

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

3 Liuqian Yu^{1,2}, Katja Fennel¹, Bin Wang¹, Arnaud Laurent¹, Keith R. Thompson¹ and Lynn

4 K. Shay³

5 ¹Department of Oceanography, Dalhousie University, Halifax, Nova Scotia, Canada

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

7 Hong Kong

8 ³Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami,

9 Florida, USA

10 E-mail: liuqianyu@ust.hk

11 **Abstract**

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

25 **1. Introduction**

26 Ocean data assimilation (DA), i.e. the incorporation of observations into ocean
27 models to obtain the best possible estimate of the ocean state, has become standard practice
28 for improving the accuracy of model predictions and reanalyses. Benefiting from the rapid
29 expansion of ocean observing platforms and advances in computing power, various ocean
30 DA applications at both regional and global scales have been developed in support of ocean
31 hindcasts, nowcasts and forecasts (e.g., see recent reviews in Moore et al. 2019 and Fennel
32 et al. 2019). Necessarily the credibility of a DA system demands rigorous validation. It is
33 straightforward to assess the assimilation impact (i.e. the differences between ocean state
34 estimates from a model run with and without assimilation), where a better fit of the model
35 state to observations following assimilation might be considered as positive. But in practice,
36 the value of such an assessment is limited because it either does not consider independent
37 observations (i.e., observations that have not been assimilated into the system) or has to
38 reduce the quantity of data used for assimilation when reserving some for independent
39 assessment.

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

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

57 The identical twin approach has been more commonly used in oceanic DA
58 applications (e.g., Counillon and Bertino, 2009b; Simon and Bertino, 2009; Srinivasan et
59 al., 2011; Song et al, 2016a; Yu et al., 2018) although it is well known from atmospheric
60 OSSEs that this approach provides biased impact assessments when the error growth rate
61 between the “truth” and forecast runs is insufficient (e.g., Arnold and Dey 1986; Atlas 1997;
62 Hoffman and Atlas 2016). This fact is not yet sufficiently recognized in applications of
63 ocean OSSEs and skill assessments of oceanic DA systems (Halliwell et al., 2014). To
64 avoid the potential bias in impact assessments associated with identical twin experiments,
65 Halliwell et al. (2014) proposed to apply a criterion that has long been used in realistic
66 atmospheric OSSEs. They suggested that the model for the forecast run should be
67 configured differently enough from that for the “truth” run so that the rate of error growth
68 between them has the same magnitude as that between state-of-the-art ocean models and
69 the true ocean. They also suggested comparing the assimilation impact in the twin
70 framework with that in a realistic configuration; if a similar impact is obtained in both twin

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

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

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

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

100

101 **2. Model description and experimental setup**

102 **2.1 The physical model**

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

117 Previous studies have highlighted two important aspects for model skill in the GOM,
118 a sufficiently high horizontal resolution for representing the mesoscale dynamics (e.g.,
119 Chassignet et al., 2005) and an accurate representation of the LC inflow through the
120 Yucatan Strait (e.g., Oey, 2003). Our model meets the two requirements. The 5-km
121 horizontal resolution is sufficient to resolve mesoscale processes (the baroclinic Rossby
122 radius is 30 to 40 km in the central GOM, see, Oey et al., 2005). And our ROMS model is
123 nested in a data-assimilative HYCOM model which simulates an accurate structure of the
124 LC and its eddies. Initial model-data comparisons showed that the model has skill in
125 statistically simulating the main features of the LC intrusion with a slight overestimation of
126 its northward penetration during the simulation period (Yu, 2018).

127 **2.2 Experimental framework**

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

$$135 \quad \mathbf{x}^a = \mathbf{x}^f + \mathbf{K}(\mathbf{d} - \mathbf{H}\mathbf{x}^f), \quad (2)$$

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

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

$$141 \quad \mathbf{K} = \mathbf{P}^f \mathbf{H}^T (\mathbf{H} \mathbf{P}^f \mathbf{H}^T + \mathbf{R})^{-1}, \quad (3)$$

142 where \mathbf{P}^f is the $n \times n$ forecast error covariance matrix (approximated by the forecast
143 ensemble), \mathbf{R} is the $m \times m$ observation error covariance, and T denotes the matrix
144 transpose. Different from the traditional EnKF (Burgers et al., 1998) which requires
145 perturbing observations to obtain an analysis error covariance consistent with that given by
146 the Kalman Filter, the DEnKF updates the ensemble mean using the analysis equation (2)
147 and ensemble anomalies with the same equation but half the Kalman gain \mathbf{K} without
148 perturbing observations, and is hence termed ‘deterministic’. Details on the DEnKF
149 derivation and implementation can be found in Sakov and Oke (2008).

150 *2.2.1 Non-identical twin experiments*

151 In non-identical twin experiments, the “truth” is generated by interpolating the daily
152 outputs of the $1/12^\circ$ data-assimilative global HYCOM (Chassignet et al., 2009) onto the
153 ROMS model grid. Synthetic observations are sampled from the “truth”, including SSH,
154 SST, and temperature and salinity profiles. Typical Gaussian observation errors of $N(0, 2$
155 $\text{cm})$ for SSH, $N(0, 0.3 \text{ }^\circ\text{C})$ for temperature (both SST and temperature profiles), and $N(0,$
156 $0.01)$ for salinity are added to the sampled data. SSH and SST are sampled weekly at every
157 fifth horizontal grid point to yield a spatial resolution of $\sim 1/4^\circ$ as such assimilation time
158 window or spatial resolution has been adopted in previous realistic DA applications (e.g.,
159 weekly gridded product of SSH used in Moore et al., 2011, Song et al., 2016b, and weekly
160 gridded product of SST in Hoteit et al. 2013). SSH in regions shallower than 300 m is not
161 used for assimilation because dynamics in shelf areas where wind and buoyancy forcing

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

169 Temperature and salinity profiles were sampled with two different sampling
170 schemes (see locations in Fig. 1a, b). The first scheme adopts the sampling dates and
171 locations used in the survey described in Shay et al. (2011). The key features of this scheme
172 are that the sampling is centered on the LC region, the majority (363 out of 472) of
173 temperature profiles are limited to the upper 400 m, and very few (34) salinity profiles were
174 collected. In the second scheme, coverage was extended such that temperature and salinity
175 profiles are sampled simultaneously over the entire central GOM down to 1000 m depth on
176 23 instead of 9 dates.

