. M., Baird, M. E., Mongin, M., Parslow, J., Skerratt, J., Margvelashvili, N., Matear,
640 R. J., Wild-Allen, K., Robson, B., Rizwi, F., Oke, P., King, E., Schroeder, T., Steven,
641 A., Taylor, J.: Use of remote-sensing reflectance to constrain a data assimilating
642 marine biogeochemical model of the Great Barrier Reef, Biogeosciences, 13, 6441–
643 6469. https:// doi.org/10.5194, 2016.

644 Ledwell, J. R., He, R., Xue, Z., DiMarco, S. F., Spencer, L. J., and Chapman, P.: Dispersion
645 of a tracer in the deep Gulf of Mexico, J. Geophys. Res.-Oceans, 121, 1110–
646 1132. doi:10.1002/2015JC011405, 2016.
647 Li, G., Iskandarani, M., Hénaff, M. Le, Winokur, J., Le Maître, O. P., and Knio, O. M.:
648 Quantifying initial and wind forcing uncertainties in the Gulf of Mexico, Computat.
649 Geosci., 1133–1153. https://doi.org/10.1007/s10596-016-9581-4, 2016.
650 Lin, X. H., Oey, L. Y., and Wang, D. P.: Altimetry and drifter data assimilations of loop
651 current and eddies, J. Geophys. Res.-Oceans, 112(5), 1–24.
652 https://doi.org/10.1029/2006JC003779, 2007.
653 Mattern, J. P., Dowd, M., and Fennel, K.: Particle filter-based data assimilation for a three-
654 dimensional biological ocean model and satellite observations, J. Geophys. Res.-
655 Oceans, 118, 2746–2760. https://doi.org/10.1002/jgrc.20213, 2013.
656 Mellor, G. L., and Ezer, T.: A gulf stream model and an altimery assimilation scheme. J.
657 Geophys. Res.-Oceans, 96, 8779–8795, 1991.
658 Moore, A. M., Arango, H. G., Broquet, G., Edwards, C. A., Veneziani, M., Powell, B. S.,
659 Foley, D., Doyle, J., Costa, D. , Robinson, P.: The regional ocean modeling system
660 (ROMS) 4-dimensional variational data assimilation systems, part II: per- formance
661 and application to the california current system. Prog. Oceanogr. 91, 50–73, 2011.
662 Moore, A. M., Martin, M. J., Akella, S., Arango, H. G., Balmaseda, M., Bertino, L.,
663 Ciavatta, S., Cornuelle, B., Cummings, J., Frolov, S., Lermusiaux, P., Oddo, P., Oke,
664 P. R., Storto, A., Teruzzi, A., Vidard, A., Weaver, A.: Synthesis of ocean observations
665 using data assimilation for operational, real-time and reanalysis systems: a more
666 complete picture of the state of the ocean, Front. Mar. Sci. 6:90. doi:

10.3389/fmars.2019.00090, 2019.

Muscarella, P., Carrier, M. J., Ngodock, H., Smith, S., Lipphardt, B. L., Kirwan, A. D., and Huntley, H. S.: Do assimilated drifter velocities improve Lagrangian predictability in an operational ocean model? Mon. Weather Rev., 143(5), 1822–1832. https://doi.org/10.1175/MWR-D-14-00164.1, 2015.

Oey, L.-Y., and Lee, H.-C.: Effects of winds and Caribbean eddies on the frequency of Loop Current eddy shedding: A numerical model study, J. Geophys. Res.-Oceans, 108(C10), 3324. https://doi.org/10.1029/2002JC001698, 2003.

Oey, L.-Y., Ezer, T., Lee, H.-C.: Loop Current, rings and related circulation in the Gulf of Mexico: a review of numerical models and future challenges, in: Circulation in the Gulf of Mexico: Observations and Models, edited by: Sturges, W., and Lugo-Fernández, A., Geophysical Monograph Series (Vol. 161, pp. 87–100). Washington, DC: American Geophysical Union, 2005.

Oke, P. R., O'Kane, T.J. (Eds.): Observing system design and assessment. Operational Oceanography in the 21st Century. Springer, Netherlands, pp. 123–151, 2011.

Sakov, P., and Oke, P. R.: A deterministic formulation of the ensemble Kalman filter: an alternative to ensemble square root filters, Tellus A, 60(2), 361–371. https://doi.org/10.1111/j.1600-0870.2007.00299.x, 2008.

Shay, L. K., Jaimes, B., Brewster, J. K., Meyers, P., McCaskill, E. C., Uhlhorn, E., Marks, F., Halliwell Jr., G. R., Smedstad, O. M., and Hogan, P.: Airborne ocean surveys of the Loop Current complex from NOAA WP-3D in support of the Deepwater Horizon oil spill, in: Monitoring and Modeling the Deepwater Horizon Oil Spill: A Record-Breaking Enterprise, edited by: Liu, Y., Macfadyen, A., Ji, Z.-G., and Weisberg, R.

H., Geophysical Monograph Series (Vol. 195, pp. 131-152). Washington, DC: American Geophysical Union. doi:10.1029/2011GM001101, 2011.

