Before replying to the reviewer, we must inform that we found a bug in a code to calculate the RMSDs relative to the KEO buoy and then corrected it (Figs. 8 and 9 in the revised manuscript). Specifically, the number in the denominator is smaller in the RMSD calculation, and the corrected results show larger RMSDs for all experiments than the previous results. However, the corrected results are qualitatively the same as the previous results. Therefore, this correction has few impacts on the conclusion in this paper. We apologize for the above.

Referee #1 (Dr. Yue Ying)

Summary: I found the revised manuscript improved in clarity and overall quality. However, the authors have not fully addressed several major issues the reviewers raised, which still limits the scientific merit of the paper. I have reiterated the major issues below, note that to address them you don't necessarily need huge amount of computational resource for additional experiments. I would strongly encourage the authors think in terms of what the readers can learn from your results and put more effort in the interpretation of results.

We thank the reviewer for reviewing again our manuscript. We have conducted ensemble forecast experiments using the forecasts from the sensitivity experiments, and confirmed that the results from the forecast accuracy are qualitatively the same as the analysis accuracy.

1. My biggest concern is that only analysis ensemble is used in diagnosing RMSD for DA performance. Is your goal only to make the best ocean reanalysis? If you are implementing an ocean prediction system I will assume the forecast accuracy does matter and shall be considered in the diagnosis. Only a few additional forecasts (10-day forecast from every 1st of each month?) will provide some evidence that the IAU+RTPP outperforms the CTRL in terms of both balance and accuracy. If you absolutely don't have additional computational resources, then consider checking the RMSD for the priors (1day forecasts before DA, you should have them already). The bottom line is that some evidence shall be shown that improved accuracy/balance at analysis time will lead to some improvement in the forecast.

We thank the reviewer for your comments on the forecast accuracy. We performed 11-day ensemble forecast experiments initialized once a month in 2016 (i.e., total 12 cases) by the forecasts from the NO INFL experiments with/without IAU,

RTPP09 experiments with/without IAU, and RTPS09 experiments with/without IAU, and then calculated the daily averaged ensemble mean forecast RMSDs relative to the independent drifter buoys. The results are shown in Figs. 4a, b and 7 in the revised manuscript, indicating that the forecast RMSDs of surface horizontal velocities are qualitatively the same as the analysis RMSDs. We have added the related descriptions to the abstract, the second paragraph in subsection 3.3, the last paragraph in subsection 4.2.1, and the first paragraph in Section 6.

2. The way imbalance is diagnosed is still problematic: currently delta_NBE is defined on the analysis increments delta_u and delta_eta (f-a), which means you assume the prior (f) fields are completely balanced and imbalance is only introduced by DA. Is this really the case? Does some imbalance still persiste after 1 day forecast, especially when ocean time scales are quite long? Could it be better to compute NBE separately for the prior and the analysis, and compare the two separately for different DA methods? If you can demonstrate that prior from the IAU+RTPP has better balance (lower NBE) and accuracy (lower RMSD) than CTRL, I will be convinced that IAU+RTPP is truly the best setup.

Although the reviewer asks to calculate ΔNBE for the forecast field, this seems not to be reasonable because the forecasts are outputs from the model and follow its dynamical theory. The surface horizontal velocities can be approximately decomposed as the geostrophic and ageostrophic velocities. Here, we assume that the ageostrophic velocities result from the wind stress curl except for vertical geostrophic shear according to the classical Ekman theory (Cronin and Tozuka 2016). In this system, the atmospheric field is not analyzed, and therefore the prior and posterior ageostrophic field would not be changed. Therefore, the analysis increments of the SSH and surface horizontal velocities are better to satisfy the geostrophic balance. The geostrophic imbalance in the analysis increments is likely to be the source of the initial shocks. Little initial shock would occur if the analysis increments satisfied the geostrophic balance. We have added the above description to subsection 2.4.1.