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

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

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

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

203 Observations are assimilated weekly from 2 April to 3 September 2010 updating the
204 3D temperature and salinity fields. On each assimilation date, the observations (regardless
205 of observation types) are assimilated simultaneously in one single step. After the last
206 assimilation step on 3 September 2010, the ensemble is run without any data assimilation
207 for 4 more weeks. Three assimilation experiments (referred to as N1, N2 and N3) are

208 conducted. N1 assimilates weekly SSH and SST, while N2 and N3 assimilate the
209 temperature and salinity profiles following the two sampling schemes described earlier (Fig.
210 1a, b) in addition to SSH and SST. Model-data misfit is quantified by computing the Mean
211 Absolute Deviations (MAD), i.e., the average of the absolute deviations, of model
212 simulations from the “truth” for the open Gulf (defined as regions deeper than 300 m). That
213 is, $MAD = \frac{1}{N} \sum_{i=1}^N |model_i - truth_i|$, where $i=1, \dots, N$ and N is the number of data pairs.
214 For ensemble assimilation runs, the forecast ensemble mean at assimilation steps is used
215 for calculating the MAD.

216 **2.2.2 Identical twin experiments**

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

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

228

229 **3. Results**

230 **3.1. Assessment of the non-identical and identical twin experiment setup**

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

245 **3.2. Impact of assimilation in non-identical twin experiments**

246 Temporally and spatially averaged MADs between the non-identical twin
247 assimilation runs and the free run are summarized in Table 1 (temporal evolution is shown
248 in Fig. 4). Assimilating SSH and SST in N1 significantly reduces the MADs of SSH (by
249 51%), temperature (by 29%) and velocity fields (by 25%), and slightly reduces MADs in
250 salinity (by 11%) (Table 1). After the last assimilation step, MADs remain low for at least
251 4 weeks (Fig. 4). Assimilating additional temperature and salinity profiles (in N2 and N3)
252 further benefits temperature and especially salinity fields, in particular in N3, where the
253 salinity MAD are reduced by 23%, but has almost no effect on SSH and velocity MAD.

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

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

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

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

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

292 **3.3. Assimilation impact in identical versus non-identical twins**

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

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

306

307 **4. Discussion**

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

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

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

323 The non-identical twin experiments show that salinity is less constrained than
324 temperature when assimilating only SSH and SST. Assimilation of additional temperature
325 profiles (experiment N2) only slightly improves salinity; inclusion of salinity profiles
326 (experiment N3) is more effective in improving salinity. This highlights the value of
327 assimilating salinity profiles to constrain model salinity fields. The importance of salinity
328 measurements has also been reported in the realistic DA configuration by Halliwell et al.
329 (2015). However, such additional benefits of assimilating temperature and salinity profiles
330 on model-simulated temperature and salinity fields are not observed in the identical twin
331 experiments, which already yield much greater improvements when assimilating SSH and
332 SST alone. It follows that, the additional information content in the subsurface observations
333 (i.e., profiles) within the identical twin system is much smaller than that for the non-
334 identical twin. We attribute this to the lack of intrinsic difference in the identical twin (e.g.,
335 physical model parameterizations, spatial resolution) between the ‘truth’ and forecast
336 model runs making it easier to correct the subsurface model fields by assimilating SSH and
337 SST alone. This close agreement of subsurface fields between the forecast model and ‘truth’
338 necessarily reduces the additional information content of subsurface observations during
339 assimilation.

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

344 in deep GOM has a positive but relatively limited impact on the circulation near the DwH
345 spill site, which is located in the transition zone between the open Gulf (where the
346 circulation is dominated by the mesoscale LC and its eddies) and the shelf (where currents
347 are largely driven by wind and density forcing). The assimilation of SSH, SST and
348 additional temperature and salinity profiles (spatial distance between profiles in experiment
349 N3 is ~70km) in our non-identical twin experiments provides limited constraints on the
350 small-scale circulation features in this region. This is consistent with Wang et al. (2003)
351 who found that assimilating SSH and SST could not accurately resolve smaller-scale eddies
352 in the DeSoto Canyon region near the DwH site. It has been suggested previously that
353 higher-resolution localized observations (Lin et al., 2007; Jacobs et al., 2014; Carrier et al.,
354 2014; Berta et al., 2015; Muscarella et al., 2015) and even finer model resolution (< 5 km,
355 Ledwell et al., 2016) are needed to better constrain these submesoscale features. In contrast
356 to the non-identical twin, the identical twin I1, which assimilates only SSH and SST, yields
357 remarkable improvements not only in the mesoscale circulation dominating the open GOM
358 but also the smaller-scale processes prevailing along the shelf breaks, including the DeSoto
359 Canyon region where the spill site is located. This is largely because in the identical twin
360 setup, the intrinsic model structures (e.g., subgrid-scale parameterizations, horizontal and
361 vertical resolution) for the “truth” and forecast model runs are identical so that an
362 improvement in large-scale processes due to assimilation of SSH and SST can readily
363 translate to an improvement in the simulated subgrid-scale processes.

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

367 underestimated. Undervaluing the information provided by a class of observational assets
368 is particularly troublesome in the context of OSSEs. While this issue is well known in the
369 context of atmospheric OSSEs (e.g., Arnold and Dey 1986; Atlas 1997; Hoffman and Atlas
370 2016), it is not yet sufficiently recognized for ocean OSSEs and skill assessments of oceanic
371 DA systems. Halliwell et al. (2014)'s set of design criteria and evaluation procedures for
372 ocean OSSEs serves as guidance for designing twin experiments for a data-assimilative
373 system. Their main criteria include 1) that the rate of error growth between simulated and
374 observed states must be similar between the twin framework and reality, and 2) that the
375 assimilation impact in the twin framework should be comparable to that of a realistic
376 configuration assimilating actual observations. We found a similar rate of error growth in
377 SSH in both twin experiments and in reality, and the impact of assimilation in the non-
378 identical twin experiment is found to be very similar to that in a realistic assimilation
379 configuration presented in Yu (2018). Thus our direct comparisons of identical versus
380 non-identical twin not only lend support to the recommendation of using the non-identical
381 over the identical twin approach, but also hint that assessing error growth in just one ocean
382 property is insufficient. Additional criteria, such as a comparative assessment of skill
383 between twin and realistic assimilation configurations as described in Halliwell et al (2014),
384 are needed to obtain a more credible impact assessment from the twin framework.