Simon, E., and Bertino, L.: Application of the Gaussian anamorphosis to assimilation in a 3-D coupled physical-ecosystem model of the North Atlantic with the EnKF: a twin experiment, Ocean Sci., 5, 495-510, https://doi.org/10.5194/os-5-495-2009, 2009.

Simon, E., Samuelsen, A., Bertino, L., and Mouysset, S.: Experiences in multiyear combined state-parameter estimation with an ecosystem model of the North Atlantic and Arctic Oceans using the Ensemble Kalman Filter. J. of Marine Syst., 152, 1–17. https://doi.org/10.1016/j.jmarsys.2015.07.004, 2015.

Song, H., Edwards, C. A., Moore, A. M., and Fiechter, J.: Data assimilation in a coupled physical-biogeochemical model of the California Current System using an incremental lognormal 4-dimensional variational approach: Part 2-Joint physical and biological data ssimilation twin experiments. Ocean Model., 106, 146–158. https://doi.org/10.1016/j.ocemod.2016.04.001, 2016a.

Song, H., Edwards, C.A., Moore, A.M., Fiechter, J.: Data assimilation in a coupled physical–biogeochemical model of the California Current System using an incremental lognormal 4-dimensional variational approach: part 3-assimilation in a realistic context using satellite and in situ observations. Ocean Model., 106, 159–172. http://dx.doi.org/10.1016/j.ocemod.2016.04.001, 2016b.

Srinivasan, A., Chassignet, E. P., Bertino, L., Brankart, J.-M., Brasseur, P., Chin, T. M., Counillon, F., Cummings, J. A., Mariano, A. J., Smedstad, O. M., and Thacker, W. C.: A comparison of sequential assimilation schemes for ocean prediction with the HYbrid Coordinate Ocean Model (HYCOM): Twin experiments with static forecast error

covariances. Ocean Model., 37(3–4), 85–111. https://doi.org/10.1016/j.ocemod.2011.01.006, 2011.

Thacker, W. C., Srinivasan, A., Iskandarani, M., Knio, O. M., and Hénaff, M. Le.: Propagating boundary uncertainties using polynomial expansions, Ocean Model., 43–44, 52–63. https://doi.org/10.1016/j.ocemod.2011.11.011, 2012.

Wang, D.-P., Oey, L.-Y., Ezer, T., and Hamilton, P.: Near-surface currents in DeSoto Canyon (1997–99): comparison of current meters, satellite observation, and model simulation, J. Phys. Oceanogr., 33(1), 313–326. https://doi.org/10.1175/1520-0485(2003)033<0313:NSCIDC>2.0.CO;2, 2003.

Wu, H., and Zhu, J.: Advection scheme with 3rd high-order spatial interpolation at the middle temporal level and its application to saltwater intrusion in the Changjiang Estuary. Ocean Model., 33(1–2), 33–51. https://doi.org/10.1016/j.ocemod.2009.12.001, 2010.

Yu, L., Fennel, K., Bertino, L., El, M., and Thompson, K. R.: Insights on multivariate updates of physical and biogeochemical ocean variables using an Ensemble Kalman Filter and an idealized model of upwelling. Ocean Model., 126, 13–28. https://doi.org/10.1016/j.ocemod.2018.04.005, 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.

Xue, Z., He, R., Fennel, K., Cai, W. J., Lohrenz, S., and Hopkinson, C.: Modeling ocean circulation and biogeochemical variability in the Gulf of Mexico. Biogeosciences, 10(11), 7219–7234. https://doi.org/10.5194/bg-10-7219-2013, 2013.

Deleted: Thyng, K. M., Greene, C. A., Hetland, R. D., Zimmerle, H. M., and DiMarco, S. F.: True colors of oceanography: Guidelines for effective and accurate colormap selection. Oceanography, 29(3), 9013. https://doi.org/10.5670/oceanog.2016.66, 2016¶

Table 1. Mean Absolute Deviation (MAD) from the "truth" of physical variables for free and data assimilation runs in non-identical twin and identical experiments. The MAD were averaged over all grid cells excluding the shelves (defined by water depths < 300 m) and daily snapshots from 1 April to 1 October 2010. At assimilation steps the forecast ensemble mean was used for the calculation. The percentage change relative to the free run is presented in parentheses.

	SSH (cm)	T (°C)	S	U (m/s)
Non-identical twin				
Free	11	0.72	0.15	0.21
N1 (satellite only)	5.3 (-51%)	0.51 (-29%)	0.13 (-11%)	0.16 (-25%)
N2 (satellite and scheme 1)	5.3 (-52%)	0.50 (-30%)	0.13 (-13%)	0.16 (-25%)
N3 (satellite and scheme 2)	5.4 (-51%)	0.48 (-33%)	0.11 (-23%)	0.16 (-26%)
Identical twin				
Free	10	0.58	0.093	0.20
I1 (satellite only)	4.2 (-59%)	0.32 (-45%)	0.073 (-21%)	0.11 (-46%)
I2 (satellite and scheme 1)	4.1 (-60%)	0.31 (-47%)	0.072 (-23%)	0.11 (-47%)
I3 (satellite and scheme 2)	4.4 (-57%)	0.29 (-50%)	0.068 (-27%)	0.11 (-46%)

