Reviewer 1

The paper presents the GOFS16 short-range ocean forecasting system run operationally at CMCC. The system is run at 1/16° resolution using the NEMO model and OceanVar 3DVar data assimilation scheme. The model, data assimilation and observation components of the system are described, as well as operational implementation details. Assessment of the operational system is presented for the 2-year period January 2022 to December 2023 using bias, RMSE and anomaly correlations of the analyses and forecasts compared against observations. A section on the impact of a tropical cyclone on the ocean state in GOFS16 is also included.
The manuscript is generally clearly written with a useful amount of detail in the system description, though some additional details could be included to strengthen the paper, as suggested below. The assessment of results follows standard procedures used in the ocean forecasting community and gives a clear picture of the quality of the GOFS16 analysis and forecasts. However, the assessment results are often compared with results from other systems published more than 10 years ago, and do not necessarily give a fair impression of the quality of GOFS16 compared to other operational systems of today. Conclusions about the impact of the high resolution of GOS16 also seem rather strong given that there isn't a direct comparison with a lower resolution version of the same system, and the fact that so many aspects of the system (apart from the resolution) can affect the forecast skill.

Response:

We thank the reviewer for the thorough and constructive assessment of our manuscript. We appreciate the positive evaluation of the overall quality of the paper, particularly the clarity of the system description and assessment methodology. We have carefully considered all comments and revised the manuscript accordingly.

Our detailed responses are provided below. Reviewer comments are shown in light black, our responses in bold black.

We have revised the discussion of assessment results to better place GOFS16 in the context of contemporary global forecasting systems and have moderated the discussion on resolution impact to more accurately reflect the evidence presented.

Regarding the first point, we can now refer to a recent international intercomparison of state-of-the-art global ocean forecasting systems that includes GOFS16 (Smith et al., 2026). This manuscript is currently under review, following minor revisions that have already been addressed; we therefore expect it to be accepted soon. We will include a reference to this intercomparison in the revised manuscript, which will provide the broader operational context identified by the reviewer and allow the performance of GOFS16 to be interpreted alongside other state-of-the-art operational systems.

Regarding the second point, we have revised the relevant passages in the Results and Conclusions sections to avoid attributing the reported improvements directly to the higher model resolution. Statements have been reformulated using more cautious wording such as "consistent with the expected benefits of higher resolution" or "likely associated with the improved representation of mesoscale features at 1/16° resolution."

We now explicitly acknowledge that a rigorous attribution of forecast skill to horizontal resolution alone would require a controlled twin experiment with an otherwise identical system at lower resolution, which is beyond the scope of the present work. We believe these revisions more accurately reflect the evidence presented while still highlighting the aspects of the results that are plausibly resolution dependent.

Major comments

1. In the abstract, it is stated that "Results indicate that GOFS16 performs within the expected range of skill for current global systems". Instead, since there isn't a direct comparison with the other current global systems included in the paper, it would be better to give some details of the actual skill of the GOFS16 system based on the assessments presented in the paper.

   We thank the reviewer for raising this point. In the revised manuscript, we have added a reference to Smith et al. (2026, under review), a study in which GOFS16 is evaluated alongside several other recent global ocean forecasting systems as part of a broader international intercomparison effort. This addition provides the external, like-for-like context that was previously missing, and justifies the statement in the Abstract as it stands. We have therefore decided not to modify the sentence.

2. The description of the data assimilation set up would benefit from some clarifications and more information. Specific comments are made in the minor comments section below.

   Following the reviewer's suggestion, we expanded the description of the data assimilation scheme and the associated references.

3. In numerous places, (e.g. L345, L363, L381, L401, L425, L433, L466) results from GOFS16 are compared with those published over 10 years ago in Ryan et al., 2015 and one single recent system description in Liu et al., 2021. The other systems referred to in Ryan et al., 2015 have been developed since that publication with eddy resolving configurations implemented by many centres which is not clear from your descriptions. More recent publications with skill information from some of the other systems is available, e.g. Mignac et al., 2025; Lellouche et al., 2018. Apart from the issue with the old reference and updates to those other systems, it is not a like-with-like comparison of the statistics in the sense that results in the older references are for a different period with different observation sampling, different quality atmospheric forcing etc.

   We agree that repeated reliance on Ryan et al. (2015) as the primary benchmark gives an incomplete picture of the current state of the art. We have revised all seven flagged locations (L345, L363, L381, L401, L425, L433, L466) to incorporate more recent references, particularly Mignac et al. (2025) and Lellouche et al. (2018a), where comparable skill statistics are available. Both systems are also among those evaluated in the intercomparison study by Smith et al. (2026), which offers a further, methodologically controlled benchmark under like-for-like conditions, over the same period and using the same set of observations. We

acknowledge that the participating systems inevitably differ in several respects, reflecting choices made independently by each operational center. Where Ryan et al. (2015) is retained for historical context, we have added explicit acknowledgement that the systems it described have since been substantially updated, with eddy-resolving configurations now operational at many centres. We have also added a clear caveat in the assessment methodology section [Section 4] noting that these older cross-system comparisons are indicative only, given differences in evaluation period, observational sampling, and atmospheric forcing.

4. Conclusions are drawn on L616 about the impact of model resolution on the results (and are implied in other places such as L329, L345, L432). However, there is no comparison of GOFS16 to a lower resolution version of the same system. The comparisons to other, much older, lower resolution systems have major flaws, particularly in that those older systems were run for much older periods with different observing systems and different atmospheric forcings available.

   We thank the reviewer for this precise and well-argued comment. We fully agree that the conclusions regarding the impact of model resolution were expressed too strongly, as the paper does not include a direct comparison with an otherwise identical lower-resolution version of the system. We have therefore revised the manuscript to adopt more cautious language throughout, replacing direct attributional claims with statements indicating that the results are consistent with the expected benefits of higher resolution, rather than demonstrating them conclusively. We have also added an explicit acknowledgement that a rigorous assessment of the isolated impact of spatial resolution would require a dedicated twin experiment with the same system configured at lower resolution. We further agree that comparisons with older, lower-resolution operational systems cannot serve as evidence of the resolution impact, since those systems differ in many aspects beyond horizontal resolution, including the observing systems and atmospheric forcing. We have revised the relevant sections to clarify that these comparisons are included only to provide historical context and should not be interpreted as demonstrating the specific contribution of spatial resolution to forecast skill.

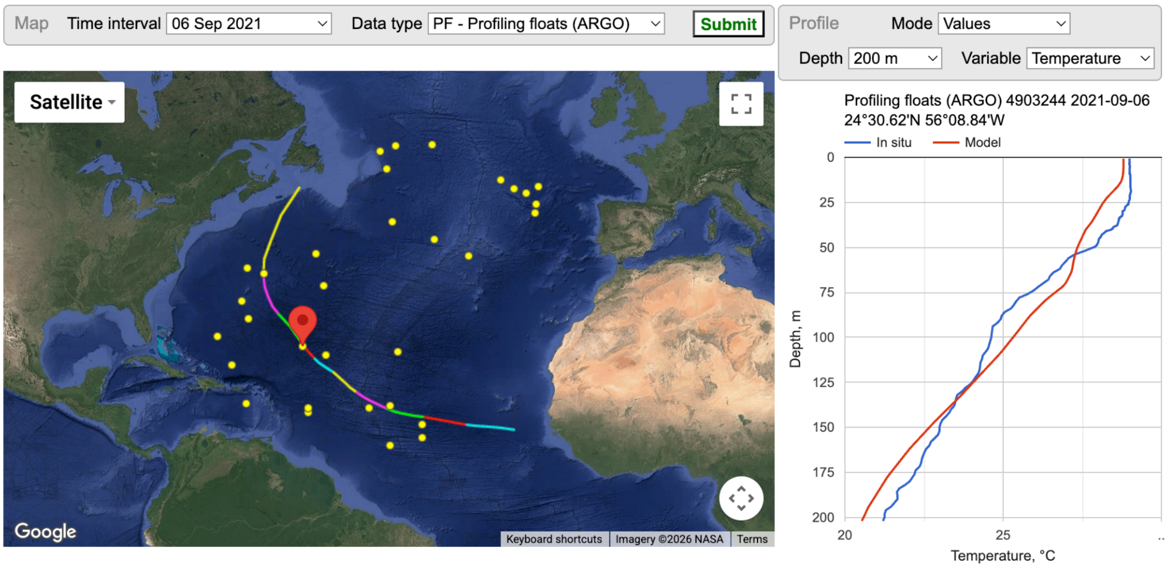
5. It is interesting to see the response of the GOFS16 system to the tropical cyclone in Section 4.4., but this section would be much stronger if there were comparisons of the model forecasts to observations, particularly if there were Argo profiles in the region of the cold wake to show how well the GOFS16 system was representing the true ocean response.

   We thank the reviewer for this suggestion. We agree that a comparison against observations would strengthen the analysis presented in this case study. We therefore searched the Argo dataset for profiles available within the cold-wake region during and following Hurricane Larry's passage and identified two suitable profiles for comparison.
   A comparison against the forecast fields could not be included because the corresponding archived forecast outputs for this event are no longer available/easily accessible. Instead, we now present a comparison between Argo

temperature profiles and the corresponding GOFS16 best estimate, evaluating the extent to which the system reproduces the observed subsurface thermal response to the cyclone's passage. Although these observations are not fully independent, as they contribute to the analysis, they provide a useful check of the consistency of GOFS16 thermal structure with in-situ observations.

Figure 1 shows the track of Hurricane Larry, with colored lines indicating its position on successive days between 1 and 10 September 2021, together with the locations of Argo profiles (yellow dots) acquired on 6 and 8 September 2021 within the cold-wake region. Along the section of the track crossed by the hurricane on 7 September, two Argo profiles were identified (red markers), acquired on 6 and 8 September 2021, they provide snapshots of the upper-ocean thermal structure immediately before and immediately after the storm passage. Their close spatial proximity minimizes the influence of spatial variability, making the comparison well suited to assess the storm-induced changes in the upper ocean.