385

386 **5. Conclusions**

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

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

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

407

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

413

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

415

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

423 **References:**

- 424 Anderson, D. L. T., Sheinbaum, J., and Haines, K.: Data assimilation in ocean models, Rep.
425 Prog. Phys., 59, 1209–1266, 1996.
- 426 Anderson, J. L., and Anderson, S. L.: A Monte Carlo implementation of the nonlinear
427 filtering problem to produce ensemble assimilations and forecasts, Mon. Weather Rev.,
428 127(12), 2741–2758, [https://doi.org/10.1175/1520-](https://doi.org/10.1175/1520-0493(1999)127<2741:AMCIOT>2.0.CO;2)
429 0493(1999)127<2741:AMCIOT>2.0.CO;2, 1999.
- 430 Arnold, C.P., Dey, C.H.: Observing-Systems Simulation Experiments : Past, Present, and
431 Future, Bull. Am. Meteorol. Soc., 67, 687–695, 1986.
- 432 Atlas, R.: Atmospheric observations and experiments to assess their usefulness in data
433 assimilation, J. Meteorol. Soc. Japan, 75, 111–130, 1997.
- 434 Burgers, G., Jan van Leeuwen, P., Evensen, G.: Analysis scheme in the Ensemble Kalman
435 Filter. Mon. Weather Rev. 126, 1719–1724 doi:10.1175/1520-0493(1998)
436 126<1719:ASITEK>2.0.CO;2. 1998.
- 437 Berta, M., Griffa, A., Magaldi, M. G., Ozgokmen, T. M., Poje, A. C., Haza, A. C., and
438 Josefina Olascoaga, M.: Improved surface velocity and trajectory estimates in the Gulf
439 of Mexico from blended satellite altimetry and drifter data, J. Atmos. Ocean Tech.,
440 32(10), 1880–1901. <https://doi.org/10.1175/JTECH-D-14-00226.1>, 2015.
- 441 Carrier, M. J., Ngodock, H., Smith, S., and Jacobs, G.: Impact of assimilating ocean velocity
442 observations inferred from Lagrangian drifter data using the NCOM-4DVAR, Mon.
443 Weather Rev., 142(4), 1509–1524. <https://doi.org/10.1175/MWR-D-13-00236.1>,
444 2014.
- 445 Chassignet, E. P., Hurlburt, H. E., Metzger, E. J., Smedstad, O., Cummings, J., Halliwell,

446 G., Bleck, R., Baraille, R., Wallcraft, A. J., Lozano, C., Tolman, H.L., Srinivasan, A.,
447 Hankin, S., Cornillon, P., Weisberg, R., Barth, A., He, R., Werner, F., Wilkin, J.: US
448 GODAE: Global ocean prediction with the HYbrid Coordinate Ocean Model
449 (HYCOM), *Oceanography*, 22(2), 64–75. <https://doi.org/10.5670/oceanog.2009.39>,
450 2009.

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

457 Counillon, F., and Bertino, L.: Ensemble Optimal Interpolation: Multivariate properties in
458 the Gulf of Mexico, *Tellus, Series A: Dynamic Meteorology and Oceanography*, 61(2),
459 296–308. <https://doi.org/10.1111/j.1600-0870.2008.00383.x>, 2009a.

460 Counillon, F., and Bertino, L.: High-resolution ensemble forecasting for the Gulf of Mexico
461 eddies and fronts, *Ocean Dynam.*, 59(1), 83–95. [https://doi.org/10.1007/s10236-008-](https://doi.org/10.1007/s10236-008-0167-0)
462 0167-0, 2009b.

463 Evensen, G.: The Ensemble Kalman Filter: Theoretical formulation and practical
464 implementation, *Ocean Dynam.*, 53(4), 343–367. [https://doi.org/10.1007/s10236-003-](https://doi.org/10.1007/s10236-003-0036-9)
465 0036-9, 2003.

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

469 Ecosystem Reanalyses and Forecasts as Tools for Monitoring and Managing
470 Ecosystem Health, *Front. Mar. Sci.*, 6, 1–9. <https://doi.org/10.3389/fmars.2019.00089>,
471 2019

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

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

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

486 Halliwell, G.R., Mehari, M.F., Le Hénaff, M., Kourafalou, V.H., Androulidakis, I.S., Kang,
487 H.S., Atlas, R.: North Atlantic Ocean OSSE system: Evaluation of operational ocean
488 observing system components and supplemental seasonal observations for potentially
489 improving tropical cyclone prediction in coupled systems, *J. Oper. Oceanogr.*, 10,
490 154–175. <https://doi.org/10.1080/1755876X.2017.1322770>, 2017.

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

492 Meteorol. Soc. 97, 1601–1616. <https://doi.org/10.1175/BAMS-D-15-00200.1>, 2016.

493 Hoteit, I., Hoar, T., Gopalakrishnan, G., Collins, N., Anderson, J., Cornuelle, B., Kohl, A.,
494 Heimbach, P.: A MITgcm/DART ensemble analysis and prediction system with
495 application to the Gulf of Mexico, *Dynam. Atmos. Oceans*, 63, 1–23.
496 <https://doi.org/10.1016/j.dynatmoce.2013.03.002>, 2013.

497 Hu, J., Fennel, K., Mattern, J. P., and Wilkin, J.: Data assimilation with a local Ensemble
498 Kalman Filter applied to a three-dimensional biological model of the Middle Atlantic
499 Bight. *J. Marine Syst.*, 94, 145–156. <https://doi.org/10.1016/j.jmarsys.2011.11.016>,
500 2012.

501 Jacobs, G. A., Bartels, B. P., Bogucki, D. J., Beron-Vera, F. J., Chen, S. S., Coelho, E. F.,
502 Curcic, M., Griffa, A., Gough, M., Haus, B. K., Haza, A. C., Helber, R. W., Hogan, P.
503 J., Huntley, H. S., Iskandarani, M., Judt, F., Kirwan, A. D., Laxague, N., Valle-
504 Levinson, A., Lipphardt, B. L., J. Mariano, A., Ngodock, H. E., Novelli, G., Olascoaga,
505 M. J., Özgökmen, T. M., Poje, A. C., Reniers, A. J.H.M., Rowley, C. D., Ryan, E. H.,
506 Smith, S. R., Spence, P. L., Thoppil, P. G., Wei, M.: Data assimilation considerations
507 for improved ocean predictability during the Gulf of Mexico Grand Lagrangian
508 Deployment (GLAD), *Ocean Model.*, 83, 98–117.
509 <https://doi.org/10.1016/j.ocemod.2014.09.003>, 2014.

510 Jones, E. M., Baird, M. E., Mongin, M., Parslow, J., Skerratt, J., Margvelashvili, N., Matear,
511 R. J., Wild-Allen, K., Robson, B., Rizwi, F., Oke, P., King, E., Schroeder, T., Steven,
512 A., Taylor, J.: Use of remote-sensing reflectance to constrain a data assimilating
513 marine biogeochemical model of the Great Barrier Reef, *Biogeosciences*, 13, 6441–
514 6469. <https://doi.org/10.5194>, 2016.

515 Ledwell, J. R., He, R., Xue, Z., DiMarco, S. F., Spencer, L. J., and Chapman, P.: Dispersion
516 of a tracer in the deep Gulf of Mexico, *J. Geophys. Res.-Oceans*, 121, 1110–
517 1132. doi:10.1002/2015JC011405, 2016.

518 Li, G., Iskandarani, M., Hénaff, M. Le, Winokur, J., Le Maître, O. P., and Knio, O. M.:
519 Quantifying initial and wind forcing uncertainties in the Gulf of Mexico, *Computat.*
520 *Geosci.*, 1133–1153. <https://doi.org/10.1007/s10596-016-9581-4>, 2016.