Deleted: fraternal

Deleted: *Fraternal*

Deleted: F1

Deleted: F2

Deleted: F3

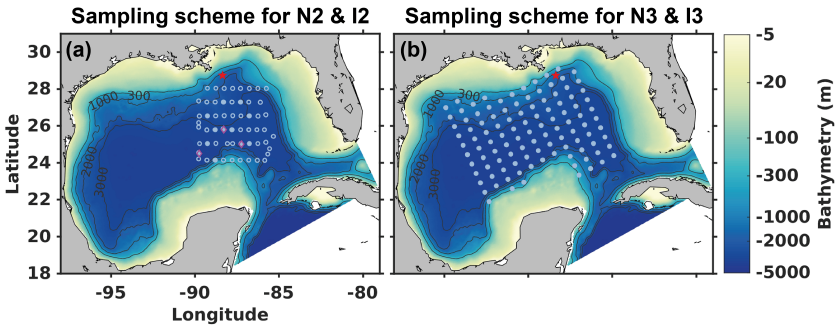
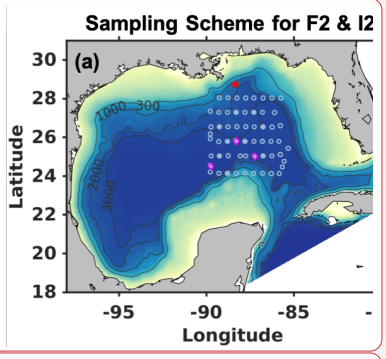


Fig. 1. Model domain and bathymetry. The red star denotes the location of the DwH oil rig. (a) Sampling scheme for twin experiments N2 and I2. The symbols represent stations where temperature (circles) and salinity (magenta diamonds) profiles were collected by Shay et al. (2011), with deep temperature or salinity profiles (down to 1000 m) marked as filled circles or magenta diamonds and shallow temperature profiles (down to 400 m) as open circles. (b) Sampling scheme for N3 and I3. The dots represent stations where temperature and salinity profiles extending to 1000 m depth were sampled from the "truth" run.



Deleted:

Deleted: F2

Deleted: F3

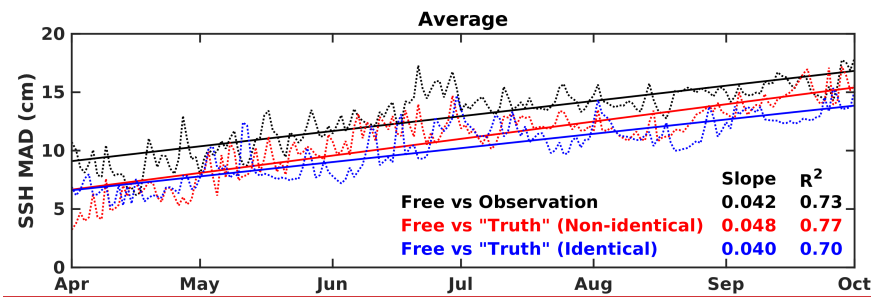
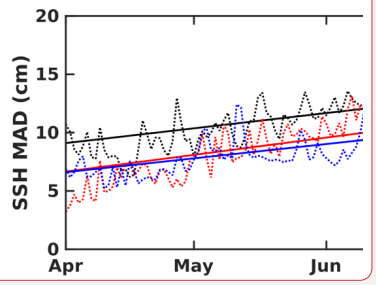
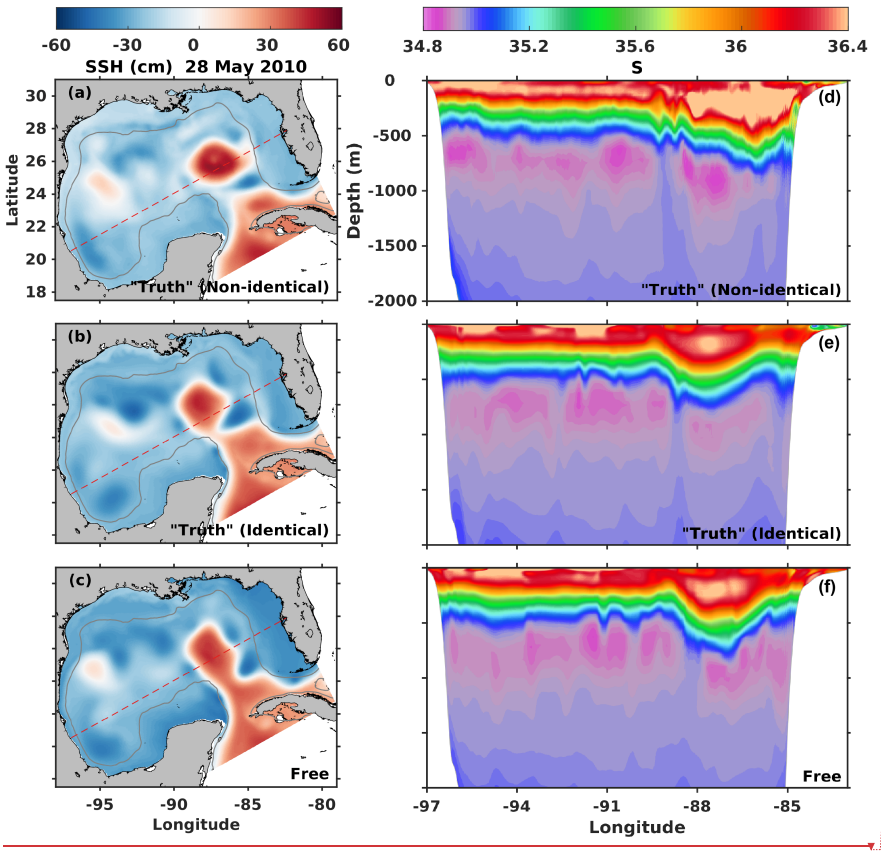