The comparison shows that GOFS16 reproduces the main temporal evolution of the temperature profile between 6 and 8 September, including the surface cooling and the deepening of the mixed layer following the passage of Hurricane Larry. The agreement between the analysed and observed profiles indicates that GOFS16 captures the storm-induced vertical redistribution of heat, with only minor differences in the thermocline, where GOFS16 tends to produce a slightly smoother thermal gradient.



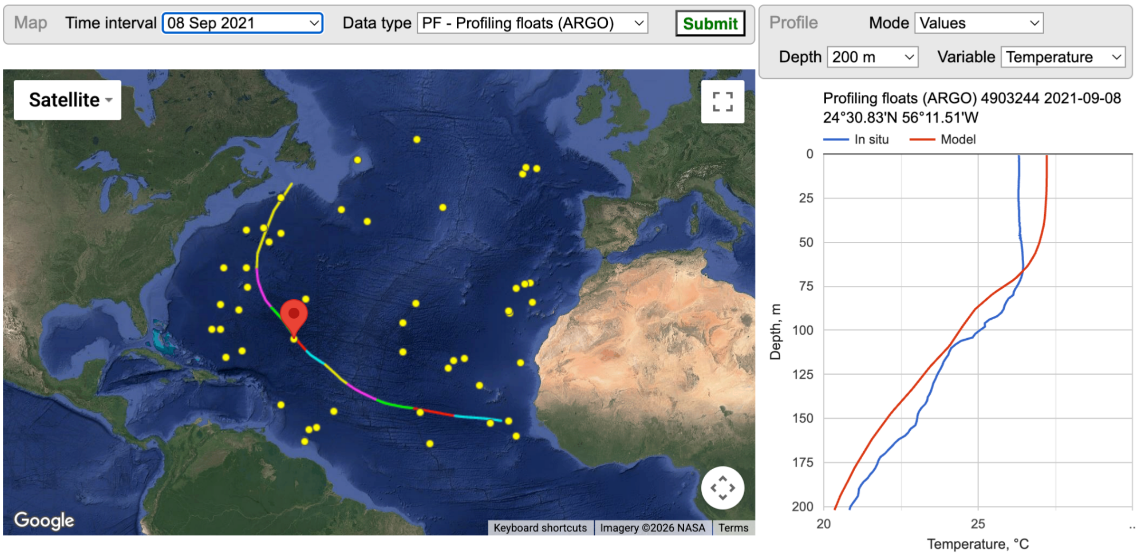


Figure 1. Left: Hurricane Larry track between 1 and 10 September 2021, with colored line segments indicating the storm position on successive days. Yellow dots denote all Argo profiles available within the cold-wake region, while red markers identify the two profiles selected for comparison with GOFS16, acquired on 6 September (top) and 8 September (bottom). Right: Comparison between the Argo temperature profiles (blue) and the corresponding GOFS16 analysed temperature profiles (red) for the two days.

Minor comments

- L44-45: "…the eddying 1/16° configuration (GLOB16) improves the representation of the boundary currents…" – improves compared to what? Do you mean compared to ¼ and 1-degree configurations as described in the Iovino 2023 paper.

  Yes, that is exactly what we meant. We modified the sentence to make it clearer. "It is worth noting that an evaluation of a set of numerical experiments following the Ocean Modeling Intercomparison Protocol (OMIP), using global NEMO configurations at increasing resolution, indicates that GLOB16 improves the representation of boundary currents and transport through key straits compared to the lower-resolution (1° and ¼°) configurations, in addition to a more energetic oceanic state and enhanced mesoscale variability (Iovino et al., 2023)."

- L47. Maybe change "prone" to "designed"?

  Done.

- L56. You could state here when the system was made operational.
  The system became operational in March 2017. We include this information in the text.

- L57. It isn't clear from this sentence which fields are available hourly or daily.
  The system works mainly with daily means. We produced some hourly datasets in smaller regions, but the global 3D hourly mean is not sustainable by our present archive system. We decided to remove the comment on hourly data to not generate confusion.

- L58. When I visited the website (https://gofs.cmcc.it) the display of the outputs was not working and it said "invalid request" over the map.

  We apologise for the inconvenience experienced when visiting the GOFS operational website. The visualisation interface is not yet fully optimised. The website at https://gofs.cmcc.it remains active and will be updated with a fully functional visualisation interface once the optimisation is complete.

- L85. 3.4 is a very old version of NEMO – are there plans to update it?

  NEMO v3.4 is not a recent release, but it was the official NEMO release when the development of GOFS16 started and was therefore adopted as the stable framework for its implementation. Since then, the underlying NEMO code has been upgraded, first to NEMO v4.0 and then to NEMO v4.2.1. The new model configuration, together with revised model physics and updated atmospheric forcing, has already been successfully tested in pre-operational mode, and will form the basis of the next GOFS16 system upgrade.

- L85. "…that integrates numerically primitive equations …" this could be reworded.

  Now it reads: "The global ocean/sea-ice configuration employs the NEMO ocean model (version 3.4), solving the primitive equations numerically with a non-linear equation of state"

- L112-113. Why is the temporal frequency of the surface forcing so low, especially for radiative fields, requiring then an imposed diurnal cycle described on L120?

  We thank the reviewer for this comment. Although higher-frequency atmospheric forcing products (e.g. hourly or 3-hourly) are available, GOFS16 was designed as an operational forecasting system, where robustness and timely delivery of the forcing fields are key requirements. The use of 6-hourly atmospheric forcing provides a good compromise between temporal resolution and operational reliability, reducing the dependency on the timely availability of high-frequency forcing products while maintaining the variability required to force the ocean model. The radiative fields are subsequently averaged to daily values during preprocessing, and the daily-mean shortwave radiation is redistributed within NEMO using the built-in diurnal cycle parameterization (Bernie et al., 2005, 2007), which conserves the daily energy input while representing the day-night cycle. This approach is implemented in the standard NEMO framework and is widely adopted across a broad range of NEMO-based applications.

- L120-121. If the mean sea level can drift in the model, how is this dealt with in the data assimilation?

  In the present configuration, the freshwater budget is not artificially closed and, consequently, the model SSH (or volume) is allowed to drift over time. From the data assimilation (DA) point of view, this situation is not fundamentally different

from an SSH that remains stable; in both cases, the SSH is not evolving realistically due to model approximation and the absence of barystatic contributions (land ice mass loss, etc). To cope with global drift, we follow the standard approach adopted in operational ocean forecasting, disentangling the global trend from DA correction. We remove the global trend by removing the global bias between model and observations every day following Storto et al (2011). This is conceptually like the approach adopted by Lellouche et al (2018a), who impose the global drift externally during the model evolution and remove any global DA correction to SSH.

We expanded the sentence:
"The mean sea level is free to drift. To disentangle the global mean sea level evolution from DA corrections (Lellouche et al 2018a), we applied a daily global unbias scheme prior to assimilation (Storto et al 2011)."

Lellouche, J.-M., Greiner, E., Le Galloudec, O., Garric, G., Regnier, C., Drevillon, M., Benkiran, M., Testut, C.-E., Bourdalle-Badie, R., Gasparin, F., Hernandez, O., Levier, B., Drillet, Y., Remy, E., and Le Traon, P.-Y.: Recent updates to the Copernicus Marine Service global ocean monitoring and forecasting real-time 1/12° high-resolution system, Ocean Sci., 14, 1093–1126, https://doi.org/10.5194/os-14-1093-2018, 2018a.

- L150-151. You talk here about non-linear observation operators. Is this an FGAT scheme then in which the model is used to generate the innovations at the correct time? If not, then which non-linearities are included in H? What time is the background field used in the calculation of the innovations valid for?

In variational schemes, the initial misfits (or departures) can be calculated using an observation operator that differs from the observation operator employed during the minimization. This last operator must be strictly linear while the one in the misfit calculation can be non-linear. In the GOFS16 configuration, the linear observation operator for SLA, used to minimize the cost function, is based on the dynamic height approximation (Cooper and Haines, 1996). On the other hand, initial misfits are calculated using the full model SSH that accounts for all the physics represented by the model. In principle, we could use the dynamic height also to calculate the misfits; however, using our approach, in the ideal case the model state exactly resembles the observations, and the corresponding corrections are identically zero.

We added in the text:

"For example, in the present configuration, the operator used to calculate the misfit (model SSH) differs from the dynamic height operator (Cooper and Haines, 1996) that is employed during the minimization"

Cooper, M., and K. Haines, 1996: Altimetric assimilation with water property conservation. J. Geophys. Res., 101, 1059–1077, https://doi.org/10.1029/95JC02902.

- L151. “construct” -> “constructing”

  Corrected

- L154. Does the vertical operator define the magnitude of the background error variances? Please describe the magnitude and spatial structure of the background error variances in more detail and perhaps show a figure of them.

  The Reviewer is correct; the vertical operator defines both the magnitude and the vertical covariances among temperature and salinity across the vertical levels. We expanded the description for clarity:

  “[...] in a non-assimilative simulation at ¼ deg, similarly to Bellucci et al 2007. The structure of the operator follows Dobricic and Pinardi (2008) and more recently presented in Storto and Masina (2016), Oddo et al 2016, Escudier et al 2021: at each point of the grid, V_v can be decomposed as:

  $$V_v = U\,\Lambda^{1/2}$$

  where the columns of U contain the eigenvectors while Λ is a diagonal matrix with eigenvalues of the EOFs. While the magnitude of the temperature and salinity variances is proportional to the eigenvalues, the vertical shape of the correlation comes from the eigenvectors. To increase the statistics for each point x,y, neighboring points at x±1 and y±1 were also used in the calculation, therefore V_v varies regionally within ½ degree box. “

  We show in Figure 2 below the structure of the covariance matrix among vertical levels of T and S in one point of the North Atlantic at lat 48N, lon -33.68E during summertime. As the EOFs are already described in many papers (as written in the text), we prefer not to add a new figure.

