RESPONSE TO REVIEWER #3

This manuscript describes an offline implementation of a well-stablished biogeochemical model, the Fennel model, and offers an evaluation of its performance against the same biogeochemical model coupled online to an ocean model configured for the Gulf of Mexico, from which physical variables are extracted to run the offline case. The offline model performance is evaluated considering different time steps. In addition, the role of three vertical mixing scheme closures is explored.

This offline model is a valuable tool that reduces the computational cost of running biogeochemical models. It is also appreciated that it is publicly available to the community. After evaluating the revised manuscript version as well as the replies to initial reviewers, I find it well-structured, relevant and fits well within the scope of the journal, for which I recommend it for publication. I have, however, a few comments that aim to enhance the clarity of the manuscript.

(Line numbering in comments below refers to the latest version of May 17 with tracked changes)

RESPONSE: We would like to thank the reviewer for the constructive and detailed comments, which have helped us improve the clarity and robustness of the manuscript. Below we address each of the points raised and indicate the corresponding changes made in the manuscript. Line numbering in refers to the updated revised version with tracked changes from 30th June.

MAJOR COMMENTS

i) I am confused about how biogeochemical tracer diffusion is implemented in the offline model regarding the three points below, which should only require some clarifications in the text.

• Authors mention the usage of three different mixing schemes, but these are vertical mixing schemes while nothing is said about horizontal mixing at subgrid scale. Please add a line or two about it.

RESPONSE: In our setup, harmonic horizontal mixing of velocities and tracers is applied along geopotential surfaces using the 'UV_VIS2', 'TS_DIF2', 'MIX_GEO_UV', and 'MIX_GEO_TS' options. We also employ the 'TS_MPDATA' advection scheme – recommended by Thyng et al. (2021) – to minimize numerical diffusion. The 'DIFF_GRID' and 'VISC_GRID' options enable grid-dependent variations in lateral diffusivity and viscosity. All simulations use this configuration to ensure full consistency and comparison between experiments.

We have included these lines for clarity in the Methods section.

L132-135: *Harmonic horizontal mixing of velocities and tracers was applied along geopotential surfaces, together with the TS_MPDATA advection scheme to minimize numerical diffusion (Thyng et al., 2021). Grid-dependent diffusivity and viscosity were also enabled via DIFF_GRID and VISC_GRID. All*

configuration files, including exact parameters for each mixing scheme, are available for full reproducibility (Crespin, 2025b).

All CPP options for each simulation are publicly archived on Zenodo (DOI: 10.5281/zenodo.14930137).

• In L126 authors refer to the GLS scheme but this is a generic scheme. Authors should specify if it was implemented as k-kl (a specific case of which is the MY25 scheme), as k-e or k-w as indicated in Warner et al. (2005), as well as the parameters used (or refer to a publication that uses the same parameters) for reproducibility of results.

RESPONSE: We implement the GLS scheme in its k–ε configuration, following Warner et al. (2005). The key exponent parameters are set to GLS_P = 3.0, GLS_M = 1.5, and GLS_N = –1.0. All configuration files for online and offline simulations (including exact parameters for each field), are openly available on the Zenodo repository (DOI: 10.5281/zenodo.14930137; https://zenodo.org/records/14930138), as well as input files, allowing full reproducibility of the study.

We have specified this in the manuscript as follows:

L129-130: *The GLS scheme in our simulations corresponds to the k–ε configuration (Warner et al., 2005), defined by the exponent values 'GLS_P' = 3.0, 'GLS_M' = 1.5, and 'GLS_N'= -1.0.*

• L161-L168. Authors mention that AKs, AKt, and AKv are used to force the offline model in GLS and MY25 mixing schemes but not the LMD one. I find this very confusing. The aim of using a vertical mixing scheme is to obtain a vertical diffusion coefficient to use for subgrid-scale vertical mixing in the biogeochemical tracer equations (e.g., eq. 9 in Fennel et al. 2022). Which of the three (AKs, AKt, and AKv) is being used to diffuse vertically biogeochemical tracers in GLS and MY25 mixing schemes? Why do you need the TKE and the generic length scale output in the offline model if the vertical diffusivity is already provided? Additionally, the text reads as no vertical diffusivity is used in the LMD scheme case, is this correct?