These two issues are essential and I hope the authors will not bypass them and try to fully address them.

Other comments:

3. A comment on novelty of the paper: while IAU and RTPP are themselves well documented already, the combination of the two in a real prediction system is novel and

should be potentially useful for the DA community. But the interpretation is the key here, you need to go beyond just showing that IAU works and RTPP works as expected, and combining them just adds the merit. Since the balance and accuracy cannot be simultaneously satisfied, is it a matter of just finding the trade off? If I am making a reanalysis and don't care about prediction, can I just forget about imbalance (see that "best method" depends on application scenario)? Also, how can the readers be guided for their own implementation which method they shall use and how can they tune the parameters? Do you believe IAU+RTPP shall always perform well for all scenarios?

We thank the reviewer for your discussion on the novelty of this paper. The IAU is adopted in ocean data assimilation systems with *3D-VAR* and *4D-VAR* (See table 2 of Martin et al. 2015), the RTPP and RTPS are well used in the EnKF-based *atmospheric* data assimilation systems. However, as described in the fourth paragraph in Section 1. "the IAU and RTPP/RTPS have not been widely used in an EnKF-based ocean data assimilation systems (Table 1).", and unfortunately there are still few results how the IAU and RTPP/RTPS affect the dynamical balance and accuracy in EnKF-based ocean data assimilation systems. Although the most novelty part is the combination of the IAU and RTPP as indicated by the reviewer, this paper would have scientific significance for the ocean data assimilation community by comprehensively investigating the impacts of the IAU and covariance inflation methods on the dynamical balance and accuracy. The results are helpful for readers to newly construct or develop EnKF-based data assimilation systems in various fields including the ocean.

Based on the RMSDs of surface horizontal velocity in Fig. 4a, b, the combination of the RTPP and IAU is not a simple superposition. It is not easy to separate the effects of the RTPP and IAU in the RTPP+IAU experiments, probably because the RTPP and IAU appear to interact with each other in terms of the balance and accuracy. However, the results from the IAU, RTPP, and RTPP+IAU experiments imply that the RTPP maintains the ensemble spread inflated by the perturbed boundary conditions and results in the improvement of the accuracy but the degradation of the balance, and at the same time the IAU improves the degraded dynamical balance by the RTPP reducing the impacts from the initial shocks. This leads to further improvement of the accuracy. We have added the description at the end of the 1st paragraph in Section 6.

As described in the third paragraph in Section 1, the impacts from the initial shocks are likely to be accumulated and degrade the accuracy if frequent data assimilation is conducted, and a suitable setting to provide well-balanced analyses is necessary to construct accurate analysis products. Therefore, it is better to monitor the dynamical

balance if the assimilation interval is quite short.

Since the appropriate setting depends on systems as described in the second paragraph of Section 6, to find the best setting, it is necessary to comprehensively investigate the impacts of IAU and covariance inflation methods. Thus, we do not think that the RTPP+IAU is the best setting for all systems, but we expect that the RTPP+IAU improves the balance and accuracy in EnKF-based frequent data assimilation systems in which initial shocks have substantial impacts.

4. Temporal evolution and spatial distribution of the imbalance and inflation seem to me as the key to understanding the behavior (interplay between IAU and RTPP), you can try to pick a time series and visualize how NBE, RMSD, and inflation (a surrogate for spread reduction or where increment occurs in DA) fields evolve, instead of just showing their time- and domain-averages. The authors seem to consider this detailed analysis to be out of scope of the current study, but I think it is relevant. If you can provide some evidence that analysis accuracy and balance leads to better priors 1 day later it will convice us to use the IAU+RTPP setup.