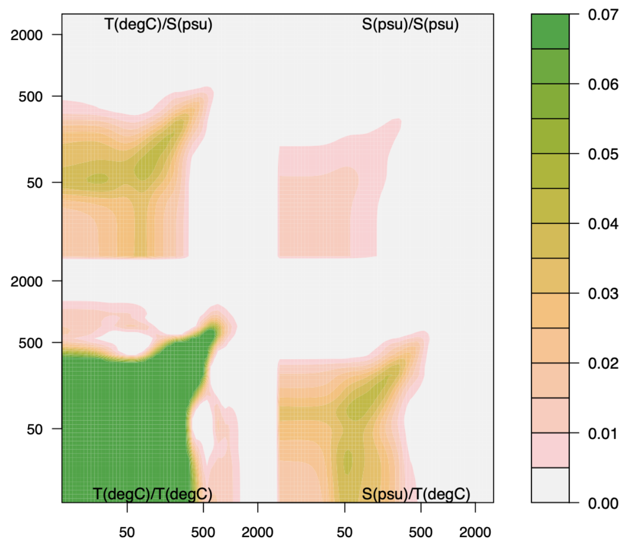


Fig. 2. Structure of temperature and salinity vertical covariance in a point of the North Atlantic at lat 48N , lon -33.68E during summertime.

Escudier R, Clementi E, Cipollone A, Pistoia J, Drudi M, Grandi A, Lyubartsev V, Lecci R, Aydogdu A, Delrosso D, Omar M, Masina S, Coppini G and Pinardi N (2021) A High Resolution Reanalysis for the Mediterranean Sea. Front. Earth Sci. 9:702285. doi: 10.3389/feart.2021.702285

Oddo, P., Storto, A., Dobricic, S., Russo, A., Lewis, C., Onken, R., and Coelho, E. (2016): A hybrid variational-ensemble data assimilation scheme with systematic error correction for limited-area ocean models, Ocean Sci., 12, 1137–1153, https://doi.org/10.5194/os-12-1137-2016.

- L158. Why are no SSH increments produced? How does the model react to T/S increments without balancing increments to SSH or velocity?

The point raised by the Reviewer is an interesting one. We are presently trying to revise such an approach that has been used in the last decade to produce reanalyses based on the C-GLORS system (Storto et al 2016). It is widely known that the initial model imbalance caused by data assimilation is a potential issue and depends on the approximation inevitably present in any DA scheme (Martin et al, 2015). The most well-known examples come from the break of balance between ocean pressure gradient and wind stress at the equator (Waters et al 2017, Balmaseda et al, 2007) or the enhancement of vertical velocity in the first timestep (Pilo et al., 2018, Bell et al., 2004).

Concerning the balance between temperature/salinity and SSH, the SLA assimilation adopts an operator based on the dynamic height that transforms the misfits into density anomaly and generates local thermosteric/halosteric correction. However, in the model equation for the SSH evolution, such steric terms are not present directly but induced by the gradient of the water velocity, specifically:

$$\frac{\partial \eta^f}{\partial t} = \frac{E+R-P}{\rho_w} - \nabla \mathbf{u}\big|_{z=\eta}$$

The model SSH is directly dependent only on the fresh-water balance (E+R-P) and the gradient of velocity. The benefit of adding corrections directly on the SSH, as well as velocity corrections, has been recently revised and discussed in the Mediterranean Sea case by Oddo et al., 2026, who implemented a barotropic operator. Similar tests are going to be extended at global scale. We included a paragraph to discuss the topic.

“[...] are given, following the configuration implemented for reanalysis products (Storto et al,2016) to reduce the impact of initial model imbalance (Pilo et 2018, Waters et al, 2017, Martin et al, 2015, Balmaseda et al, 2007, Bell 2004). The benefits of a new approach including direct corrections in SSH and UV currents have been recently revised by Oddo et al. (2026) in the Mediterranean Sea. Tests at global scales are underway.”

Martin, M. J., Balmaseda, M., Bertino, L., Brasseur, P., Brassington, G., Cummings, J., et al. (2015): "Status and future of data assimilation in operational oceanography." J. Operat. Oceanogr. 8(Supp. 1), s28–s48. doi: 10.1080/1755876X.2015.1022055

Balmaseda, M. A., Dee, D., Vidard, A., and Anderson, D. L. (2007): "A multivariate treatment of bias for sequential data assimilation: application to the tropical oceans". Q .J. R. Meteorol. Soc. 133, 167–179. doi: 10.1002/qj.12

Bell, M.J., Martin, M.J. and Nichols, N.K. (2004), Assimilation of data into an ocean model with systematic errors near the equator. Q.J.R. Meteorol. Soc., 130: 873-893. https://doi.org/10.1256/qj.02.109

Waters, J., Bell, M.J., Martin, M.J. and Lea, D.J. (2017): "Reducing ocean model imbalances in the equatorial region caused by data assimilation." Q.J.R. Meteorol. Soc., 143: 195-208. https://doi.org/10.1002/qj.2912

Oddo, P., Adani, M., Carere, F., Cipollone, A., Goglio, A. C., Jansen, E., Aydogdu, A., Mele, F., Epicoco, I., Pistoia, J., Clementi, E., Pinardi, N., and Masina, S. (2026): "Implementation and evaluation of sea level operators in OceanVar2.0: an open-source oceanographic three-dimensional variational data assimilation system", Geosci. Model Dev., 19, 423–445, https://doi.org/10.5194/gmd-19-423-2026.

Pilo G. S., Oke P. R., Coleman R., Rykova T., Ridgway K. (2018): Impact of data assimilation on vertical velocities in an eddy resolving ocean model,Ocean Modelling, 131, 71-85, https://doi.org/10.1016/j.ocemod.2018.09.003.

- L161. Is the recursive filter performed in 1Dx1D or 2D? How are the normalisation factors calculated to ensure unit correlations at zero separation?

  We expanded the description:
  The horizontal diffusion with non-uniform length-scale follows Storto et al. (2014) whose mathematical formulation has been discussed by Mirouze and Storto (2016) who also proposed a different treatment of the boundaries at ¼° for future versions.
  The V_H is defined as:
  $V_h = Λ_y G_y Λ_x G x$
  where G represents the recursive filter operators in different directions, while Λ is a diagonal matrix with normalization coefficients that account for resolution and correlation scales needed to have correlations with unit amplitude at zero separation (Mirouze and Weaver 2010; Yaremchuk and Carrier 2012).

Mirouze, I. and Storto, A. (2016), Handling boundaries with the one-dimensional first-order recursive filter. Q.J.R. Meteorol. Soc., 142: 2478-2487. https://doi.org/10.1002/qj.2840

Mirouze I and Weaver A.T. (2010): Representation of correlation functions in variational assimilation using an implicit diffusion operator. Q. J. R. Meteorol. Soc. 136: 1421–1443. DOI:10.1002/qj.643

Yaremchuk, M., and M. Carrier (2012): On the renormalization of the covariance operators. Mon. Wea. Rev., 140, 637–649, doi:10.1175/MWR-D-11-00139.1.

- L162. Are the length-scales are coming from a ½ degree resolution model? How would that affect the results of DA using 1/12 degree model?

We agreed that the optimal length-scales should come from an evaluation of the correlation among spatial anomalies extracted from a long-term simulation at the same 1/16° horizontal resolution. Such a time series has recently become available (Iovino et al. 2023), and we plan to investigate the potential improvement associated with deriving the background-error correlation length-scales directly from this eddy-rich configuration.
However, during the 2022-2023 period, GOFS16 production employed horizontal length-scales derived from Storto et al 2014, and subsequently reshaped to access small-scale variability. While the length-scales were kept unchanged in the tropics, in the other regions the length scales were reduced by a factor proportional to the model resolution with a smoothing transition between the two regimes. This approach, originally proposed and evaluated in Cipollone et al. (2020), shows that the use of such rescaling is beneficial when assimilating dense observations (SLA and SST). Following the discussion in Cipollone et al. (2020), the figure below compares the increments coming from the assimilation of SST obtained using two different grids, at ¼° and 1/16°. We used the same horizontal length-scale but rescaled according to the model resolution when applied to 1/16°. The GOFS16 grid has a higher number of degrees of freedom coming from the higher resolution and lower length-scales, thus being better suited for dense observations. The overall SST RMSE was smaller at 1/16°, providing that the resolution of the model engine was identical in the two experiments.

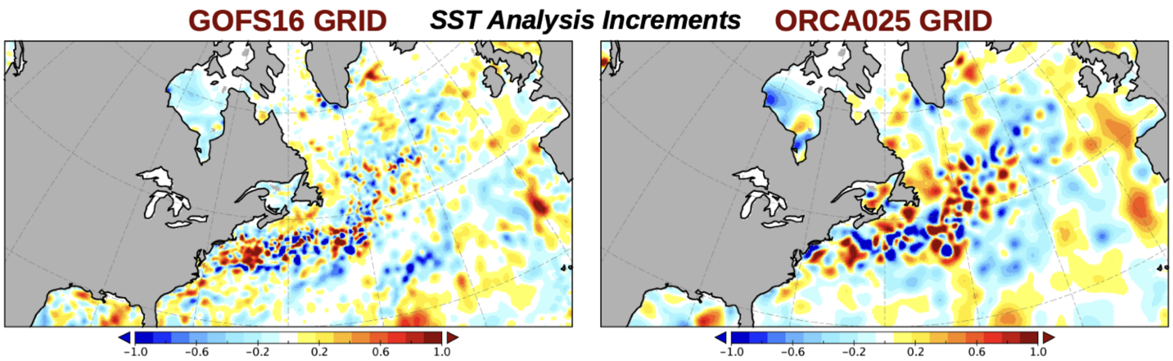


Fig. 3. SST increments for one day of analysis with different horizontal lengths scales and grid. Left panel, SST increments for GLOB16 configuration. Right panel, similar but using ORCA025 configuration.