RESPONSE: We thank the reviewer for this important question and apologize for the confusion. All details on the use of the AKXCLIMATOLOGY and MIXCLIMATOLOGY flags are documented in Thyng et al. (2021). Below we clarify their behavior in the offline implementation:

With these flags enabled (GLS & MY25 offline simulations):

- The offline model reads the pre-computed fields (AKs, AKt, AKv, TKE, and GLS) directly from the online parent simulation at 3-h intervals.
- Of these, AKt (vertical temperature diffusivity) is used for subgrid-scale vertical mixing of all passive tracers, including biogeochemical variables. AKs (vertical salinity diffusivity) and AKv (vertical viscosity) influence only salinity and momentum, respectively, and do not modify passive tracer

advection (Thyng et al., 2021). TKE and GLS are imported for diagnostic consistency.

Without climatology flags (LMD):

- The offline model recomputes its own vertical mixing coefficients (AKs, AKt, AKv) internally from the LMD closure, using the exact same parameter values as in the online run.
- Thus, vertical diffusivity remains active, computed on-the-fly rather than read in.

All three configurations therefore supply the necessary vertical diffusivity for biogeochemical tracers: via climatology read-in for GLS/MY25, and via internal computation for LMD. We retain AKt, AKs, AKv, TKE, and gls in the GLS/MY25 cases solely to adhere to the original offline framework (Thyng et al., 2021) and for diagnostics. In Thyng et al. (2021) they report that only AKs had a minor impact on tracer accuracy, while Akt, Akv, TKE, and GLS did not.

We have updated the manuscript for clarity:

L171-182: *For the GLS and MY25 simulations, the offline model was forced by additional vertical mixing parameters – namely vertical salinity diffusion (AKs), temperature vertical diffusion coefficient (AKt), and vertical viscosity coefficient (AKv), –, all of which influence sub-grid-scale vertical mixing. These fields are obtained from the online parent run via the 'AKXCLIMATOLOGY' CPP flag, which ingests the 3-h climatologies of AKs, AKt, and AKv. In addition, the 'MIXCLIMATOLOGY' flag is used to import the generic length scale (GLS) and turbulent kinetic energy (TKE) coefficients from the online simulation.*

In contrast, LMD simulations omit these climatology flags, so the offline model recomputes its own AKs, AKt, and AKv internally by using the same turbulence-closure parameters defined in the online configuration. This ensures that vertical diffusivity remains active under all schemes while enabling a direct test of sensitivity to externally prescribed versus internally computed mixing fields. This treatment also mirrors the approach of Thyng et al. (2021) for the GLS and MY25 cases; for full implementation details of these flags, refer to Thyng et al. (2021).

ii) Reviewer 2 had a concern about the additional value of showing the RMSE besides the SS. The concern arises because the SS is the RMSE normalized by the online values, therefore both properties quantify differences in amplitude between the online and offline. Authors could replace Table 1 by a Taylor diagram (Taylor 2001), offering a visual and faster comparison between simulations (DT and vertical mixing scheme) to the reader. This diagram shows RMSE but also the correlation coefficient, a metric on how the online and offline values covary in time and/or space, which is currently lacking in the evaluation. In this case, statistics for both p and RMSE should be weighted by volume given that it varies among grid cells (I believe depth levels are unevenly spaced in the model) and tracer concentrations are per unit volume. Volume-weighted averages should be also used for the computation of the SS time series in Fig. 2.

I also agree with reviewer 2 that the results shown in Fig. 3, and its related discussion (L286-L295), are applicable to Fig. 6, which also seems more relevant to me. The only differences are the discrepancies in PO_4 and NO_3 offshore, which are not very relevant to the manuscript and would likely be unveiled in Fig. 6 if considering the full water column.

Finally, related to model evaluation of results shown in Fig. 4, how is equation (1) applied when two offline simulations are compared? That is, which values are used in the denominator to normalize the RMSE? Could this be related to the fact that the heatmap is not exactly symmetric?