Fig. 2. Time series of MAD error (cm) averaged over the open Gulf (excluding shelf regions shallower than 300 m) for free run's SSH in relative to the SSH from the satellite observation (black dashed line), the "truth" in the non-identical (red) and identical (blue) twin experiments, respectively. The corresponding colored solid lines are linear regressions of the time series, where the slope values represent the respective MAD error growth rate in unit of cm/day.

Deleted:



Deleted: fraternal



Deleted:

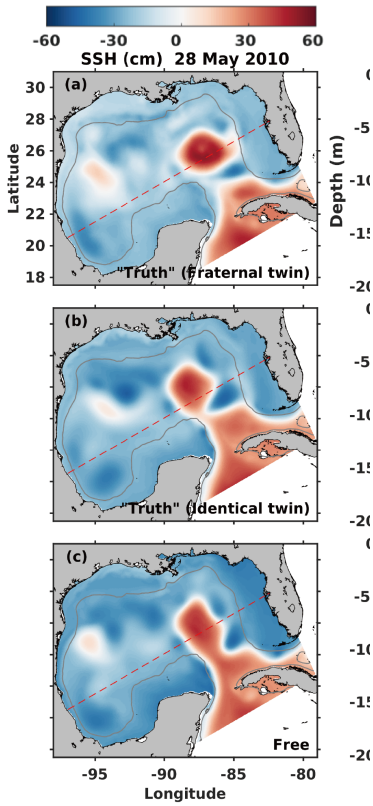


Fig. 3. Sea surface height (SSH, cm) and transect of salinity (S) on 28 May 2010. Panels (a) and (d) are from HYCOM and used as the "truth" in the non-identical twin experiments. Panels (b) and (e) are from ROMS and used as "truth" in identical twin experiments. Panels (c) and (f) are from the free ROMS run. The gray contour in the SSH maps marks the bathymetric depth of 300 m, and the red dashed line shows the position of the transect in panels (d-f).

Deleted: fraternal

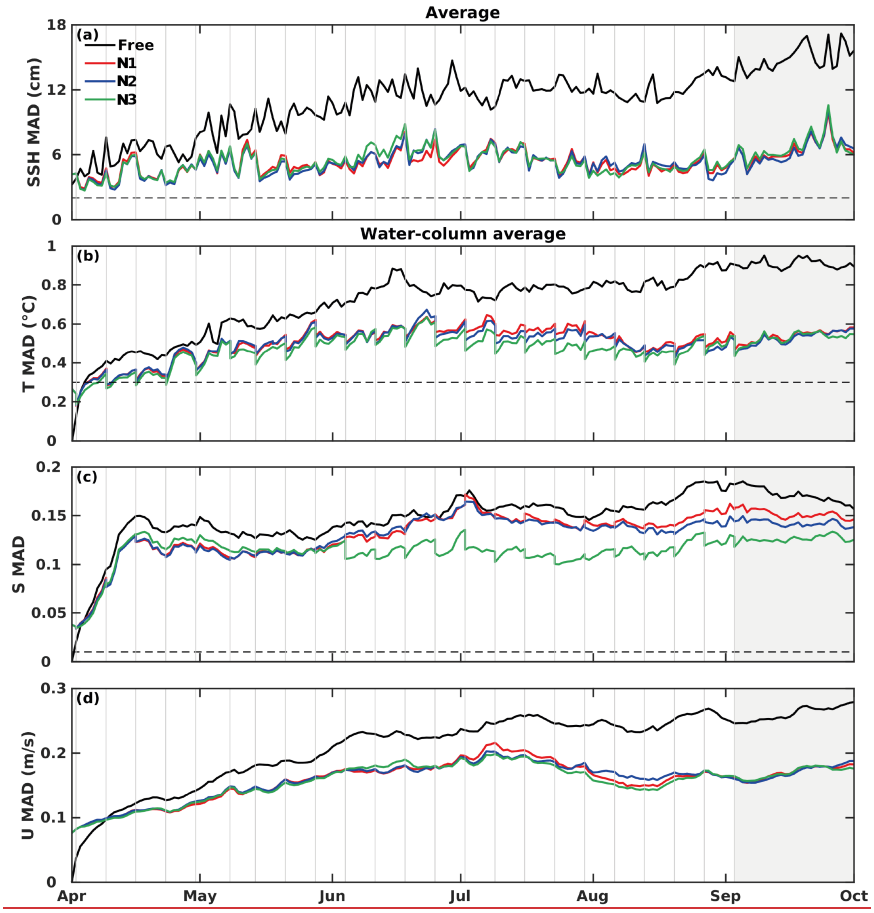
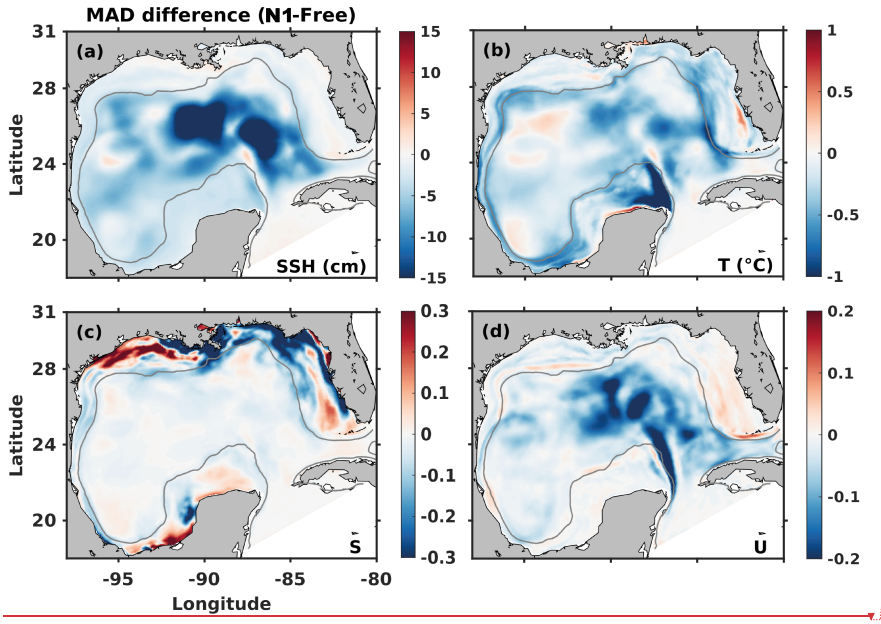