We expanded the discussion as follows:

“The latter were estimated from the dataset of anomalies of temperature and salinity (Storto et al. 2014) and subsequently reshaped to capture the small-scale variability. While in the tropics the length-scales have been kept constant, in the other regions, we divided the original length scales by a factor proportional to the final resolution with a smoothing linear transition. Smaller horizontal length-scales are beneficial especially for dense observations such as SST retrievals, as discussed in Cipollone et al 2020.”

- Table 1. For SLA you list both L3 and L4 products – why is that?

  While we assimilate observation from L3, the bias is usually defined by comparing the model field against L4 maps. In this way all the regions contributed equally to the definition of bias thus reducing the imbalance among areas populated differently.

- L170-176. I found it hard to understand the reasoning behind the SLA bias correction scheme. Are the innovation bias and other terms calculated at each model grid point or some lower resolution grid? Is the resulting bias estimate d_overbar then removed from the SLA obs on the current day? Why is the bias adjusted in the way it is depending on the observation and model variability? If the obs or model variability is large then the bias is reduced to zero, but in that case do you then assimilate the raw (un-bias-corrected) observations?

  Following the comments of the two Reviewers, we expanded the discussion on this scheme. Altimetry unbias procedures originate by the choice of assimilating anomaly rather than total sea level, thus requiring a field that acts as an average dynamic topography in the observation operator (Lea et al, 2009). The scope of unbiased scheme is twofold: i) reducing the impact of errors in the MDT and ii) improving the corresponding ocean data assimilation. Different unbiased approaches, coming from a non-perfect MDT, are adopted in present global eddy-resolving forecasting systems based either on online variational approaches (Barbosa et al 2024) or simple spatial and temporal filtering in the definition of SLA (Lellouche et al, 2018b). In our present system the MDT is calculated as the SSH average of 1993-2012 insitu-only assimilative run, similarly to Balmaseda et al. 2013, with the same model but at a lower resolution 1/4°. Similarly to the other eddy-resolving systems, we cannot exclude the presence of local drifts coming from a suboptimal definition of the mean circulation. These drifts are systematic biases generated by the difference between the idealized MDT and the practical choice made.
  A hint of local biases can be seen from the long-term biases in SLA shown in Fig 5.b. In our system the unbiased scheme follows the univariate optimal interpolation of the misfits adopted in Storto et al, 2011, and Dobricic 2005, to correct the MDT based on a long-term run. This approach considers a factor, based on optimal weighting that multiplies the misfits similarly to our first term

$\frac{\sigma_\eta^2}{\sigma_\eta^2+\sigma_O^2}$ d , that is the ratio between the model variance and the sum of the model variance and the square of observation error. We also apply a second term $\frac{\underline{d}^2}{\underline{d}^2+\sigma_\eta^2}$ to reduce the bias when the absolute mean initial bias $\underline{d}$ is less than the model variability σ_η. This correction is used online in GOFS16 to identify the daily bias based on the statistics of the previous week and they are removed from the misfits. The calculation of bias is based on the differences between the model maps and altimetry L4 product rather than the misfits to assign a value of the bias in each point. In the future we are considering either to move in the direction of an MDT based on a simulation at 1/16° (Iovino et al 2023) or a hybrid version that includes observation-based MDT (Jousset et al, 2026, Rio et al, 2014).
About the last question, the assimilation of a single SLA observation ultimately depends on the magnitude of the misfit, if it is too large it is removed by the background quality check.

We modify the manuscript as follows:

'SLA observations are treated using a spatial bias correction procedure based on the statistics of the misfits of the last seven days. Different approaches are used in eddy-resolving forecasting systems to deal with altimetry biases, from online variational corrections (Barbosa et al, 2024) to online simple spatial temporal filtering in the definition of SLA (Lellouche., et al., 2018b). The main reason is the presence of errors in the MDT that can generate spurious local drifts thus spoiling the local dynamics and with profound impact on the thermodynamics in the case of SLA observation operator based on the dynamic height formulation. In our system the unbiased scheme follows the structure of a univariate optimal interpolation that was used in literature to correct the MDT from the misfits (Storto et al, 2011, Dobricic 2005). This approach considers two factors: $(\frac{\sigma_\eta^2}{\sigma_\eta^2+\sigma_O^2})$, $(\frac{\underline{d}^2}{\underline{d}^2+\sigma_\eta^2})$ based on optimal weighting of variances, and that multiplies the initial $\underline{d}$ .The first factor $(\frac{\sigma_\eta^2}{\sigma_\eta^2+\sigma_O^2})$ used to correct the MDT is the optimal weightening discussed in Storto et al, 2011, the ratio of model variance and the sum of model variance and observation error squared. A second factor,$(\frac{\underline{d}^2}{\underline{d}^2+\sigma_\eta^2})$,compares the initial bias $\underline{d}$ with the model variability σ_η^2 ,i.e. the bias is reduced when the variability is much larger than the squared bias. These corrections are used online in GOFS16 to identify the daily bias during the model run and are removed from the misfits. The corrections are applied at the validation step as well. In the future we will consider either to move in the direction of a corrected MDT based on a simulation at 1/16° (Iovino et al 2023) or an hybrid version that includes observation-based MDT (Jousset et al, 2026, Rio et al, 2014).

Balmaseda, M. A., Mogensen, K., and Weaver, A. T.: Evaluation of the ECMWF ocean reanalysis system ORAS4, Q. J. Roy. Meteor. Soc., 139, 1132–1161, https://doi.org/10.1002/qj.2063, 2013a

Dobricic, S., 2005: New mean dynamic topography of the Mediterranean calculated from assimilation system diagnostics. Geophys. Res. Lett., 32, L11606, doi:10.1029/2005GL022518.

Lellouche, J.-M.., et al., 2018b: The Mercator Ocean global high-resolution monitoring and forecasting system. In "New Frontiers in Operational Oceanography", E. Chassignet, A. Pascual, J. Tintoré, and J. Verron, Eds., GODAE OceanView, 563-592, doi:10.17125/gov2018.ch20.

Jousset, S., Mulet, S., Greiner, E., Wilkin, J., Vidar, L., Chafik, L., Raj, R., Bonaduce, A., Picot, N., and Dibarboure, G.: New global mean dynamic topography CNES-CLS-22 combining drifters, hydrography profiles and high frequency radar data, Earth Syst. Sci. Data, 18, 2285–2303, https://doi.org/10.5194/essd-18-2285-2026, 2026.

Rio, M.-H., Mulet, S. and Picot, N. (2014): "Beyond GOCE for the ocean circulation estimate: synergetic use of altimetry, gravimetry, and in situ data provides new insight into geostrophic and Ekman currents. Geophysical Research Letters, 41, 8918–8925, https://doi.org/10.1002/2014GL061773

- L176. Suggest removing "instead".

  We agree, thank you

- L178. How is the SST diurnal signal treated when calculating the innovations? Presumably you are not aiming to correct the diurnal signal since the increments are added over a whole day, so how do you avoid diurnal cycle errors from contaminating the innovations?

  We agree with the Reviewer, the treatment of the SST assimilation is probably too simple, in the SST section we already pointed out the possibility to add a specific unbiased scheme similarly to other systems (Barbosa et al, 2024) Presently, the impact of unresolved diurnal variability is mitigated by tuning (increasing) the SST observation error to account for physical processes not fully represented. Practical tuning of the model/observation error is a standard procedure largely adopted during the implementation of a production chain (Desroziers et al, 2005, Oke and Sakov 2008). In this context, we defined a minimum threshold of 0.3degC for all SST observation errors to reduce strong warming or cooling of the water column generated by an over-confidence in the observations.

  We added in the manuscript:

  "We impose a lower bound to the SST observation error by artificially increasing all the errors below 0.3°C to be exactly 0.3°C. This is a practical choice that has been implemented to reduce strong warming or cooling generated by an over-confidence in the observations that contain skin/subskin processes not resolved

by the model. This simple approach to SST assimilation is planned to be revised by including a dedicated ML-assisted SST operator (Broccoli and Cipollonel, 2025).”

Oke, P. R., and P. Sakov, 2008: Representation Error of Oceanic Observations for Data Assimilation. J. Atmos. Oceanic Technol., 25, 1004–1017, https://doi.org/10.1175/2007JTECHO558.1.

Desroziers, G., Berre, L., Chapnik, B. and Poli, P. (2005), Diagnosis of observation, background and analysis-error statistics in observation space. Q.J.R. Meteorol. Soc., 131: 3385-3396. https://doi.org/10.1256/qj.05.108

- L183. Why do you think in this way rather than averaging to create superobs?

  We agree with the Reviewer that creation of superobs is a possible alternative. However, as a general idea, we prefer not to alter the original observations, and instead apply thinning or subsampling which avoid generating “artificial” observations Yet we agree the two approaches are broadly equivalent; our choice mainly reflects a preference for preserving direct traceability to the original measurements.

- L187-189. This description of the error variance would fit better with the earlier error covariance descriptions, e.g. around L162.

  We delete these sentences that are now included in the answer to L162.

  L189. What do you mean here by the horizontal length-scales being scaled to the high-resolution grid?

  We move the description above and delete this part. See answer to L162.

- L191-192. Why do you need to include a nudging scheme for SST during the analysis when you are assimilating SST observations? Is the SST nudging applied in the forecast?