RESPONSE: We appreciate the reviewer's insightful suggestions regarding the evaluation metrics used in our study. In response to each of the reviewer's points, we have made the following enhancements:

1. **Volume-weighted metrics:** We acknowledge the importance of using volume-weighted averages for both RMSE and SS, especially given the variability in grid cell depths and tracer concentrations. We have implemented these changes in our calculations, ensuring that all skill metrics are now computed using volume-weighted averages. This adjustment is detailed in the Methods section (Subsection 2.5, Lines 212-242), where we explain the weighting procedure applied to each metric as follows:

L213-242: *We evaluate the Offline Fennel model using two complementary, volume-weighted diagnostics: a skill score (SS) and the root-mean-square error (RMSE). Both metrics account for the true physical volume ($V_{i,j,k}$) of each grid cell, thereby avoiding biases due to varying horizontal areas or layer thicknesses.*

Each cell volume ($V_{i,j,k}$) is computed as the product of the horizontal ROMS grid spacings (Δx_i, Δy_j) and the vertical thickness ($\Delta z_{i,j,k}$), which is calculated from the difference in model layer depths (z_w) at each horizontal location (i,j).

*SSs are a widely used metric for evaluating model performances (Bogden et al., 1996; Hetland, 2006). To assess the performance of our Offline Fennel model, we applied the following equation (**Eq. 1**), adapted from Thyng et al. (2021):*

$$SS = 1 - \sqrt{\frac{\sum_{i,j,k} V_{i,j,k} \times \left(C(t) - C_{ref}(t)\right)^2}{\sum_{i,j,k} V_{i,j,k} \times C_{ref}(t)^2}} \quad \textit{(Eq. 1)}$$

where C(t) and C_{ref}(t) are the concentrations of a tracer of a tracer at time t in the compared and reference simulations, respectively – typically representing offline (uncoupled) and online (coupled) configurations. The sums are performed over all spatial and vertical dimensions, and results are volume-weighted. This yields a time series of SS values that tracks the temporal evolution of model performance. A time-mean SS can be computed

by averaging over the simulation period, providing a single, scalar measure of overall model accuracy.

*To complement the SS analysis, RMSE was equally employed as a metric to assess the accuracy of offline simulations compared to online results (**Eq. 2**). RMSE provides insight into the magnitude of errors by measuring the square root of the average squared differences between offline and online simulation results.*

$$RMSE = \sqrt{\frac{\sum_{i,j,k} V_{i,j,k} \times \left(C - C_{ref}\right)^2}{\sum_{i,j,k} V_{i,j,k}}}$$

(Eq. 2)

where C and C_{ref} are the time-averaged concentrations of a tracer on the 3D grid for the offline and online simulations, respectively. Because each grid cell's contribution is proportional to its volume, this RMSE reflects the true three-dimensional error structure.

Together, these volume-weighted SS and RMSE metrics provide a robust evaluation of model performance, capturing both relative and absolute discrepancies while avoiding biases caused by unequal grid cell sizes.

Moreover, the figures based on these updated metrics have been revised accordingly: Figs. 2, 3, and 4, as well as Supplementary Figs. S1, S2, and S3, and Table 1.

While the overall results remain consistent with the previous version, we have adjusted some numerical values in the following paragraphs of the Results section to reflect the new metrics. The changes are highlighted in **bold** for clarity:

L263-286: *Table 1 summarizes the mean SSs computed using Eq. 1 for key biogeochemical tracers. Across all mixing schemes, the simulations demonstrate high accuracy, with minimal differences between configurations. GLS slightly outperforms the others in some tracers, with scores **above 95% for NO_3, PO_4, and O_2, and a mean SS of 92.92% across all tracers.** CHL scores hover around **91%,** highlighting the scheme's ability to capture primary production dynamics effectively. However, NH_4 exhibits lower SSs, ranging from **83.21% to 83.53%.***

*The LMD scheme similarly produces robust results, particularly for **O_2, PO_4, and NO_3 with SSs of 98.68%, 96.00% and 95.44%, respectively**. Its mean SS is slightly lower (**92.79%),** reflecting solid tracer performance overall, although NH_4 again shows the weakest accuracy, with SSs between **82.53% and 82.94%**.*

*Results for the MY25 scheme align closely with those of GLS and LMD, yielding SSs of **96.12%** for NO_3 and **96.86%** for PO_4. The mean SS for MY25 is **92.86%,** underscoring its comparable performance. While NH_4 again exhibits lower SSs, the values still represent good model performance, as they remain above **80%.***

L332-335: ***In relation to typical tracer concentrations in the study area,*** *O_2 systematically shows the lowest errors across the domain, with only small and localized increases near the coast. […]*

L337-338: *Finally, NH_4 presents the highest error levels **when considering typical concentrations in the region,** which range from 0 to 5 mmol·m^{-3} at the surface and can reach up to 20 mmol·m^{-3} at depth. […]*

L344-345: *In contrast, comparisons across different mixing schemes show decreased SSs, dropping **to 92%** in some cases between GLS and LMD simulations.*

2. **Addition of a Taylor diagram:** we concur that a Taylor diagram offers a comprehensive visual representation of model performance by integrating RMSE, correlation coefficients, and standard deviation metrics. To enhance our analysis, we have added this diagram to complement Table 1, providing a clearer comparison of the models' performance metrics across spatial and vertical dimensions.