521 Lin, X. H., Oey, L. Y., and Wang, D. P.: Altimetry and drifter data assimilations of loop
522 current and eddies, *J. Geophys. Res.-Oceans*, 112(5), 1–24.
523 <https://doi.org/10.1029/2006JC003779>, 2007.

524 Mattern, J. P., Dowd, M., and Fennel, K.: Particle filter-based data assimilation for a three-
525 dimensional biological ocean model and satellite observations, *J. Geophys. Res.-*
526 *Oceans*, 118, 2746–2760. <https://doi.org/10.1002/jgrc.20213>, 2013.

527 Mellor, G. L., and Ezer, T.: A gulf stream model and an altimetry assimilation scheme. *J.*
528 *Geophys. Res.-Oceans*, 96, 8779–8795, 1991.

529 Moore, A. M., Arango, H. G., Broquet, G., Edwards, C. A., Veneziani, M., Powell, B. S.,
530 Foley, D., Doyle, J., Costa, D. , Robinson, P.: The regional ocean modeling system
531 (ROMS) 4-dimensional variational data assimilation systems, part II: per- formance
532 and application to the california current system. *Prog. Oceanogr.* 91, 50–73, 2011.

533 Moore, A. M., Martin, M. J., Akella, S., Arango, H. G., Balmaseda, M., Bertino, L.,
534 Ciavatta, S., Cornuelle, B., Cummings, J., Frolov, S., Lermusiaux, P., Oddo, P., Oke,
535 P. R., Storto, A., Teruzzi, A., Vidard, A., Weaver, A.: Synthesis of ocean observations
536 using data assimilation for operational, real-time and reanalysis systems: a more
537 complete picture of the state of the ocean, *Front. Mar. Sci.* 6:90. doi:

538 10.3389/fmars.2019.00090, 2019.

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

543 Oey, L.-Y., and Lee, H.-C.: Effects of winds and Caribbean eddies on the frequency of
544 Loop Current eddy shedding: A numerical model study, *J. Geophys. Res.-Oceans*,
545 108(C10), 3324. <https://doi.org/10.1029/2002JC001698>, 2003.

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

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

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

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

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

563 Simon, E., and Bertino, L.: Application of the Gaussian anamorphosis to assimilation in a
564 3-D coupled physical-ecosystem model of the North Atlantic with the EnKF: a twin
565 experiment, *Ocean Sci.*, 5, 495-510, <https://doi.org/10.5194/os-5-495-2009>, 2009.

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

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

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

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

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

586 Thacker, W. C., Srinivasan, A., Iskandarani, M., Knio, O. M., and Hénaff, M. Le.:
587 Propagating boundary uncertainties using polynomial expansions, *Ocean Model.*, 43–
588 44, 52–63. <https://doi.org/10.1016/j.ocemod.2011.11.011>, 2012.

589 Wang, D.-P., Oey, L.-Y., Ezer, T., and Hamilton, P.: Near-surface currents in DeSoto
590 Canyon (1997–99): comparison of current meters, satellite observation, and model
591 simulation, *J. Phys. Oceanogr.*, 33(1), 313–326. [https://doi.org/10.1175/1520-
592 0485\(2003\)033<0313:NSCIDC>2.0.CO;2](https://doi.org/10.1175/1520-0485(2003)033<0313:NSCIDC>2.0.CO;2), 2003.

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

597 Yu, L., Fennel, K., Bertino, L., El, M., and Thompson, K. R.: Insights on multivariate
598 updates of physical and biogeochemical ocean variables using an Ensemble Kalman
599 Filter and an idealized model of upwelling. *Ocean Model.*, 126, 13–28.
600 <https://doi.org/10.1016/j.ocemod.2018.04.005>, 2018.

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

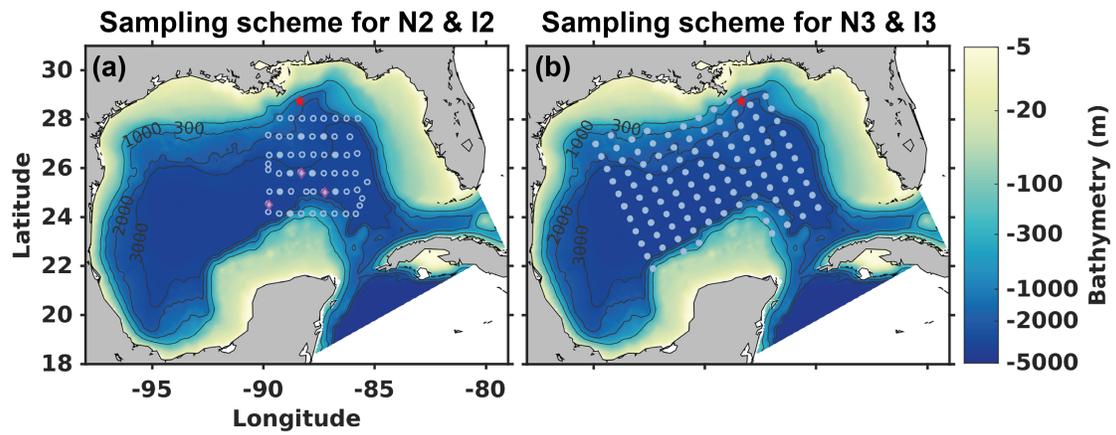
604 Xue, Z., He, R., Fennel, K., Cai, W. J., Lohrenz, S., and Hopkinson, C.: Modeling ocean
605 circulation and biogeochemical variability in the Gulf of Mexico. *Biogeosciences*,
606 10(11), 7219–7234. <https://doi.org/10.5194/bg-10-7219-2013>, 2013.

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

613

	SSH (cm)	T (°C)	S	U (m/s)
<i>Non-identical twin</i>				
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%)
<i>Identical twin</i>				
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%)

614



615

616 **Fig. 1.** Model domain and bathymetry. The red star denotes the location of the DwH oil rig.

617 (a) Sampling scheme for twin experiments N2 and I2. The symbols represent stations where

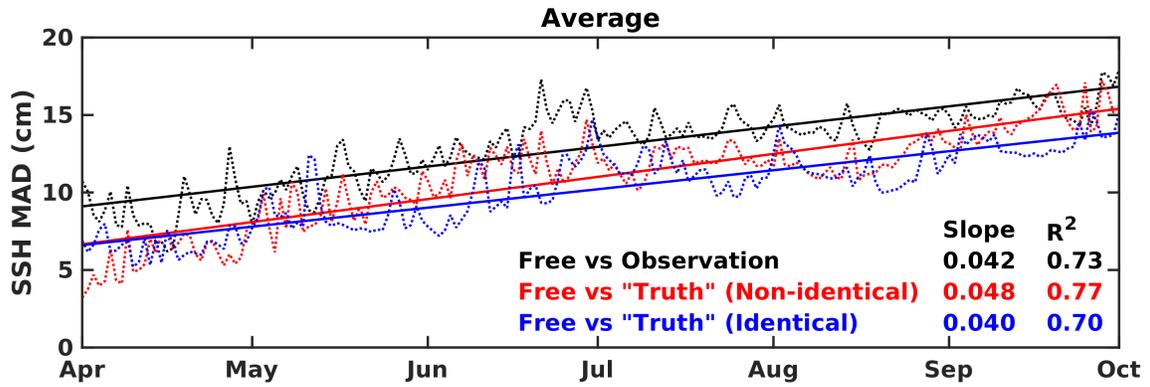
618 temperature (circles) and salinity (magenta diamonds) profiles were collected by Shay et al.