  We agree with the Reviewer that a better approach can be considered for SST assimilation that is presently done both at DA and at model step. At DA step, the assimilation of SST is performed together with insitu temperature to provide incremental corrections that are coherent between surface and subsurface. However, the ingestion of such increments into the model temperature is not able to fully constrain the heat flux during the NEMO integration and requires the addition of a dedicated heat flux correction to make them as realistic as possible. In Broccoli and Cipollone (2025) the authors developed a more complex operator that can be further used also to correct the flux into the model at daily scale.

  We added the discussion as follows:
  “The analysis scheme of GOFS16 includes a nudging scheme to correct the surface heat and freshwater fluxes. DA corrections to surface model temperature does not sufficiently constraint the heat flux in the NEMO model and we include a nudging

scheme to have a realistic air-sea exchange towards the gridded sea-surface temperature analyses provided by NOAA that uses infrared retrievals from AVHRR instruments, drifter and buoy in-situ measurements (Reynolds et al., 2007). Such a solution is going to be revised in the future, following a more complex SST operator (Broccoli and Cipollone, 2025).”

- L193. Is the SSS nudging scheme included in the forecast?

  No, forecasts are forced only by atmospheric and runoff boundary conditions.

- L201. “...reduce the artificial spatial correlation among observations that are inevitably introduced at the length scale of grid spacing”. Including observations on the model grid at this high resolution will surely give you observations which do contain spatially correlated errors, and more so than if you thinned to lower resolution model grids.

  The sentence was slightly ambiguous, and we rephrased it to clarify the meaning. The DA imposes artificial correlation at the length scale of grid spacing. This implies that you cannot have multiple observations for a single grid box. The final number of independent observations depends on the resolution of the grid for dense observation. This can be also discussed theoretically following Li et al. (2015). Suppose that our B matrix can be split in two uncorrelated terms, B=B_L + B_S, where S is the length scale of the grid spacing and L represents everything above that length scale; then the final correction can be seen as

$$x^a = x^b + B_L H \frac{1}{HB_LH^T + (R + HB_SH^T)} d$$

  The H B_s H^T term acts to correct the error covariance matrix R at the scale S. To limit this problem of multiple observations, we use a thinning procedure that retains only one observation (the one closest to the middle of the assimilation window) per grid box.
  We rephrase the sentence:

  “The GOFS16 system is designed to share the same 1/16° grid for both the model and the DA scheme, although this choice is computationally intensive and not strictly required (Lellouche et al, 2018a). The use of the same grid for model integration and DA maximizes the amount of information that can be extracted from the datasets, especially for observations that are dense on the grid (Cipollone et al, 2020). “

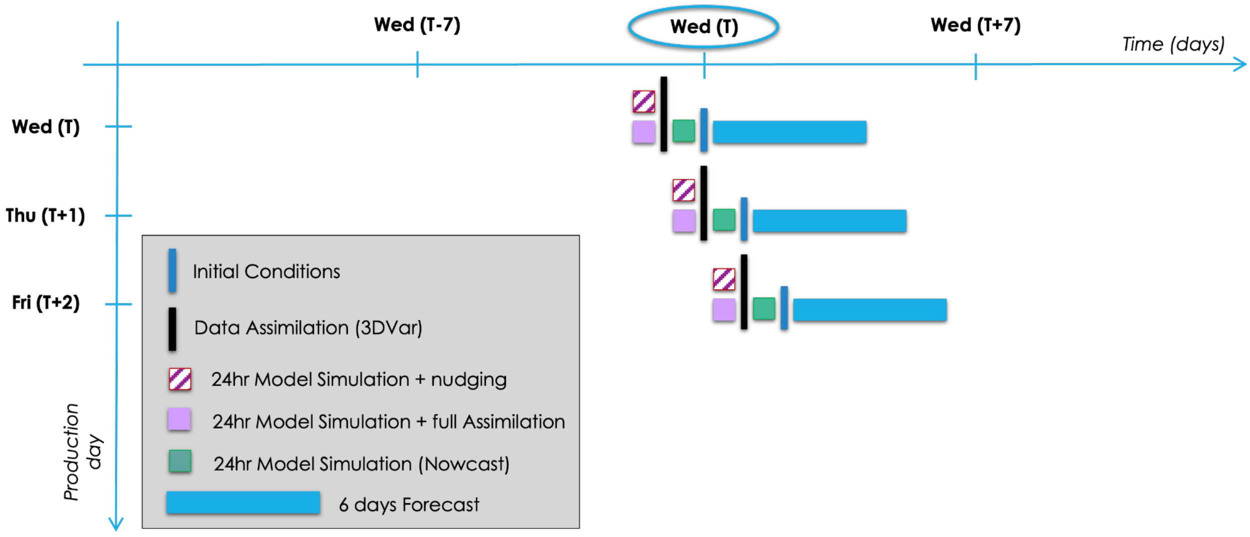
  Li, Z., J. C. McWilliams, K. Ide, and J. D. Farrara, 2015: A multi-scale variational data assimilation scheme: Formulation and illustration. Mon. Wea. Rev., 143, 3804–3822, https://doi.org/10.1175/MWR-D-14-00384.1.

- L202. Typo “explicitly”.
  Done

- L213. By “inspected” do you mean they are manually inspected as part of the operational production/dissemination? Is this done every day including weekends?
  By “inspected”, we mean that the operational outputs are manually checked by the operational team before dissemination to verify that no anomalies or unexpected issues are present. This manual inspection is routinely performed on working days. During weekends and public holidays, the production chain is automatically monitored through operational control procedures, while manual inspection is carried out only if anomalies are detected or if intervention is required.

- Fig 1. Since every day the pattern is the same, perhaps you could zoom in on only two or three of the days so that it is easier to see the details of each cycle.

  We have produced a new Fig.1 as suggested.



- L221-222. I don’t agree that this procedure is analogous to nowcasting in the atmosphere. In nowcasting, high resolution observations are used with the latest analysis to provide a very short-range estimate, whereas here there are no observations included.

  We agree with the Reviewer that in the NWP community there is a specific definition of nowcast, on the other hand in ocean prediction such definition is somehow relaxed. The definition of daily nowcast has been used in literature to define a period that include present time, and it spans over 24h of more (Castillo-Trujillo et al, 2023, Mehra et al 2010) and this terminology is also used in available website (NOAA, 2026). Following the Reviewer’s comment, we remove the link to the hourly nowcast and add these references:

  “Following the definition of daily nowcast for operational ocean systems (Castillo-Trujillo et al, 2023, Mehra et al, 2010) we refer to the model output between T-24h to T as daily nowcast .”

Carolina Castillo-Trujillo, Alma, Young-Oh Kwon, Paula Fratantoni, et al. «An evaluation of eight global ocean reanalyses for the Northeast U.S. Continental shelf». Progress in Oceanography 219 (dicembre 2023): 103126. https://doi.org/10.1016/j.pocean.2023.103126.

Mehra, A. and I. Rivin, 2010: A real time ocean forecast system for the North Atlantic Ocean. Terr. Atmos. Ocean. Sci., 21, 211-228, doi: 10.3319/TAO.2009.04.16.01(IWNOP)

NOAA ,2026: Operational Forecast System (OFS) Frequently Asked Questions (FAQs), https://tidesandcurrents.noaa.gov/ofs/ofs_faq.html

- L231. Why did you choose to use GFS atmospheric forcing rather than ECMWF?

  GFS atmospheric forcings are a reliable, robust and open product that is used by many systems and is suitable for our scope. We agree with the Reviewer that ECMWF analysis and forecasts seem to have better skill scores, but they are not open and free. In 2026, a new low-resolution version of ECMWF forecasts has become freely available. We are planning to test this new dataset in the future.

- The spatial and temporal resolution of the forcing is very low compared to the ocean model resolution.

  As answered to the above comment, the choice of using GFS instead of ECMWF at the very beginning of the operationalization of GOFS16 was due to the fact that GFS was the only open and free atmospheric forcing available at that time. Since April 2026 ECMWF has made freely available a new low-resolution version of atmospheric forecasts and it is in our plan to assess them.

- L238. The EN4 objective analysis is not available close to real time so with what delay is the SSS field available and how do you deal with the time offset?

  The nudging of SSS is not meant to assimilate observations but rather to avoid drifts in the model salinity coming from prescribed atmospheric forcings that can generate unbalanced freshwater flux locally. The approach is similar to what is done in the OMIP exercise to reduce possible unrealistic effects over the overturning circulation.
  Rather than using directly the monthly climatology of the chosen month J, we compute the SSS anomaly of the previous month (J-1) and add it to the climatology of month J, so

  $$SSS_J = SSS^{CLIM}_J + (SSS_{J-1} - SSS^{CLIM}_{J-1})$$

  We prefer to use SSS_J instead of SSS^{CLIM}_J to be more consistent with the present ocean state. SSS_J is used to produce the best estimate, with a timescale of 1 month. It is switched off in forecasting mode.

We expand the description:

"SSS nudging is not meant to assimilate observations but rather to avoid large drift in the model salinity and reduce unrealistic effect on the overturning circulation. Rather than using a climatology, for the month J the following formula is used:

$\mathbf{SSS_J = SSS^{CLIM}_J + (SSS_{J-1} - SSS^{CLIM}_{J-1})}$

The reference for month J, is constructed based on the climatology of the month J plus the anomaly of the month J-1."

- L247. Is the first guess field a daily average, instantaneous in the middle of the day or something else? How is the interpolation of the model values done in the vertical for profile observations?

  First guess fields are daily mean, interpolation scheme is bilinear horizontally and linear in the vertical. We added this information into the description.

- L252. Do you mean that "the increments are added to the model fields with an ..." and not added to the restart fields?

  Correct, the increments are added through IAU at each step directly into the model variables, we clarify the point in the text:

  "The model is then integrated a second time for the same time-period (T-48h to T-24h) with the DA increments added during the model integration directly to the variable field with an incremental analysis update (IAU) algorithm".