Figure R1 shows spatially averaged ΔNBE, analysis RMSDs of surface zonal velocity relative to the drifter buoys, and absolute SSH increment averages for each ensemble. Since ΔNBE results from the geostrophic imbalance (the difference between the SSH gradient and surface horizontal velocity in the analysis increment fields), only the SSH and horizontal velocity increments cannot explain how ΔNBE undergoes spatiotemporal variations. As shown in Fig. R1a, c, the timeseries of the SSH increment does not correspond to that of the ΔNBE. Since the accumulated initial shocks result in the degradation of the accuracy in frequent data assimilation, it is reasonable that the timeseries of the RMSDs is not consistent with that of the ΔNBE (Fig. R1a, b). Better balance and higher accuracy essential for frequent assimilation are well maintained in the RTPP09+IAU experiment than in the NO INFL experiment. We have already described the implication for interplay between the RTPP and IAU in the second paragraph in the reply to the third major comment. Furthermore, as replied to the first comment, we have demonstrated that the combination of the IAU and RTPP09 is the best for the forecast accuracy as in the analysis accuracy.

5. I fully understand the limit in computational resources, and in this case will not insist on tuning for the optimal parameters. The time evolution of averaged delta_NBE indicates a clear seasonal cycle, so I would expect the best parameter will also have a seasonal

cycle anyway. Maybe you can make a similar time series for the estimated inflation factor (you have shown the mean as 1.08, 1.11, and so on) to show that in practice an adaptive inflation/relaxation is better.

The best relaxation parameters would depend on time and space, and the adaptive RTPP and RTPS (Yue et al. 2015; Kotsuki et al. 2017) might be useful. This is a topic in future studies. Since the estimated MULT is used to investigate how much the inflation in the RTPP09+IAU experiment corresponds to the MULT parameter as described in the first sentence in Section 5, the estimated MULT does not enable us to estimate the suitable tuning parameter for the MULT and RTPP09+IAU from the estimated MULT.

Reference:

Cronin MF, Tozuka T (2016) Steady state ocean response to wind forcing in extratropical frontal regions. Sci Rep 6:28842. https://doi.org/10.1038/srep28842

Referee #3

I don't think the authors have addressed my previous comments. To be scientifically sound and a valuable contribution for the ocean DA community, the manuscript definitely need further and detailed explanations for the results.

We thank the reviewer for reviewing again our manuscript. We have confirmed the SSH increments and forecast accuracy as replied to the second and fourth comments, respectively, and also discussed the difficulty to decompose the interplay between the IAU and RTPP in the third comment.

1. To my previous comment 1, "Compared to NOINFL, IAU in NOINFL+IAU degrades the accuracy. Why IAU degrade the accuracy for ocean assimilation that has longer time scale than atmosphere?" The authors explain that "The main difference between without and with the application of the IAU is directly updated the SSH or not. Temperature, salinity, horizontal velocities, and SSH analyses are used for the initial conditions for model integration within the assimilation window if the IAU method is not applied, whereas the analysis increments of temperature, salinity, horizontal velocities except for the SSH are distributed if the IAU method is applied." I don't understand why different update variables are used for experiments with and without IAU. Please give the detail configurations of the experiments in the text, and also explain the reasons for the different choices of updated variables for different assimilation experiments.

We thank the reviewer for your comments on updated variables in the experiments with and without the IAU. In the experiments without the IAU (i.e., standard method), all analysis variables (SSH, temperature, salinity, and horizontal velocity) are used for initial conditions to keep the consistency. As seen in table 2 of Martin et al. 2015, there are three types of the IAU in ocean data assimilation systems using the following analysis variables:

(i) temperature and salinity
(ii) temperature, salinity, and horizontal velocity
(iii) temperature, salinity, horizontal velocity, and SSH

All methods have advantages and disadvantages, and there are still discussions about which is better. We have chosen the second method in this study. It is because the SSH depends on the density, and SSH might be overcorrected if the analysis increments of the SSH, temperature, and salinity are applied to the IAU at the same time. We have added the above description at the end of subsection 2.1.