619 (2011), with deep temperature or salinity profiles (down to 1000 m) marked as filled circles

620 or magenta diamonds and shallow temperature profiles (down to 400 m) as open circles. (b)

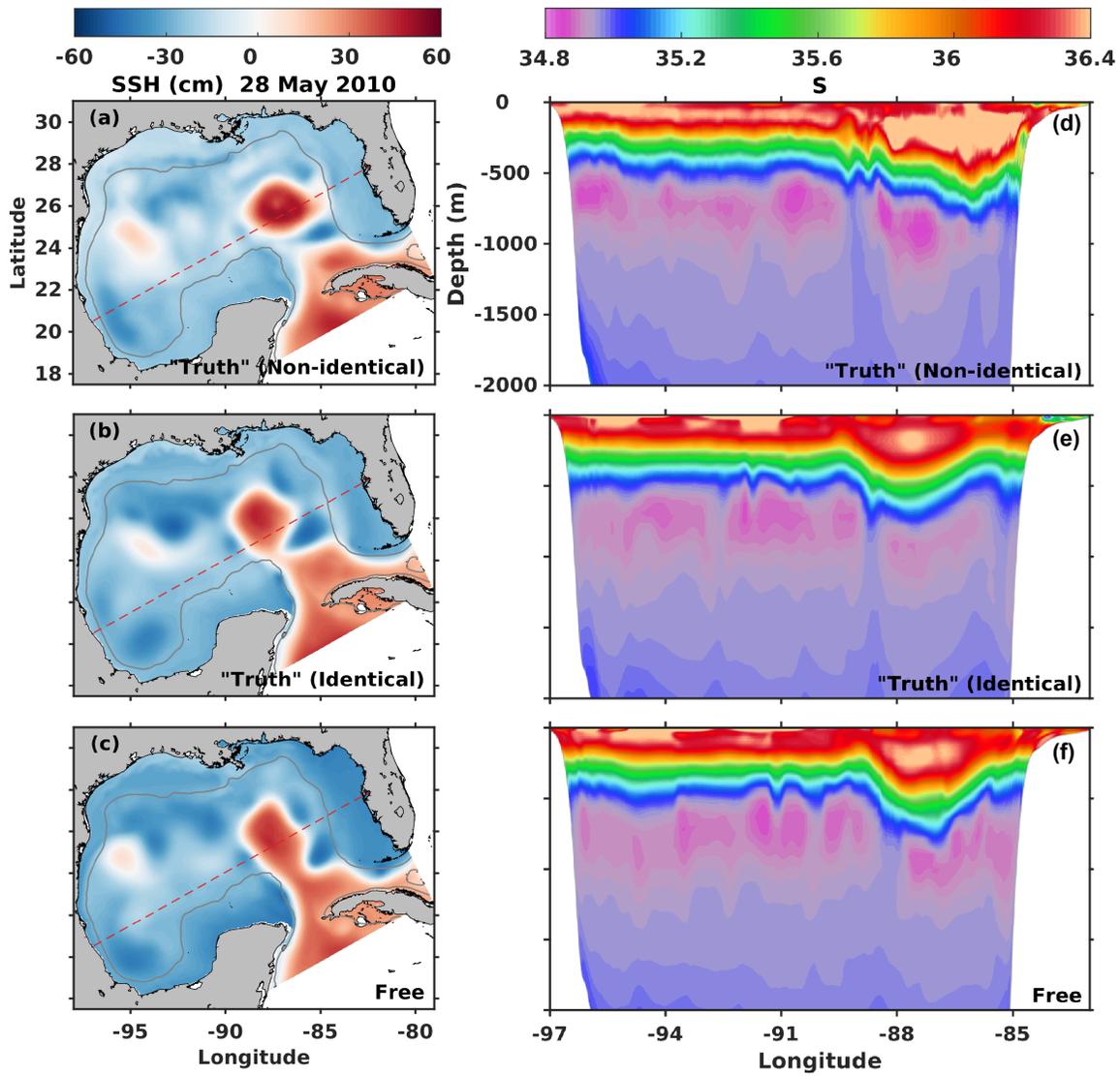
621 Sampling scheme for N3 and I3. The dots represent stations where temperature and salinity

622 profiles extending to 1000 m depth were sampled from the “truth” run.