Fig. 4. Time series of MAD averaged over the open Gulf (excluding shelf regions shallower than 300 m) for (a) SSH (cm), (b) temperature (T, ℃), (c) salinity (S), and (d) velocity (U, m/s) from the free run and non-identical twin runs. MAD of all physical variables except SSH were averaged over the entire water column. Black dashed lines in (a, b) denote the values of observation errors. Gray vertical lines indicate the assimilation steps. The gray area marks the 4-week period without data assimilation.



792

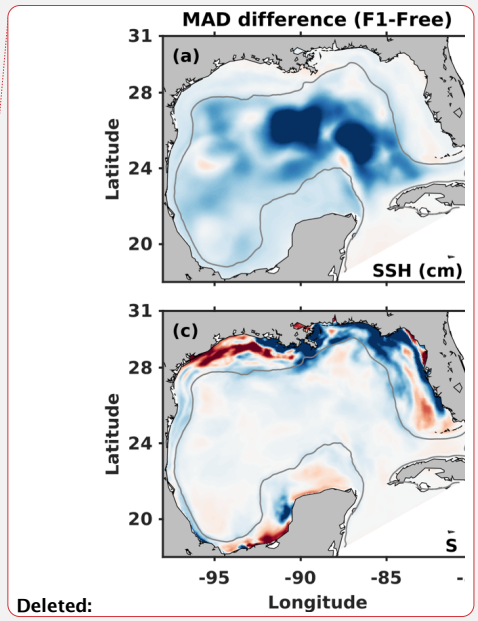
793 **Fig. 5.** The difference of physical variables' time-averaged (daily snapshots from 1 April

794 to 1 October) MAD between non-identical twin N1 and the free run. MAD of temperature

795 and velocity were averaged over the entire water column. Negative values (cold colors)

796 correspond to a decrease in MAD compared to free run, whereas positive values (warm

797 colors) correspond to an increase. The gray contour marks the bathymetric depth of 300 m.



Deleted:

Deleted: 4

Deleted: fraternal

Deleted: F1

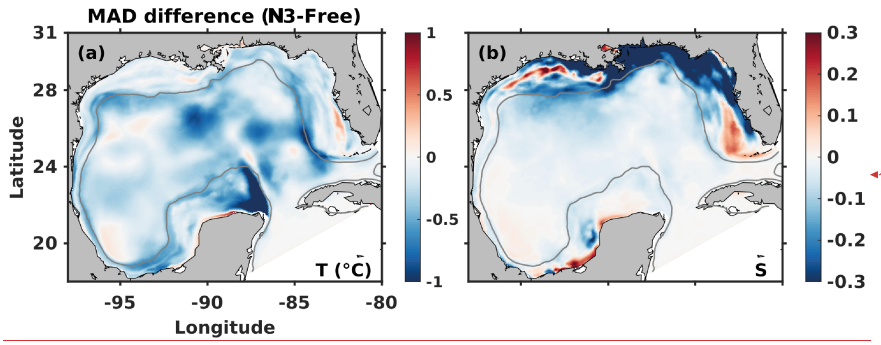


Fig. 6. The difference of physical variables' time- and water-column-averaged (daily snapshots from 1 April to 1 October) MAD between non-identical twin N3 and the free run. Negative values (cold colors) correspond to a decrease in MAD compared to free run, whereas positive values (warm colors) correspond to an increase. The gray contour marks the bathymetric depth of 300 m.