- L255-258. You discuss in situ observations being available close to real time, but satellite SST observations are also available within a few hours.

  The Reviewer is correct, we remove the reference to insitu and generalize it to all the observations:

  "[...] some observations are already available for the interval T-24h to T-4h. As a potential further improvement, an additional cycle of simulation/assimilation could be added in the future to assimilate these available data, provided that their inclusion demonstrates a significant benefit to the forecasting skill score."

- L259. Which product data are hourly and which daily? Or do you provide all data at the full 3D resolution at hourly and daily frequencies?

  The system works mainly with daily means. We produced some hourly datasets in smaller regions, but the global 3D hourly mean is not sustainable by our present archive system. We decided to remove the comment on hourly to not generate confusion.

- L262. "0.33 wall clock hours". Do you mean that you run the two 1-day NEMO runs (background trajectory and IAU) at 1/16 degree resolution and the OceanVar step all within 20 minutes elapsed time?

  The production of the best estimate takes one hour with one cycle of assimilation and two cycles of NEMO for one day.

- L274. Is the assessment against observations carried out using the model best estimate fields? In that case, please clarify that the observations have already been assimilated and are not independent.

  Following the Reviewer suggestion we added a specific comment:

  "This tool focuses on the skill score of the best estimate fields that can be considered only semi-independent. The best estimate in fact includes corrections from observations disseminated within 24h with respect to the reference date, while late coming observations are not included. "

- L319. Did any of the bug fixes or continuous improvements have a large enough impact to describe what they are here?

  We agree with the Reviewer that the comment is probably too general. The configuration presented includes all the changes adopted through the years (assimilation of new altimetry observations, new quality checks, new data format etc). One of the first relevant changes concerns the adoption of IAU algorithm rather than direct ingestion (DI) that contributed to stabilizing the production of best estimate at the beginning. The implementation of a double integration (without and with increments) to generate the analysis (Figure 1) was a second relevant improvement. Pre-processing and quality checks also include some modifications in the treatment of quality flags (we decide to assimilate only good observations from near-real time datasets). The vertical check on Argo profiles, described in L185, was implemented to eliminate dubious corrections coming from profiles that have only good observations at depth. Several consistencies check on SIC and SST were included to avoid the use of products disseminated without proper consistency checks (NOAA SIC and SST in several occasions). Near real time insitu data disseminated by CMEMS significatively changed formats several times. We also updated the number of available altimetry satellites and the passage from metopA to metopB SST observations.

  We expand the description in the text accordingly:

  "The configuration presented in this paper includes all the changes adopted through the years to maintain/improve the system. One of the first relevant changes concerns the adoption of IAU algorithm rather than direct ingestion (DI) that contributed to stabilizing the production of best estimate at the beginning. The implementation of a second cycle of NEMO to produce the analysis (see Figure 1) is another relevant improvement. Vertical checks on Argo profiles, described

above, were crucial to remove dubious corrections from profiles that have observations flagged as "good" only at depth. We also add observations from new satellites disseminated throughout the years."

- L325. What is the cold bias in SST due to?

  Looking at Fig 4, we can associate the cold bias mainly to a specific region, that is the Eastern North Pacific. In the last decade, starting from 2014 several heatwaves covered that area, known as the North Pacific warm blob (Shi et al, 2024) or https://oceanview.pfeg.noaa.gov/projects/mhw/videos
  The impact of these heatwaves is not included in our BEC and this means that we constantly underestimate the model variability over the region, i.e. DA corrections are not as large as they should be. Other possible causes could be identified by ad hoc experiments, and it is in our plans to reduce the cold SST bias.

  Shi, J., Huang, H., Fedorov, A.V. et al. Northeast Pacific warm blobs sustained via extratropical atmospheric teleconnections. Nat Commun 15, 2832 (2024). https://doi.org/10.1038/s41467-024-47032-x

- L331. What is happening during 2023 in Fig. 3a? There's a different behaviour in the bias during May-Aug 2023 compared to 2022 with a positive rather than negative bias evolution, particularly in the 5-day forecast. Also, why is there a gap in Oct 2023?

  Figure 3 shows GOFS16 skill-scores against the SST/SLA observations that are collected and released by GODAE/OceanPredict with a delay of five days with respect to present time. The dissemination activities are supported by experienced Production Units (PUs), such as UKMO or MOI, but they are not funded directly by OceanPredict. Therefore, PUs cannot guarantee the prompt recovery after failures or the temporal homogeneity of the timeseries, especially in case of changes in the dissemination steps that are not under their control. For example, in October 2023 there was a lack of data that PUs is not able to fill yet. About Figure 3 RMSE and BIAS of SST shows a marked boreal seasonality that depends on the distribution of the SST drifters exploring mostly the northern hemisphere rather than the southern one.
  The figure below shows the spatial SST BIAS in the period May-July 2022 and May-July 2023 for the best estimate and the 5th day of forecast. Observations are cumulated over 1°x1° bin. The large negative bias in 2022 is generated by the drifter prolonged exploration of the severe North Pacific warm blob (Shi et al, 2024) that is less impacting in 2023.

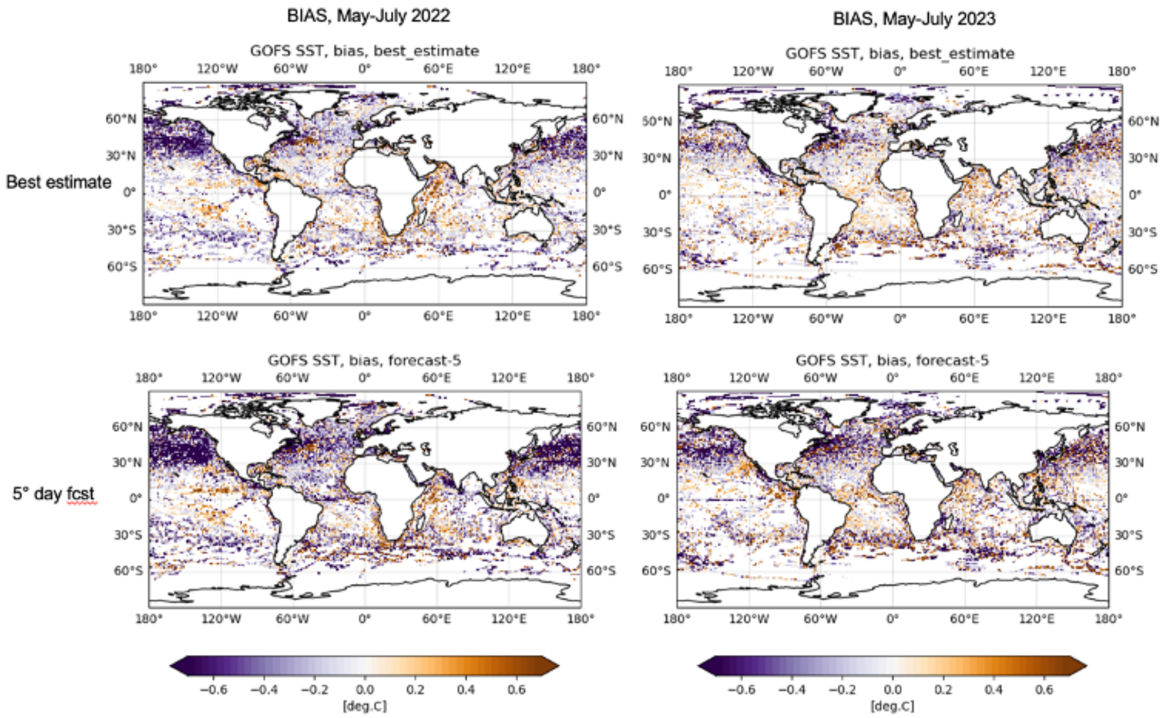


Figure 4. SST BIAS for best estimate (first row) and the 5th day of forecast (second row) in May-July 2022 (first column) and May-July 2023 (second column).

- L340. The bias values are negative, not positive as stated in the text.

  We thank the reviewer for pointing out this mistake. The SST bias values are negative, and the text has been corrected accordingly in the revised manuscript: "The median bias approximately doubles in magnitude after three forecast days, changing from about −0.1°C in the best estimate to about −0.2°C, after which it changes only gradually."

- L352. Any idea why there are degradations in the N. Pacific?

  See also our response to L335. Starting from 2014 the North Pacific underwent several heatwaves called "warm blob" that repeated during the latest years. Such phenomena are of great interests for the ocean community (Xu et al 2026, Shi et al, 2024, Ren et al 2023) and there are several "blob tracker" showing the anomaly year by year:
  https://oceanview.pfeg.noaa.gov/projects/mhw/videos
  https://www.integratedecosystemassessment.noaa.gov/regions/california-current/california-current-marine-heatwave-tracker-blobtracker

  Ren, X., Liu, W., Capotondi, A. et al. The Pacific Decadal Oscillation modulated marine heatwaves in the Northeast Pacific during past decades. Commun Earth Environ 4, 218 (2023). https://doi.org/10.1038/s43247-023-00863-w
  Xu, T., Newman, M., Shin, SI. et al. Persistent Northeast Pacific marine heatwaves are sensitive to the seasonality of tropical and North Pacific dynamics. Commun Earth Environ 7, 528 (2026). https://doi.org/10.1038/s43247-026-03442-x

Shi, J., Huang, H., Fedorov, A.V. et al. Northeast Pacific warm blobs sustained via extratropical atmospheric teleconnections. Nat Commun 15, 2832 (2024). https://doi.org/10.1038/s41467-024-47032-x

- L358. Why is there a bias in the SLA differences, despite the bias removal process? Could this be related to the drift in the global mean SSH you mentioned earlier?

  We do not apply directly the unbiased scheme on the SLA dataset used for validation. The unbiased process originates from the comparison between the model and L4 altimetry maps to equally balance all different ocean regions and removing any dependency on the observation density. The small global bias that is present can probably come from the difference in the datasets.