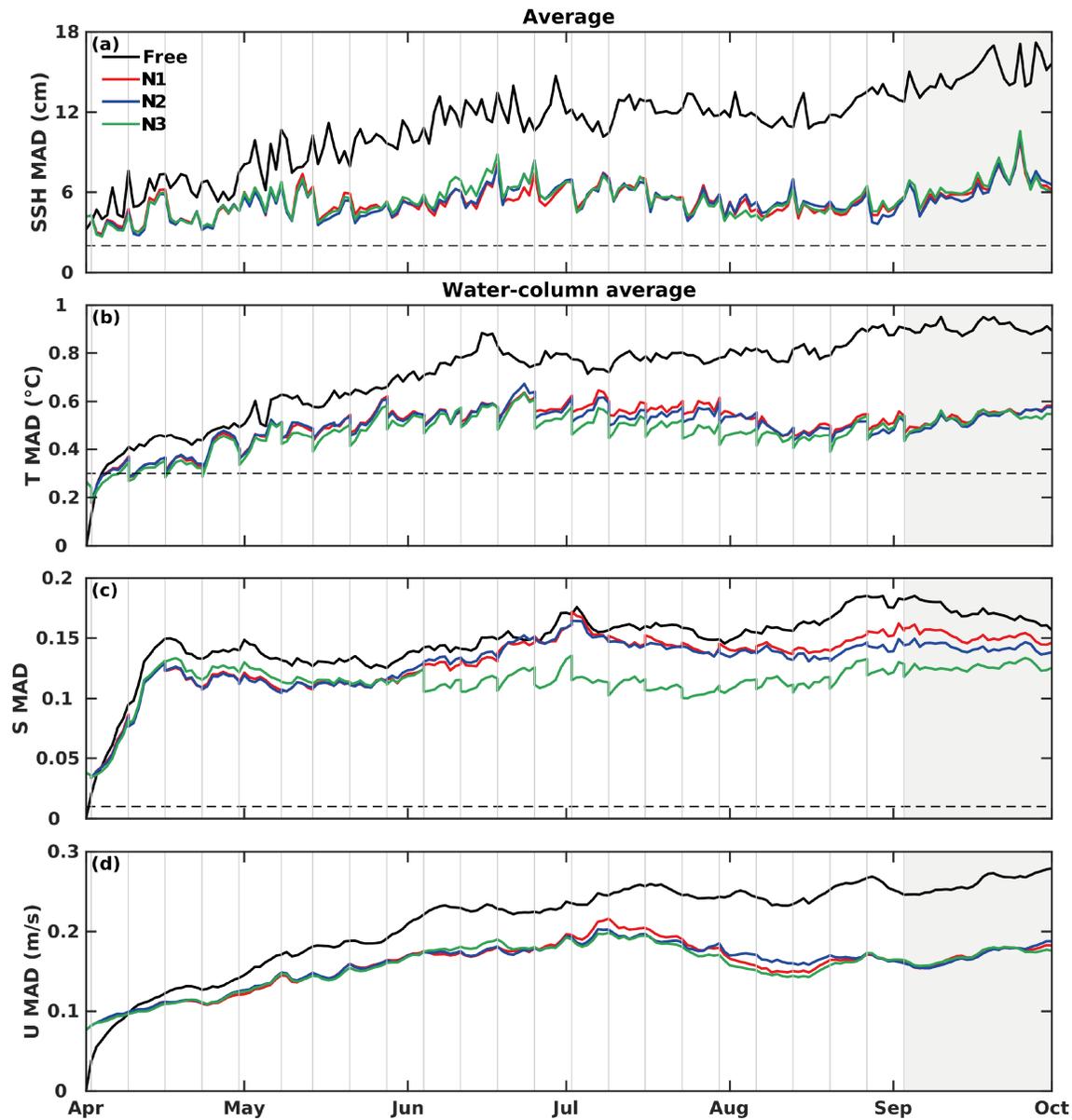
623

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



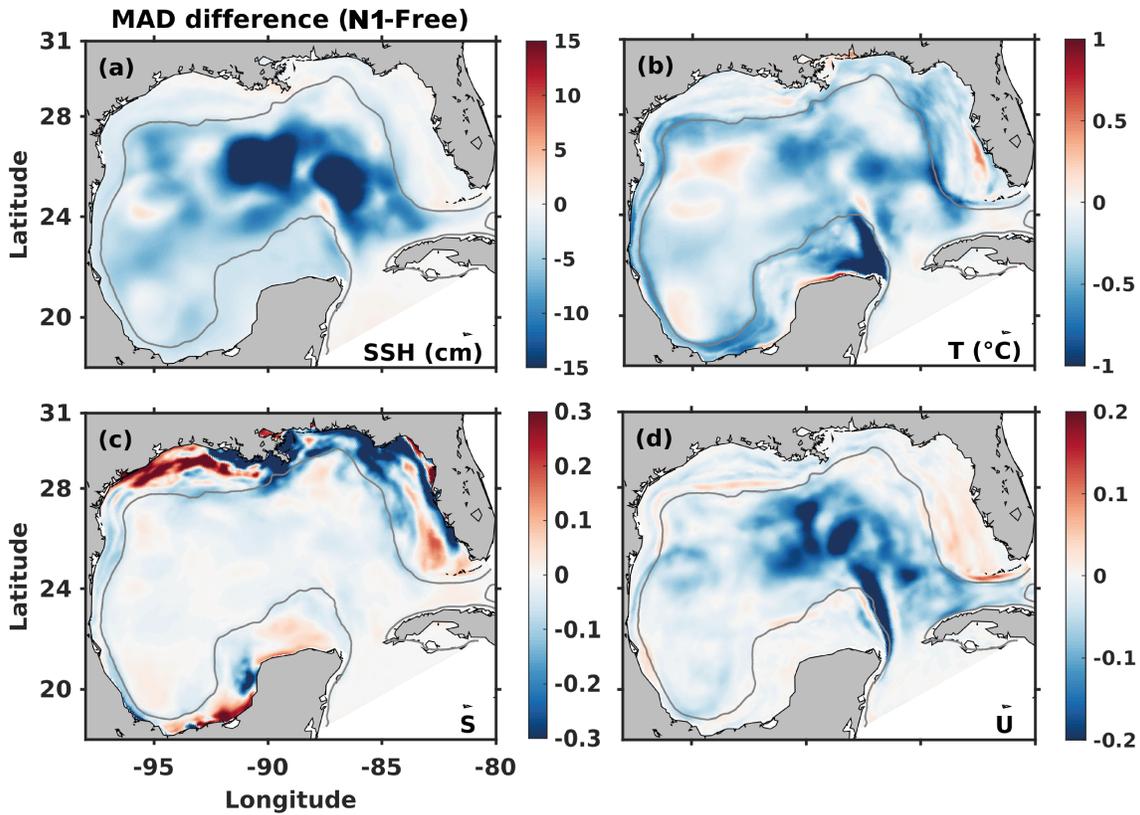
630

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



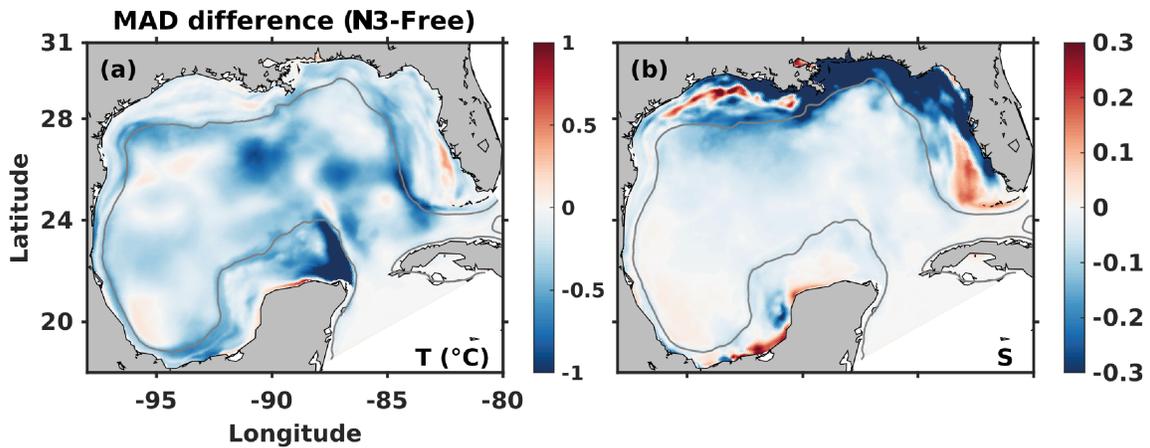
637

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



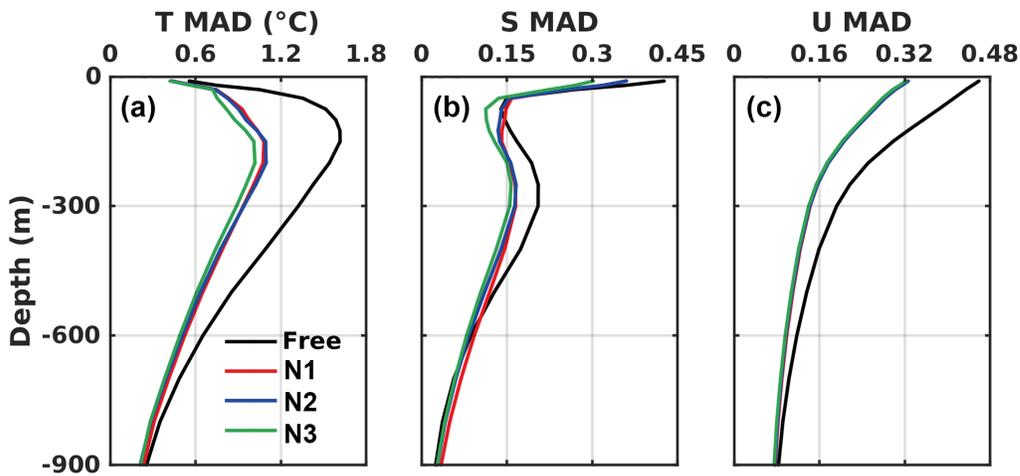
644

645 **Fig. 5.** The difference of physical variables' time-averaged (daily snapshots from 1 April
 646 to 1 October) MAD between non-identical twin N1 and the free run. MAD of temperature
 647 and velocity were averaged over the entire water column. Negative values (cold colors)
 648 correspond to a decrease in MAD compared to free run, whereas positive values (warm
 649 colors) correspond to an increase. The gray contour marks the bathymetric depth of 300 m.