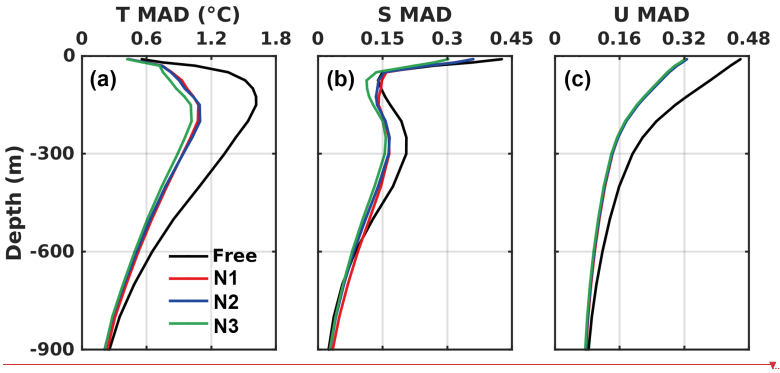
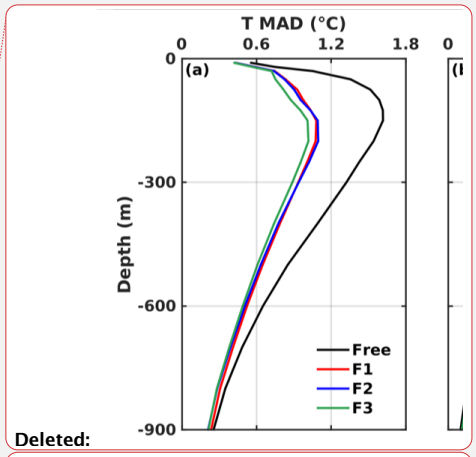


Fig. 7. Profiles of MAD averaged over the open Gulf (excluding shelf regions shallower than 300 m) and daily snapshots from 1 April to 1 October 2010 for (a) temperature (T, °C), (b) salinity (S), and (c) velocity (U, m/s) from the free run and the non-identical twin runs.