- L366. "…with barely any differences among them." Not sure what you mean here.

  We agree that the original sentence was unclear. We have rephrased it to clarify that the spatial patterns of SLA bias and RMSE patterns are very similar in the best estimate and nowcast products.
  "The bias and RMSE spatial patterns of SLA confirm that the largest errors are confined to mesoscale-rich regions in both the best estimate (Fig. 5) and the nowcast (not shown), with negligible differences elsewhere."

- Fig. 7. The climatology SST RMSEs on this figure are very different in magnitude to the ones in Fig. 6 of Blockley et al., 2014, especially for the N. Pacific where you have a value of over 3 degrees whereas Blockley et al. have it at about 1.2 degrees. Can you explain this large discrepancy?

  Following the discussion to the answer L352, the North Pacific has undergone repeated summer heatwaves in the last years that are not present in the climatological fields of Blockley et al., 2014. Such climatology is defined as the average value of surface temperature fields calculated from EN3 dataset between 2004 and 2008. In the figure below we compare the annual SST anomaly for 2012 (one year covered by Blockley) and 2023. The anomaly is calculated against the 2004-2008 mean from EN4.2.2 datasets that is the latest version of the same UKMO-based dataset employed by Blockley. SST from 2023 is much warmer than 2004-2008, with large regions exceeding 2 degrees, while 2012 was much similar. The discrepancies with Blockley mainly come from the period analyzed, the present state of the ocean in 2023 is farther from the climatological field than the ocean state in 2012 analyzed by Blockley.

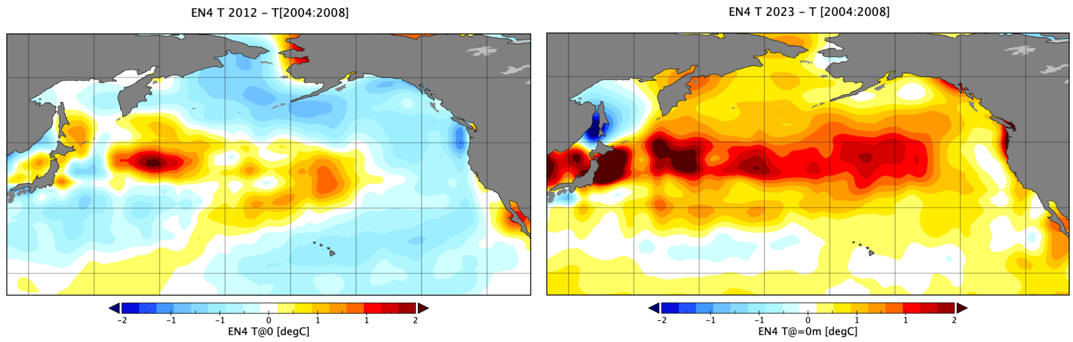


Fig. 5. Left panel: Difference between yearly average SST for 2012 and reference SST from Blockey et al 2014 (average SST between 2004-2008) extracted from EN4. Right panel: Difference between yearly average SST for 2023 and reference SST from Blockey et al 2014 (average SST between 2004-2008) extracted from EN4.

- L409. Why do you do the assessment down to 1500 m depth and not the lowest depth of Argo floats at 2000 m?

  Figure below extends the analysis of global T and S vertical RMSE profiles for the best estimate and forecast fields down to 2000m. In the lowest vertical level (around 2000m) the impact of DA is minimal, infact the best estimate and forecast RMSE are similar around 2000m. At such depth 3D damping towards climatology starts to be more relevant than DA itself and in the present system no 3D damping is applied. Therefore, we preferred to keep the discussion in the first 1500m, we will add a comment on the paper about this point.

  "Validation is limited to 1500m. Below 1500m DA corrections seem to be marginal and the skill scores of the best estimate are like the forecast ones (see discussion with Review). One of the possible reasons is that the large amount of surface data reduces the sensibility of the minimizer to the sparse observations at depth."

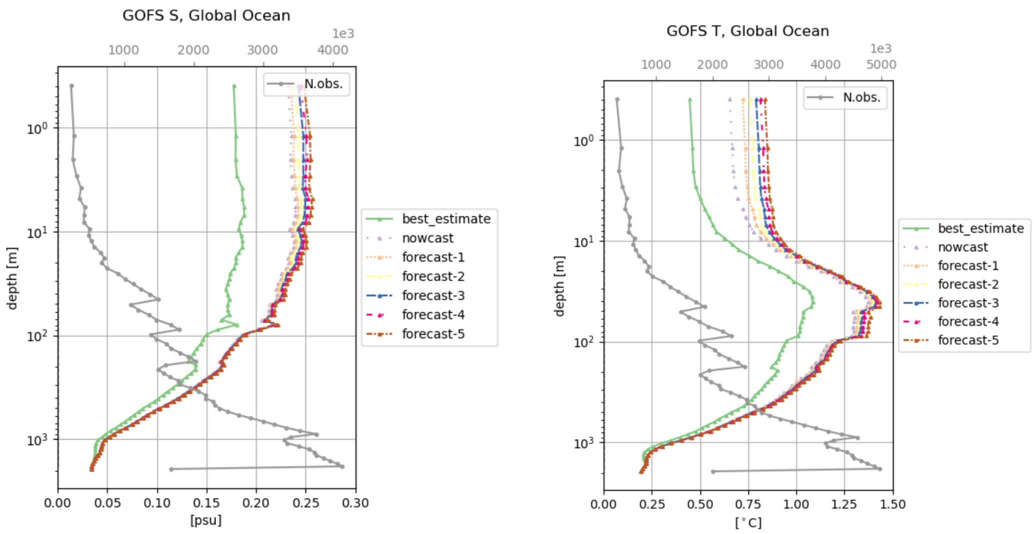


Fig. 6. Vertical RMSE profile for Temperature (left) and Salinity (right) for best estimate and forecasts extended down to 2000m, averaged over 2022-2023.

- L413. What happens to the RMSE in T and S at the start of 2023 where there is a discontinuity in the RMSE time-series plots of Fig. 9, including also a jump in the salinity bias?

  Following the answer to L331, the validation datasets are sustained by experienced product units (PUs) such as UKMO and MOI (recently ECCC too) but are not funded directly by any project. Therefore, the PUs cannot guarantee the temporal homogeneity of the time series, which is also affected by changes in each entity's dissemination chain — changes that are not under the control of the PUs. Looking at the number of observations (FIg. 9) it can be noted that for T observations there have been several jumps during the period under investigation. About salinity the number of observations had a drop by halving the available observations. Those changes affect the BIAS or RMSE as well.

- Fig 9 caption. Part b is for T profiles not SST. Also a typo "nowvast".

  We thank the reviewer for pointing out these errors. Indeed, panel b shows temperature statistics, not SST, and the typographical error ("nowvast") has also been corrected in the revised manuscript.

- L433-435. You claim here that it is the high vertical resolution is giving improvements, but the maximum of the temperature profile RMSE is not any lower than more recently published results from other systems with fewer vertical levels.

  The Reviewer is right. The maximum of the temperature profile RMSE is about 1°C, in the range of the other systems (Ryan et al., 2015), despite the GOFS16 higher vertical resolution. We rephrased the sentence accordingly.

  "The maximum RMSE of the best-estimate temperature profile is about 1°C at the thermocline depth, similar to other systems (Ryan et al., 2015) and despite the GOFS16 higher vertical resolution."

- L435-436 and also L454. Do you have evidence that it is the vertical mixing in the model causing the large degradation in temperature and salinity? Other explanations could include an unbalanced initial state or overfitting of the profile observations in the analysis.

  The same point has been raised by the second Reviewer. Looking at Fig. 10, it is a matter of fact that the model is not able to sustain the improvements brought by DA in the best estimate, similarly to other old (Ryan et al 2015) and new systems (Smith et al 2026). The reason can be the inconsistency of DA increments with the model dynamics. Another reason can be that theDA increments are too localized and do not sufficiently spread the information around. This is another point to be investigated, we added the comment below:

  "[...] suggesting that the model vertical mixing is not able to maintain an accurate initial state. Other possible reasons can span from an inconsistency of the

increments with the model dynamics to increments that are too localized and do not spread the information around."

- L446. I think you mean Fig. 10b?

  We apologize for this incorrect figure reference. It has been corrected to Fig. 10b in the revised manuscript.

- L474. Argo profiles are now provided close to real time so it would be useful to provide evidence if you think this is not the case.

  We rephrased the sentence that was not clear enough. The first group of insitu observations for a date T arrives with a delay of a few hours. However, the same date T is continuously populated by observations coming late. GOFS16 production system uses observations for date T at T+24h, therefore it is only using part of the datasets and does not assimilate a non-negligible fraction of the total observations for date T . We changed the text as follows:

  "[...] salinity, and it uses for the date T all the data available at T+24h, therefore missing a non-negligible part of the data that arrives later than 1 day. Those are usually more critical for the subsurface levels where observations are less abundant than in the surface ones."

- L496. The observations of drifter-derived velocities from CMEMS do have some temporal filtering which allow a fairer comparison with models.

  The methodology employed in CMEMS is referenced in the associated QUID, https://documentation.marine.copernicus.eu/QUID/CMEMS-INS-QUID-013-044.pdf. They mostly rely on Elipot et al 2016 designed for the GDP with the addition of a correction for wind slippage in dense areas. Such a product seems to be in evolution, and we plan to use it in the next future, but the main differences described in the text remains.

  Elipot, S.; Lumpkin, R.; Perez, R.C.; Lilly, J.M.; Early, J.J.; Sykulski, A.M., 2016: A global surface drifter data set at hourly resolution. J. Geophys. Res. Oceans, 121, 2937–2966. https://doi.org/10.1002/2016JC011716

- L497. The NEMO model should include near-inertial oscillations (see for example Waters et al., 2024; https://doi.org/10.3389/fmars.2024.1383522).