2. To my previous comment 2, “The authors state that IAU reduces the spread and accuracy of DA. But MULT, RTPP and RTPS have totally different impacts on the spread and accuracy when IAU is applied. Why MULT that also inflate the ensemble spread has the opposite impacts on spread and accuracy than RTPP and RTPS? Since the results with different inflation methods are inconsistent, it would be helpful to understand the roles of different inflation methods, especially the interactions with IAU.” The authors referenced an under-review manuscript that is not available for the reviewers. And the explanation is “The RTPP and RTPS relax the analysis ensemble perturbations toward the forecast ensemble perturbations. This implies that the analysis increments in the RTPP and RTPS would be smaller than the MULT, and the above degradation process might be suppressed.”

It would be helpful to see samples of analysis increments and subsequent forecasts, using MULT, RTPP and RTPS. However, the differences of analysis increments and subsequent forecasts with different inflation methods still cannot explain the interactions between IAU and inflation. Please provide understandings to the interactions of different inflation methods with IAU, which could help future design of data assimilation frameworks.

We apologized for not citing Ohishi et al. (in review) in the reply for Referee #3 and the revised manuscript, although it is cited in the previous reply comments for Referee #2. Ohishi et al. (in review) have been available through the GMD’s website (https://gmd.copernicus.org/preprints/gmd-2022-91/gmd-2022-91.pdf). We have added the citation to reference in this reply comments for Referee #3 and the revised manuscript.

As shown in Figs. R1c and R2, the SSH analysis increments in the RTPP09 and RTPP09+IAU experiments are substantially smaller compared with the other experiments over most of the period. Here, we do not show the increments in the MULT and MULT+IAU experiments because they are exponentially increased and exceedingly large.

Although we have conducted sensitivity experiments on the IAU, covariance inflation methods, and their combination, their combination is not a simple superposition as shown by the analysis RMSDs of the surface horizontal velocity (Fig. 4a, b). Therefore, it is not easy to separate the effects of the IAU and RTPP in the RTPP+IAU experiments, probably because the IAU and RTPP interact with each other for balance and accuracy. This is true for understanding air-sea interaction. However, the IAU, RTPP, and RTPP+IAU experiments imply that the large relaxation parameter maintains the ensemble spread induced by perturbed boundary conditions and leads to the improvement of accuracy but the degradation of dynamical balance, and at the same time the IAU

improves the degradation of the dynamical balance by the RTPP. As a result, this would lead to further improvement of the accuracy by reducing the initial shocks in frequent data assimilation. We have added the description at the end of the first paragraph in Section 6.

3. To my previous comment 3, "Previous studies of IAU (e.g., Lei and Whitaker 2016, He et al. 2020) showed that IAU has more advantages for variables that are more influenced by imbalances that variables that are less influenced by imbalances. However, results here are inconsistent with the previous findings. IAU improves the accuracy of wind field more than the accuracy of height field (Figures 3 and 4). Please provide explanations or insights for these counter-intuitive results." The authors replied "The degradation of the accuracy by the IAU is consistent with He et al. (2020) who demonstrated that the accuracy of most variables is worser in the 3D-IAU experiment than experiment without the IAU when the assimilation windows are short of 1 and 3 hours [See table 3 of He et al. (2020)]; Lei and Whitaker (2016) who indicated that the accuracy of temperature and wind speed is worser in the 3D-IAU experiment than the experiment without the IAU using NCEP GFS experiments with assimilation of real observations [See fig. 8 of Lei and Whitaker (2016)]; and Yan et al. (2014) who showed that the IAU degrades the accuracy in twin experiments using an EnKF-based ocean data assimilation system [See table 3 of Yan et al. (2014)]."

First, Table 3 of He et al. (2020) showed that for the surface height that is more sensitive to imbalances than the other variables, 3DIAU is better than NoIAU for DA frequencies of 12h, 6h and 3h, while 3DIAU is worse than NoIAU for DA frequency of 1h. I don't think the 1h DA frequency can be extrapolated here, since here less frequent observations are assimilated, and the oceanic model has much longer time scales than the QG model. Second, He et al. (2020) showed that IAU can impact the surface height more than the wind, since the latter is less sensitive to imbalances. But in this study, IAU improves the accuracy of wind field more than the accuracy of height field (Figures 3 and 4). Please provide dynamical explanations for this result.