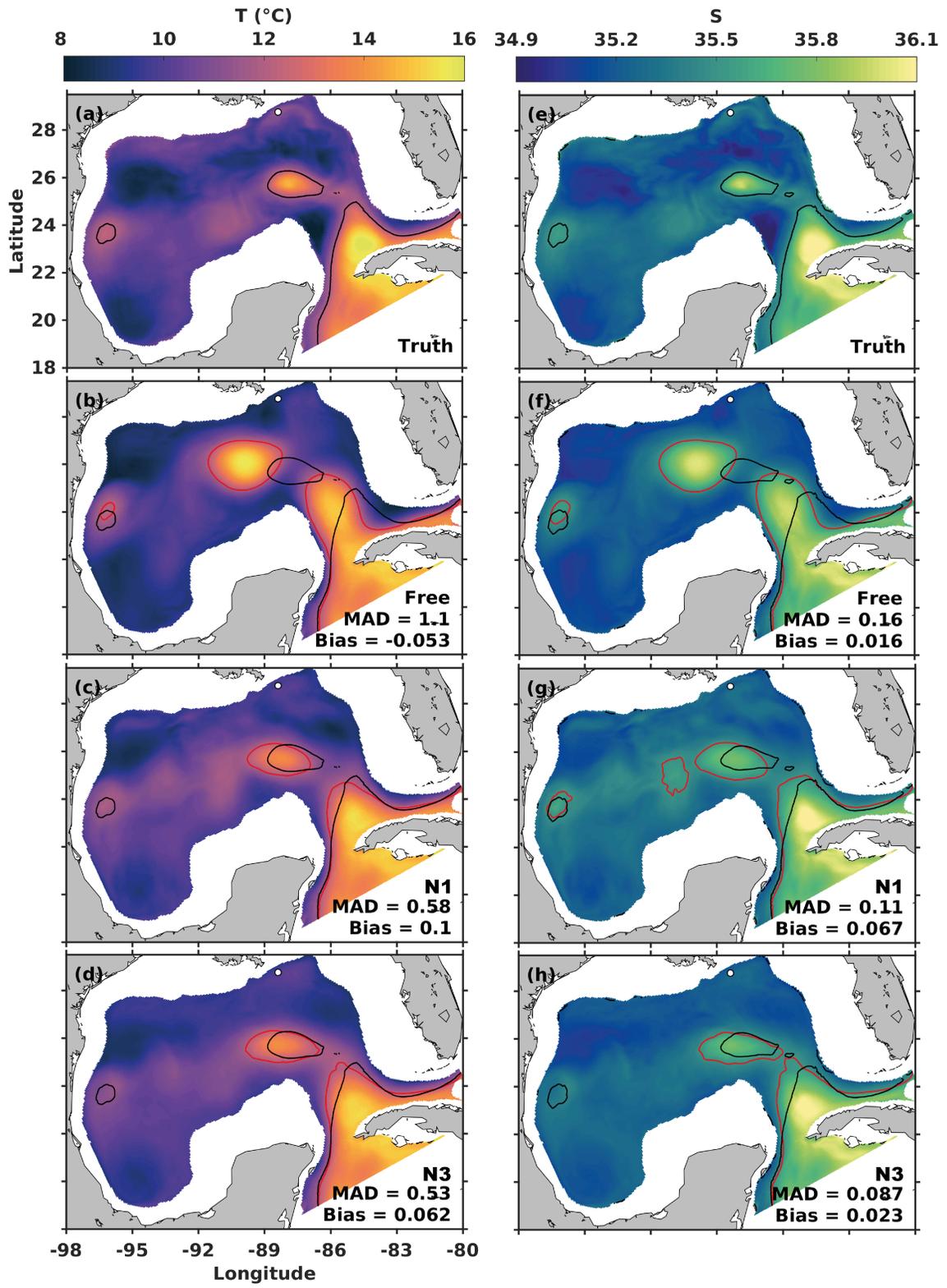
650

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

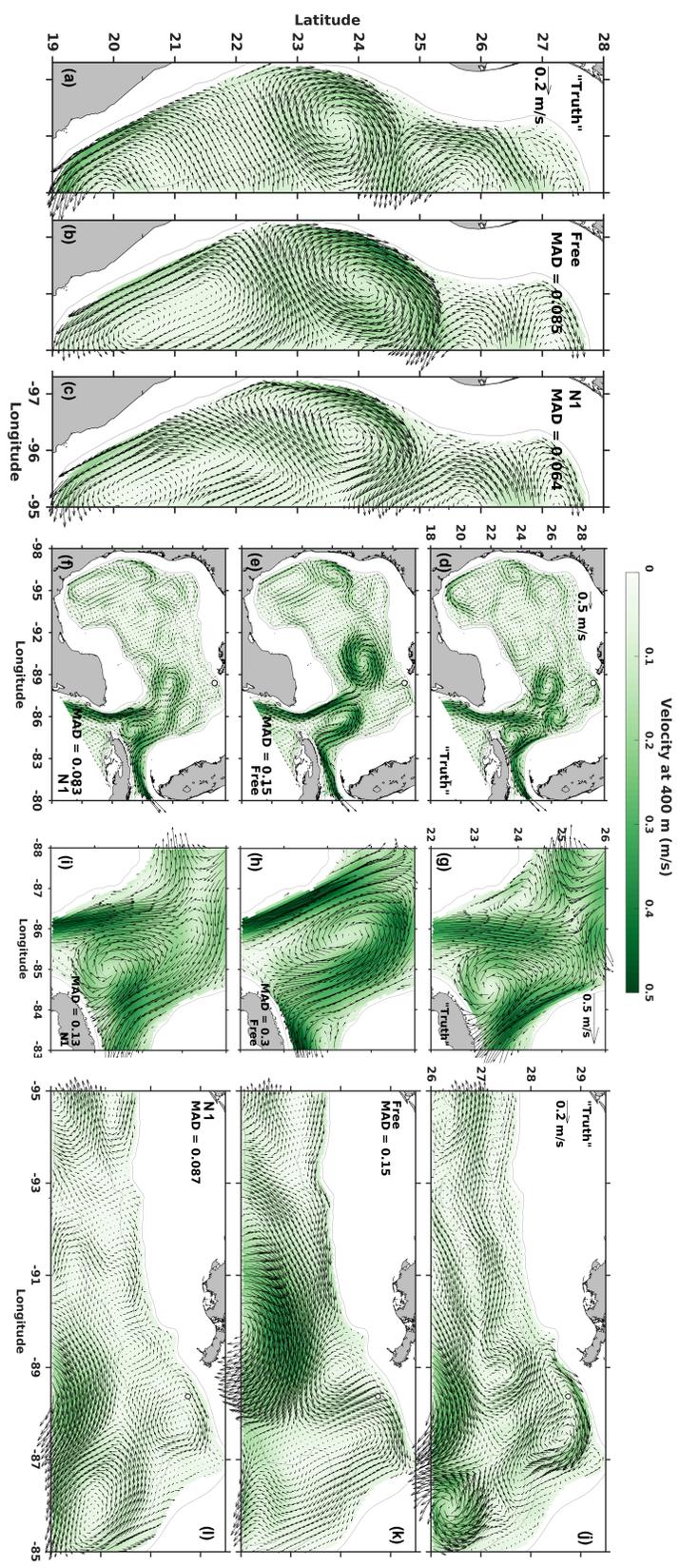


656

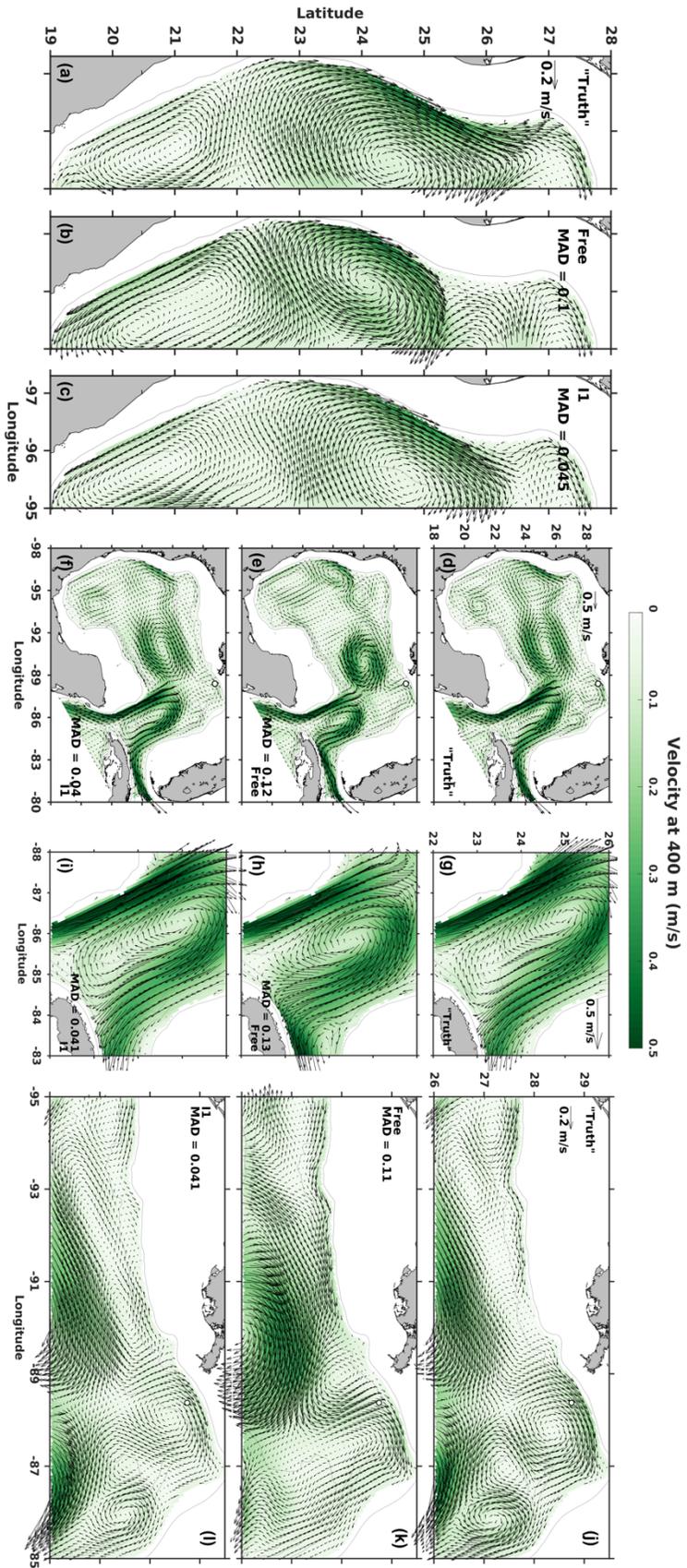
657 **Fig. 7.** Profiles of MAD averaged over the open Gulf (excluding shelf regions shallower
 658 than 300 m) and daily snapshots from 1 April to 1 October 2010 for (a) temperature (T, °C),
 659 (b) salinity (S), and (c) velocity (U, m/s) from the free run and the non-identical twin runs.



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



669 **Fig. 9.** August-mean velocity at 400 m in the (a, d, g, j) “truth”, (b, e, h, k) free and (c, f, I,
670 l) N1 run in non-identical twin experiments. Panels in the 1st, 3rd and 4th columns are zoomed
671 into the western shelf, central Gulf, and norther shelf, respectively. The white dot denotes
672 the location of the DwH oil rig, and gray contours mark the bathymetric depths of 300,
673 1000, 2000 and 3000 m, respectively.



675 **Fig. 10.** August-mean velocity at 400 m in the (a, d, g, j) “truth”, (b, e, h, k) free and (c, f,
676 I, l) I1 run in identical twin experiments. Panels in the 1st, 3rd and 4th columns are zoomed
677 into the western shelf, central Gulf, and norther shelf, respectively. The white dot denotes
678 the location of the DwH oil rig, and gray contours mark the bathymetric depths of 300,
679 1000, 2000 and 3000 m, respectively.