  We thank the reviewer for this comment and for pointing us to this relevant reference. We agree that the original wording was imprecise. Near-inertial oscillations are represented by NEMO when the model configuration and forcing allow their generation. Waters et al. (2024) further demonstrate the importance of accurately representing and preserving this high-frequency variability within an operational data assimilation framework. Our intention was not to imply that near-inertial oscillations are absent from the model dynamics, but rather that they are

not adequately represented in the daily-averaged velocity fields used in the present comparison against instantaneous drifter observations, therefore contribute to the mismatch with instantaneous drifter observations. Since the near-inertial period is of the order of one day over much of the mid- and high-latitude ocean, a substantial fraction of this variability is filtered out by the daily averaging, even though it may be present in the underlying model solution. We have revised the sentence accordingly.
We changed the original lines 495-497 "The direct comparison is largely biased by the difference in time-averaging between model (daily) and observations (instantaneous), not to mention several phenomena that are not resolved by the model such as Stokes drift, tidal currents, near-inertial oscillation, etc." into
"The direct comparison is largely affected by the difference in temporal averaging between the model (daily-averaged currents) and the observations (instantaneous drifter measurements), as well as by several processes that are either not represented by the model, such as Stokes drift, or are not adequately represented in the time-averaged model output used here, such as tidal currents and near-inertial oscillations (Waters et al., 2024)."

- L506. "Data are vertically averaged". Do you mean the GOFS16 data?

  We rephrase the wording, the OSCAR maps are provided as "an average over an assumed well-mixed top 30 m" (https://podaac.jpl.nasa.gov/dataset/OSCAR_L4_OC_FINAL_V2.0)

  "Daily maps of meridional and zonal currents are provided at 0.25° grid representing the average subsurface circulation of an assumed well-mixed top 30 m"

- L508. Is the correlation an anomaly correlation as stated in the caption of Fig. 12? If so, what climatology is used to calculate the anomalies?

  We thank the reviewer; the correlation is an anomaly correlation that follows equation 4. We include the word "anomaly". The monthly climatology comes from the average over the same two years.

  "[...] focusing on the biases and anomaly correlation (Eq.4) for both zonal (U) and meridional (V) components. The monthly climatology comes from the average over the same two years."

- Fig. 13. Is this the GOFS16 best estimate SST field? Perhaps you could also include the field from before the TC (e.g. 30th August) to show the change in SST more clearly.

  Yes, this is the GOFS16 best-estimate SST field.
  We thank the reviewer for this suggestion. During the analysis of Hurricane Larry, we examined the daily evolution of the GOFS16 SST from before to after the storm passage. We ultimately chose to show only the 11 September field in the manuscript as it clearly highlights the cold-wake signature left by the hurricane across most of the basin. Prior to the passage of Larry, the pre-storm field exhibits

the typical expected large-scale spatial distribution of temperature over the North Atlantic, with no distinct features associated with the approaching storm. We have therefore decided not to modify the manuscript, as Figure 13 effectively conveys the intended message. For completeness, we include below the GOFS16 daily best-estimate SST fields every two days from 1 September, illustrating the evolution of the upper-ocean thermal response before, during, and after the passage of Hurricane Larry, to confirm progressively cold-wake development along the hurricane track.

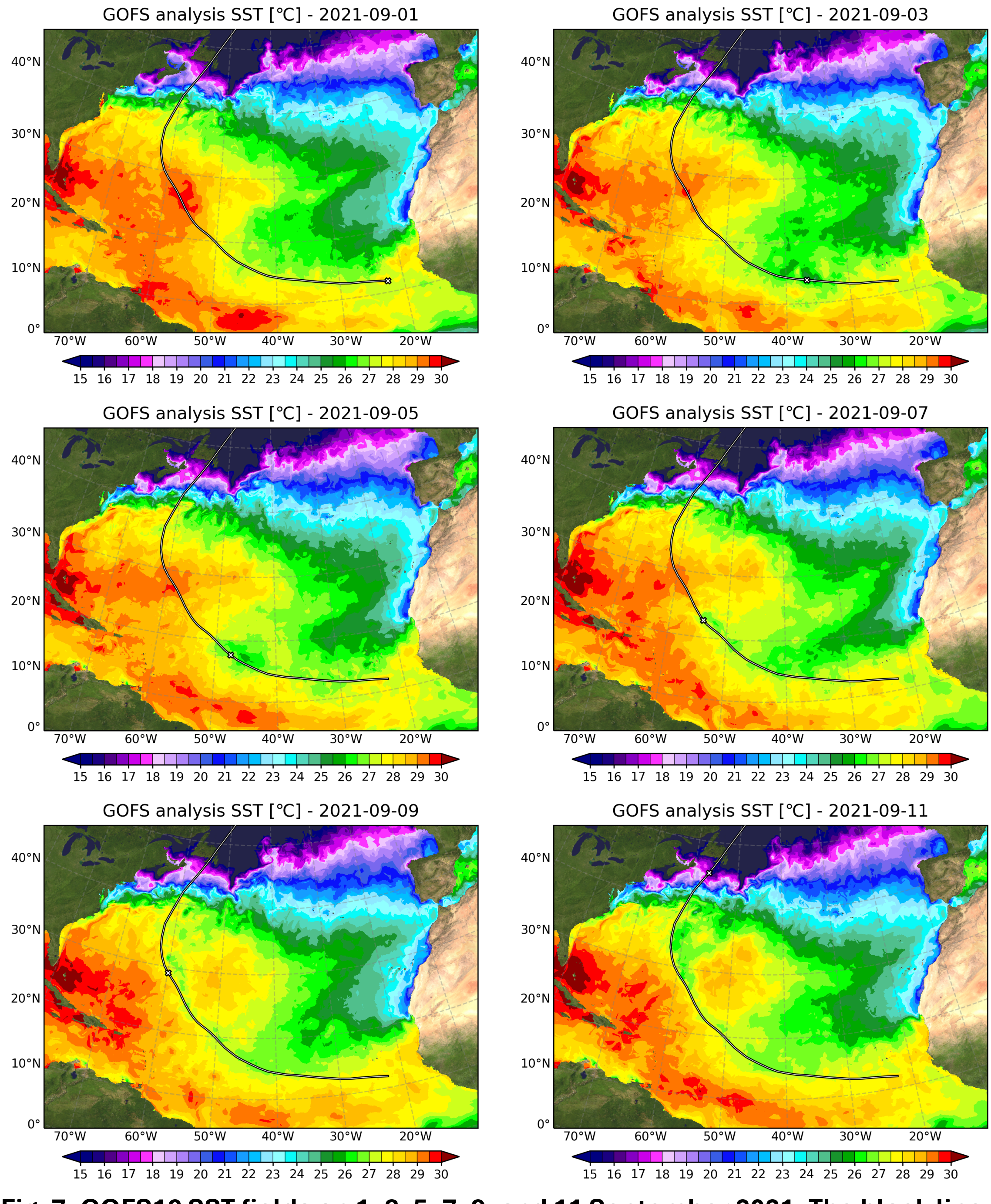


Fig. 7. GOFS16 SST fields on 1, 3, 5, 7, 9, and 11 September 2021. The black line indicates the hurricane Larry track, and the crossmarks the storm position corresponding to the date shown in each panel.

- L660. You could also mention here more sophisticated data assimilation algorithms such as 4DVar and hybrid ensemble/variational methods which could use the limited observations in a better way.

  We expand the text including additional methodologies. We believe the hybrid methods could have a big role in the evolution of the forecasting system, while the 4DVar probably shows its best application for reanalysis production when the analysis of a specific period can benefit from the availability of past and future observations at the same time. We modify the text as follows:
  "The other way might be offered by coarser resolution ensemble forecasting methods which should complement the current effort on the higher-resolution forecasting system. The development of hybrid schemes that mix climatological B with ensemble-derived "error of the day" are promising (Lea et al, 2022) as well as possible extension to 4DVar (Fenty et al 2017, Storto et al 2019)."

  Lea, D.J., While, J., Martin, M.J., Weaver, A., Storto, A. & Chrust, M.(2022): A new global ocean ensemble system at the Met Office: Assessing the impact of hybrid data assimilation and inflation settings. Quarterly Journal of the Royal Meteorological Society, 148(745), 1996–2030. Available from: https://doi.org/10.1002/qj.4292

  Fenty, I., Menemenlis, D. & Zhang, H. (2017): Global coupled sea ice-ocean state estimation. Clim Dyn 49, 931–956 https://doi.org/10.1007/s00382-015-2796-6

  Storto, A., Oddo, P., Cipollone, A., Mirouze, I., Lemieux-Dudon, B. (2018): Extending an oceanographic variational scheme to allow for affordable hybrid and four-dimensional data assimilation, Ocean Modeling, 128, 67-86, https://doi.org/10.1016/j.ocemod.2018.06.005

- L696. Multi-scale methods don't have to be done in two steps, see e.g. Mirouze et al., 2016 (https://doi.org/10.3402/tellusa.v68.29744).

  We believe the Reviewer intended to discuss line 669 (rather than line 696 that corresponds to the reference section). We thank the Reviewer and expand the sentence following also the last comment of the second Reviewer:

  "In addition, we propose to start investigating multi-scale analysis technique (Li et al., 2015; Chamberlain et al., 2021) in which two assimilation steps correct the ocean at separate length scales: the first assimilation step is performed on a coarser grid to ingest sparse insitu observations. The second step employs a higher resolution grid and assimilates dense datasets to generate increments at smaller scales. Similar multi-scale analysis can be also performed in one single step (Mirouze et al, 2016b)."

  Mirouze I, Blockley EW, Lea DJ, Martin MJ, Bell MJ. (2016b). A multiple length-scale correlation operator for ocean data assimilation. Tellus A 68: 29744.