To focus on the points to be discussed, we only compare the results from the assimilation experiments at the finest interval of 1 hour in table 3 of He et al. (2020) with this study, because 1-day assimilation is a quite frequent interval for the ocean data assimilation systems. As seen in table 3 of He et al. (2020), the 3DIAU results in *larger* RMSEs and *degrades* the accuracy of all atmospheric variables (upper- and lower-layer

wind, interface height, and surface height). However, the reviewer incorrectly described "IAU *improves* the accuracy of wind field more than the accuracy of height field" in the previous and present comments, although we have indicated this in the previous comment. Therefore, the IAU *degrades* the accuracy of the SSH and surface horizontal velocity, and this is qualitatively the same as He et al. (2020). We note that He et al. (2020) conducted assimilation experiments at a 1-hour interval using hourly observations, and therefore their results are not extrapolated.

In the second point, we again note that the IAU *degrades* both SSH and surface velocity fields. As described in subsection 2.4.3, the SSH/SSHA from the AVISO and the surface horizontal velocity from the drifter buoys are completely different. Since the AVISO is constructed by summing optimally interpolated satellite SSHA and mean dynamical ocean topography estimated from the atmospheric datasets and drifter buoys, the AVISO is not independent dataset. In contrast, the surface drifter buoys are independent observations. Therefore, we cannot quantitatively nor directly compare the accuracy between the SSH and surface horizontal velocity. If forecast/analysis errors are accurately estimated by conducting twin experiments, the direct comparison is possible.

4. To the last question of my previous comment 4, "The RMSD is computed for the prior or posterior? How the RMSD is computed for experiments with IAU?" and my previous comment 5, "Since assimilation is conducted at a daily frequency, both the daily prior and free forecast at longer forecast lead times worth to check." The authors replied "To perform a free forecast after every assimilation cycle, all experiments must to be integrated again, and the huge amounts of the computational resources are required. Consequently, this is an issue in future studies." I totally understand the computational cost. But since cycling assimilation experiments are already done, it should be straightforward to calculate the verifications with priors, since no additional computations needed. Moreover, just several samples of long free forecasts from different assimilation experiments could be useful to draw some conclusions.

Because of the limitation of the storage, we have conserved only instantaneous forecasts and analyses (restart files) at the 1st and 16th days for each month, and daily averaged ensemble mean and spread throughout the whole period. Therefore, we have to conduct all experiments from the beginning if daily forecast RMSDs are calculated throughout the analysis period. Instead, we performed 11-day ensemble forecast experiments for each month in 2016 following the first major comment from Referee #1 (Fig. 7 in the revised manuscript). The results show that the forecast RMSDs are

qualitatively the same as the analysis RMSDs, and that the forecast accuracy is the best in the RTPP09+IAU experiment. Therefore, the combination of the IAU and RTPP is the most suitable for not only constructing the analysis products but also conducting ensemble forecast.

Reference:

Ohishi, S., Miyoshi, T., and Kachi, M.: An EnKF-based ocean data assimilation system improved by adaptive observation error inflation (AOEI), Geosci. Model Dev. Discuss. [preprint], https://doi.org/10.5194/gmd-2022-91, in review, 2022.