L261-271: *To assess model performance, we first present a Taylor diagram (**Fig. 2**) (Taylor, 2001) that illustrates the volume-weighted and normalized statistics averaged across all biogeochemical variables. This diagram highlights a strong agreement between the offline simulations and the online parent model. The offline configurations show higher standard deviation values compared to the coupled reference (0.26 across all vertical mixing schemes), exhibiting relative standard deviations slightly exceeding 1.*

*GLS and LMD demonstrate remarkably similar performance, characterized by standard deviations ranging from 1.107 to 1.123, low centered root mean square error (RMSE) values around 0.25, and high correlation coefficients ($r > 0.98$). The MY25 scheme shows slightly higher RMSE values (up to 0.29) and marginally lower correlation coefficients, but these differences are minimal and do not significantly detract from its overall performance. Furthermore, differences between DTs are negligible across all cases (**Fig. 2**).*

3. **Relevance of Fig. 3 and Fig. 6:** Following the update of the skill metrics equations to be volume-weighted (as detailed in point 1), we have chosen to retain both figures in our analysis. This decision aims to provide readers with a more comprehensive understanding of the various metrics and the differences observed across the different experiments and simulations.

4. **Comparison of offline simulations:** In relation to the application of Equation 1 for comparing two offline simulations, we would like to clarify that the denominator used for normalizing the RMSE is based on the SS reference series defined by the first simulation in the comparison. We have revised Equation 1 to incorporate this clarification, defining C and C_{ref} to reduce potential confusion for readers (please see the updated equation in point 1).

MINOR COMMENTS

L15. Comma after fields.

RESPONSE: Noted and corrected.

L20. Express the time reduction in percentage since it is more meaningful as a general message. Authors could also mention in the abstract the fact that NH4 is a challenging tracer to simulate accurately offline since its timescale of change is faster than other tracers. This is a nice general result applicable to other offline biogeochemical model implementations.

RESPONSE: Thank you for the suggestion. We have included the time reduction in percentage and now also mention the challenging aspect about modelling NH4 as follows (changes in **bold**):

L19-20: *By leveraging physical hydrodynamic outputs, we ran the Offline Fennel model using various time-step multiples from the coupled configuration, significantly enhancing computational efficiency* ***and reducing simulation computational time by up to 87%.***

L25-26: ***A significant challenge identified was the simulation of ammonium (NH_4), which exhibited the largest discrepancies due to its rapid turnover timescale compared to other tracers.***

L90. Which type of shift? A time shift?

RESPONSE: Yes, thank you. We have clarified that the shift refers to time, as follows (changes in **bold**):

L90-91: *In the previous model version, a* ***time*** *shift occurred when processing climatology fields, leading to a bias that propagated from the bottom toward the surface, affecting tracer concentrations.*

L121. Add “vertical” before “mixing”.

RESPONSE: Noted and added (in **bold**).

L123: *To evaluate the offline model performance, online simulations were conducted using three different* ***vertical*** *mixing schemes.*

L131-L132. Instead of listing all variables, authors could say that the Fennel model variables are included except those involved in carbon pools (which imply using ALK). This is a more direct way for the reader to understand what is exactly included and avoids redundancy since all Fennel model variables are listed in L100-102.

RESPONSE: Noted and corrected as follows (changes in **bold**).

L137-139: *For the biogeochemical implementation, we used the same configuration for both online and offline simulations to ensure comparability. The state variables incorporated* ***all Fennel model tracers except those involved in carbon pools (e.g., alkalinity).*** ~~*from the biogeochemical model include NO_3, NH_4, PO_4, CHL, phytoplankton, zooplankton, nitrogen detritus divided into large, small, and river-derived, and O_2.*~~

L147. Climatology over which time period?