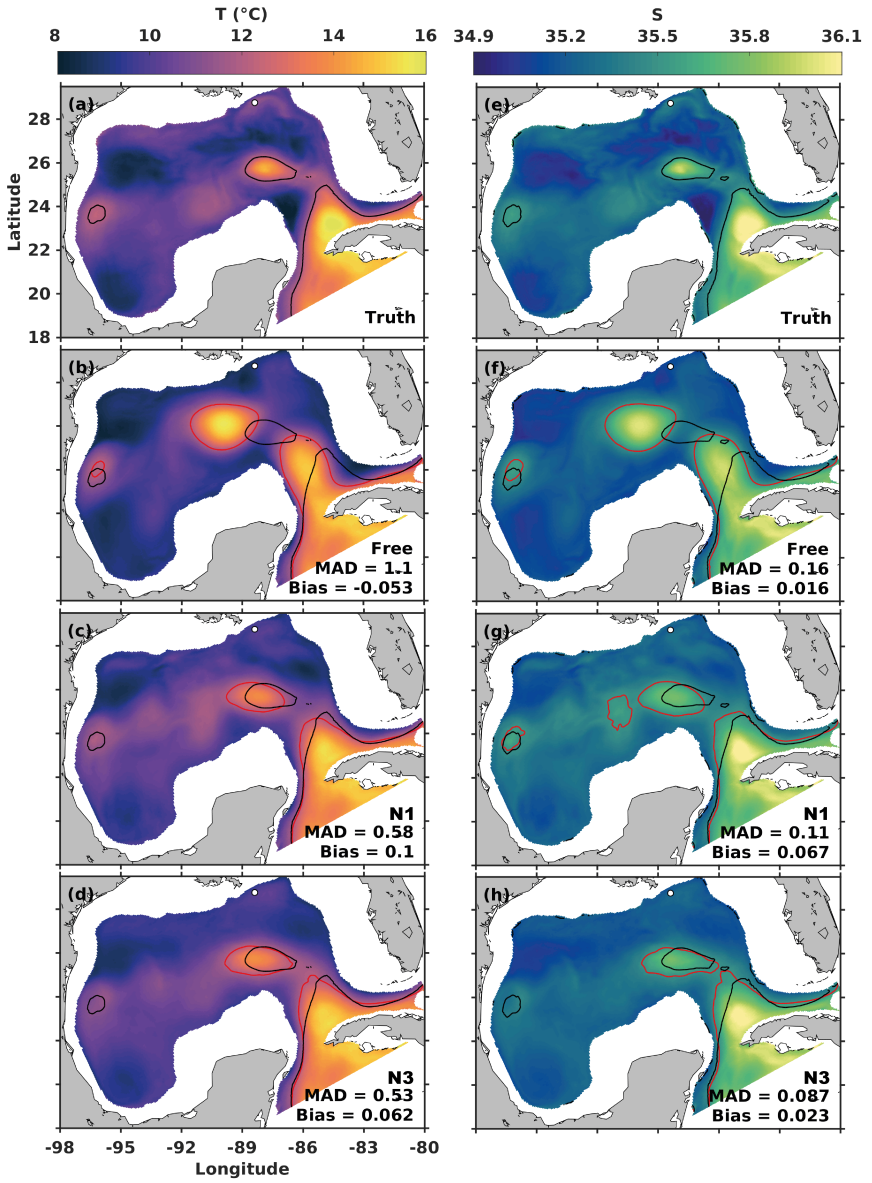


Formatted: Line spacing: Double

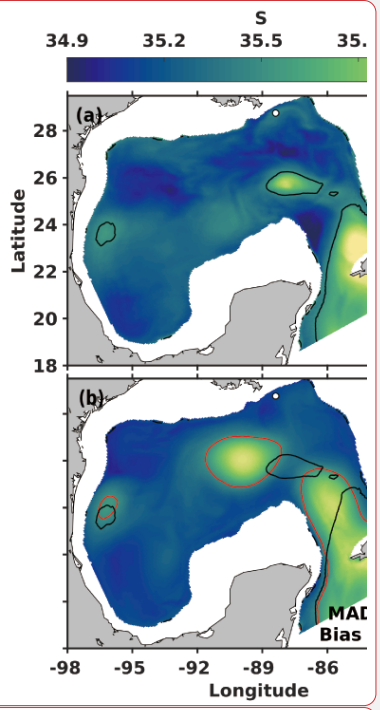
Deleted:

Deleted: 5

Deleted: fraternal



816



Deleted:

Formatted: English (US)

Fig. 8. August-mean (a, b, c, c) temperature (T, ℃) and (e, f, g, h) salinity (S) at 400 m from the "Truth", Free, N1 and N3 run in non-identical twin experiments. The white dot denotes the location of the Deepwater Horizon oil rig. The contours mark the 12 ℃ isotherm and 35.5 isohaline, respectively, where the black contours denote the isotherm or isohaline for the "truth" while red contours denote those for the actual simulation in each panel. The horizontal domain averaged MAD and Bias values at 400 m for each experiment in relative to the "truth" are also presented in respective panel.

Deleted: 6

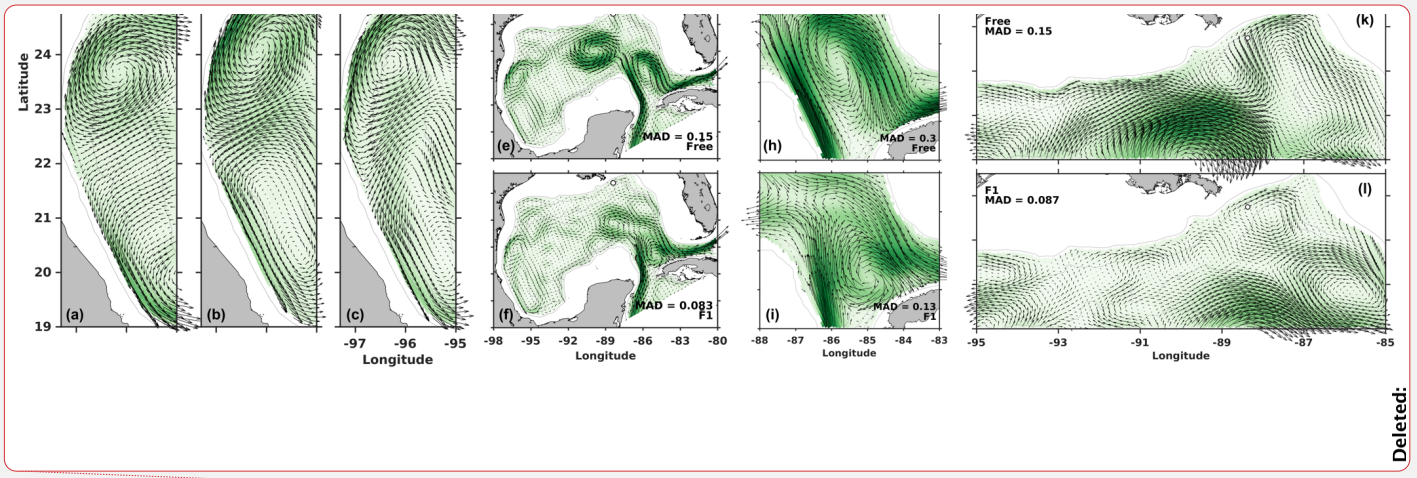
Deleted: (a)

Deleted: (b)