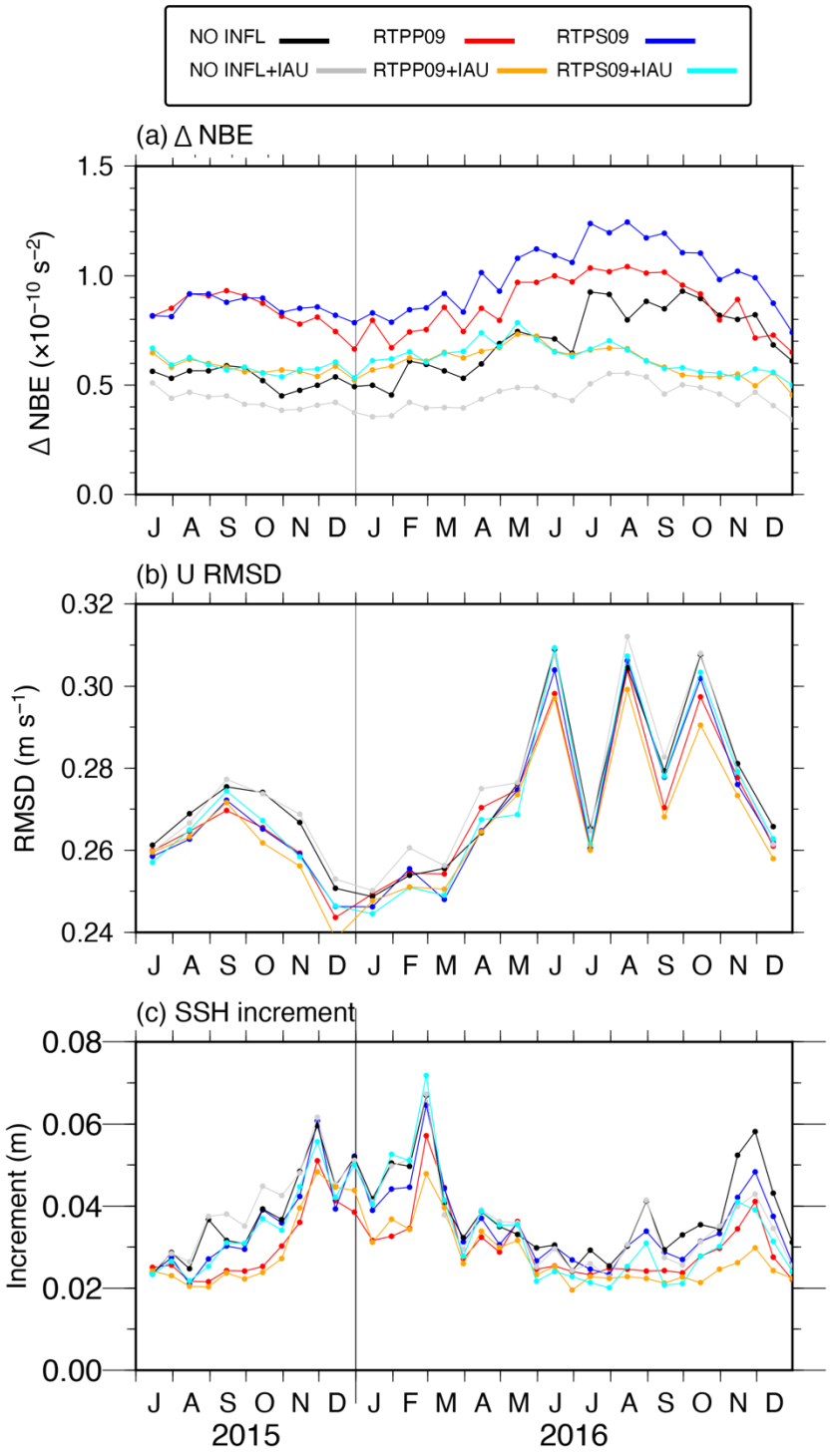

Figure R1: Spatial averaged (a) ΔNBE, (b) analysis RMSDs of the surface zonal velocity relative to the drifter buoys, and (c) absolute SSH increments over the whole domain in the NO INFL (black), RTPP09 (red), RTPS09 (blue), NO INFL+IAU (gray), RTPP09+IAU (orange), and RTPS09+IAU (orange) experiments.