RESPONSE: This file is not a traditional climatology but rather a 3-hourly forcing file that is incorporated in ROMS through the climatology mechanism (Thyng et al., 2021). We have clarified the time period in the manuscript, as follows (changes in **bold**):

L159-160: *Following the recommendation of Thyng et al. (2021), a 3-hourly hydrodynamic input frequency was selected to run the offline simulation* ***for both the physical and climatology forcing files.***

The details of each forcing file are described in the Appendices.

L147. Explain what zeta, ubar- and vbar- velocity properties are.

RESPONSE: Noted.

L154-155: *The climatology forcing incorporated variables such as* ***free surface elevation (zeta), vertically integrated u-momentum component (ubar), vertically integrated v-momentum component (vbar)****, […]*

L174-L175. Delete last sentence to avoid too much redundancy. The value of the barotropic time step is already stated in L107.

RESPONSE: Noted and removed as follows:

L183-188: *Offline simulations were run with varying multiples of the online DT (x1, x3, x5, x10, and x15) to improve computational efficiency, until the results became unstable. Given that the baroclinic DT of the online simulation was 60 seconds, these corresponded to offline baroclinic DTs of 60 s, 180 s, 300 s, 600 s and 900 s. A DT 15 times longer than the online time-step led to unstable writing of solutions. As such, while this case was initially tested, it was excluded from analysis figures and tables to avoid misleading interpretations.* ~~*Thus, the offline DT could be 1, 3, 5, or 10 times the online DT. Our offline simulation*~~

~~*experiments extended up to 600 s for the baroclinic DT and 75 s for the barotropic DT ('NDTFAST').*~~

L124. Delete "subgrid-scale turbulent mixing closure scheme" since all three schemes are.

RESPONSE: Noted and removed as follows:

L124-125: *the Large–McWilliams–Doney (LMD) mixing scheme, also known as the K-Profile Parameterization~~, which is a subgrid-scale turbulent mixing closure scheme~~ (Large et al., 1994) […]*

L309. Remove "(GLS, LMD, or MY25)" to avoid redundancy.

RESPONSE: Noted and removed as follows:

L343-345: *The figure illustrates that simulations using the same mixing scheme ~~(GLS, LMD, or MY25)~~ exhibit the highest similarity, with SSs […]*

L330. Specify the time length of the averaged output.

RESPONSE: The revised sentence now specifies the time resolution clearly: (changes in **bold**):

L371-372: *For consistency, all plots were generated using **3-hourly** 'avg' output files from ROMS**, which were then daily averaged.***

Figure 8. I think panels C are redundant differences from B are shown in A. Also, in caption remove "(x5 DT)" after "offline" to avoid repetition. Finally, consider a colorblind friendly colorbar in panels B such as the cmocean ones (Thyng 2016).

RESPONSE: We acknowledge the redundancy but decided to retain both panels B and C to help readers unfamiliar with typical chlorophyll levels in the region better interpret the differences shown in panel A. As suggested, we have removed "(x5 DT)" from the caption and replaced the colormap with the colorblind-friendly 'algae' from the cmocean package (Thyng et al., 2016).

Figure 8 caption: *Seasonal maps of chlorophyll concentrations (CHL) for Generic Length Scale (GLS) mixing simulations using x5 time-step (DT). (A) Displays the difference in concentrations between offline (x5 DT) and online simulations, as indicated by the coolwarm color scale on the right. (B) Shows chlorophyll concentrations for the online simulation, while (C) presents concentrations for the offline simulation (x5 DT). **Panels (B) and (C) share the same color scale from the cmocean package ('algae')**. Seasonal designations are as follows: DJF (December-January-February, winter), MAM (March-April-May, spring), JJA (June-July-August, summer), and SON (September-October-November, fall).*

REFERENCES

Fennel, K., Mattern, J.P., Doney, S.C. et al. (2022). Ocean biogeochemical modelling. Nat Rev
Methods Primers 2, 76. https://doi.org/10.1038/s43586-022-00154-2.
Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram, J.
Geophys. Res., 106(D7), 7183–7192, doi:10.1029/2000JD900719.
Thyng, K. M., Greene, C. A., Hetland, R. D., Zimmerle, H. M., & DiMarco, S. F. (2016). True
colors of oceanography. Oceanography, 29(3), 10.
Warner, J.C. et al. (2005). Performance of four turbulence closure models implemented using a
generic length scale method, Ocean Modelling, Volume 8, Issues 1–2, 2005, Pages 81-113.