Deleted: (c) F

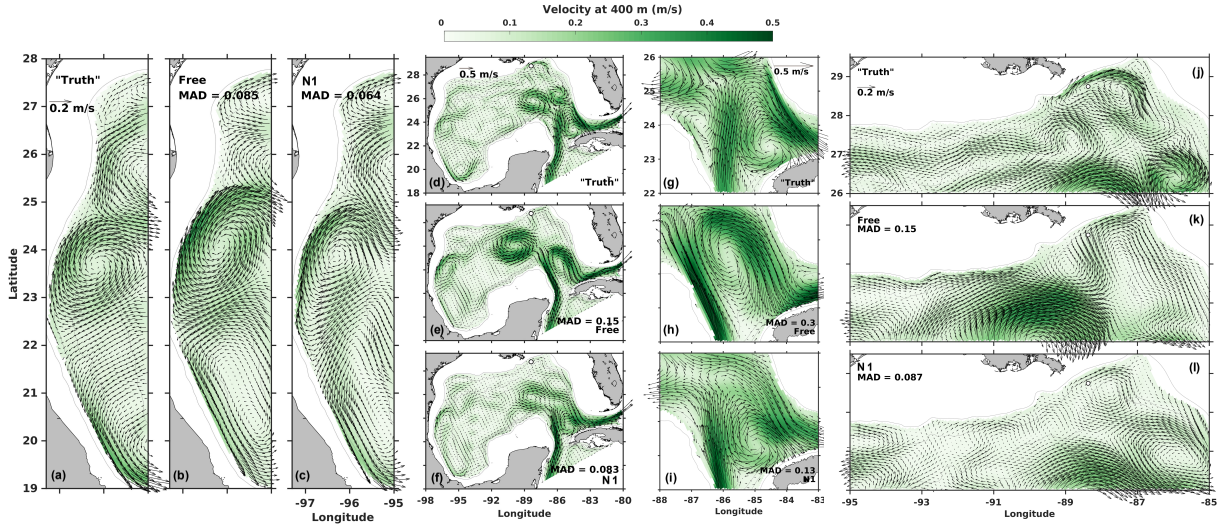
Deleted: (d) F

Deleted: fraternal



Deleted:

38



831

Fig. 9. August-mean velocity at 400 m in the (a, d, g, j) "truth", (b, e, h, k) free and (c, f, I, l) N1 run in non-identical twin experiments. Panels in the 1st, 3rd and 4th columns are zoomed into the western shelf, central Gulf, and norther shelf, respectively. The white dot denotes the location of the DwH oil rig, and gray contours mark the bathymetric depths of 300, 1000, 2000 and 3000 m, respectively.

Deleted: 7

Deleted: F1

Deleted: fraternal

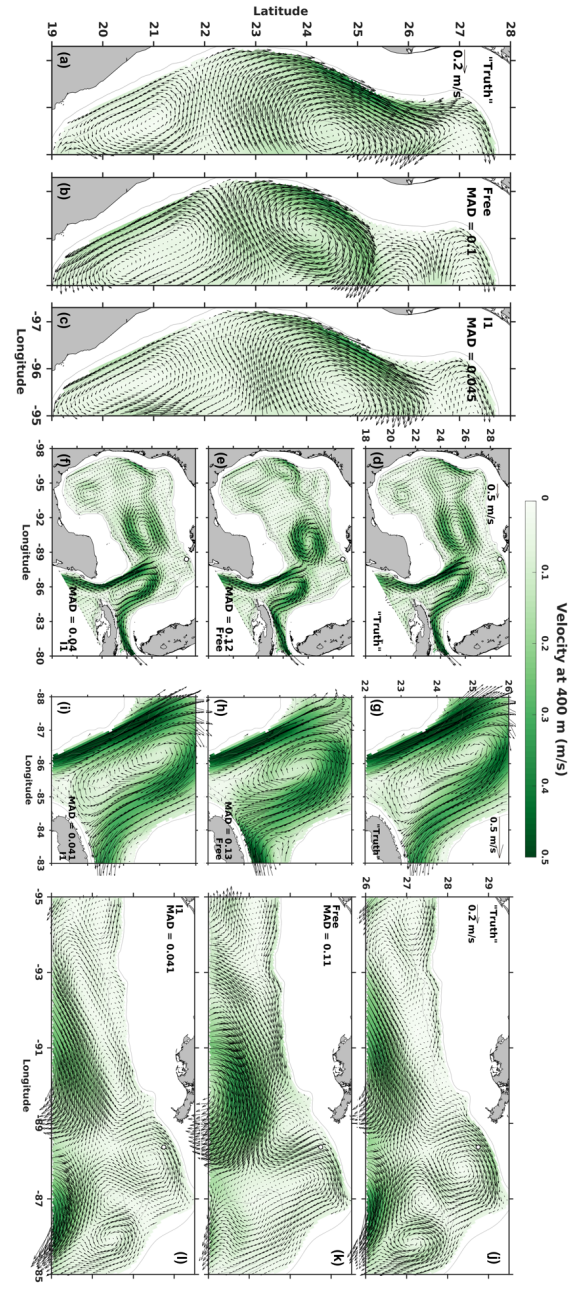

Velocity at 400 m (m/s)
0
0.1
0.2
0.3
0.4
0.5
"Truth"
0.2 m/s
Free
MAD = 0.11
i1
MAD = 0.041
0.5 m/s
"Truth"
MAD = 0.12
Free
MAD = 0.041
i1
Latitude
Longitude
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
(l)

Fig. 10. August-mean velocity at 400 m in the (a, d, g, j) "truth", (b, e, h, k) free and (c, f, I, l) I1 run in identical twin experiments. Panels in the 1st, 3rd and 4th columns are zoomed into the western shelf, central Gulf, and norther shelf, respectively. The white dot denotes the location of the DwH oil rig, and gray contours mark the bathymetric depths of 300, 1000, 2000 and 3000 m, respectively.