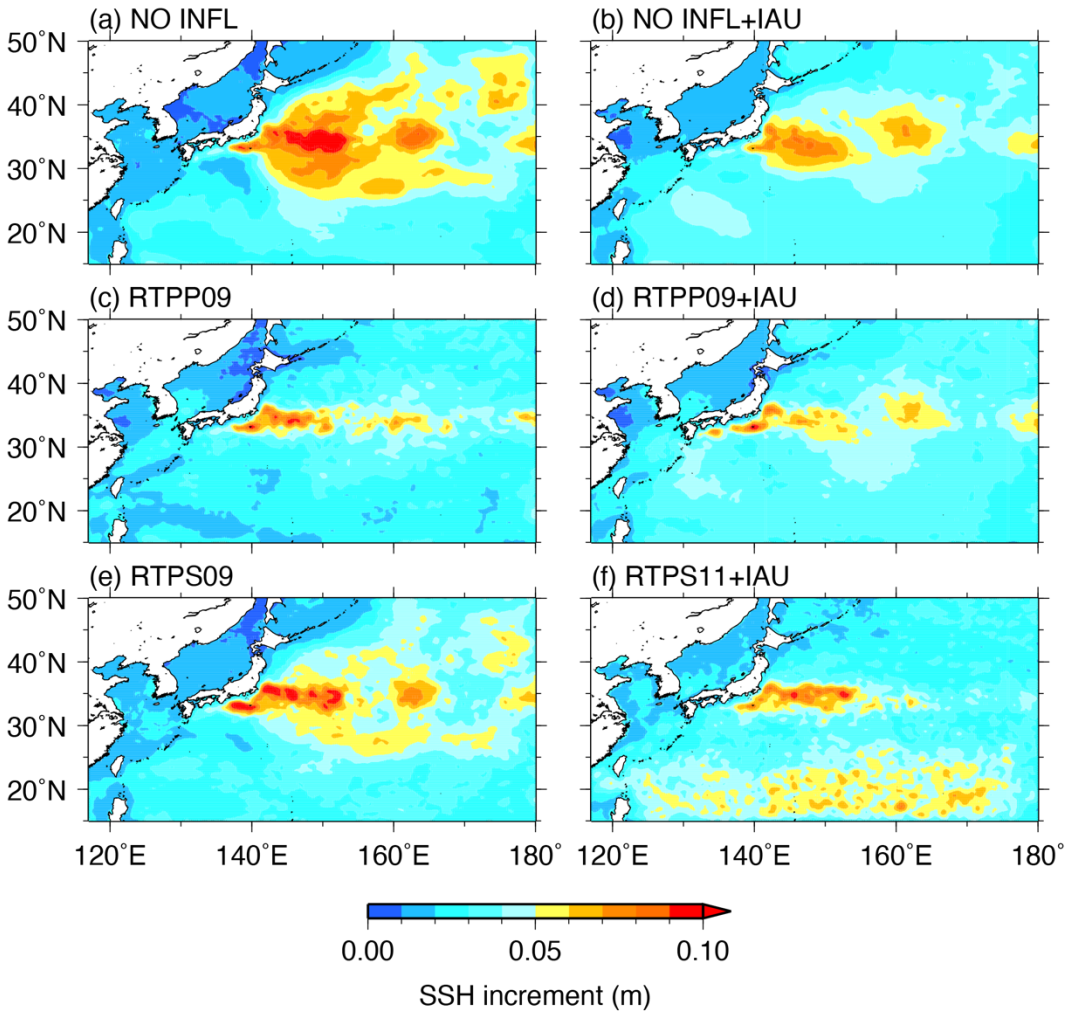

Figure R2: Absolute SSH increments averaged over 2016 in (a) NO INFL, (b) NO INFL+IAU, (c) RTPP09, (d) RTPP09+IAU, (e) RTPS09, and (f) RTPS11+IAU experiments.