Comment from Referee#2: General comments

This paper presents an assessment of a hybrid variational-ensemble scheme and looks at the impact of systematic error correction and the inclusion of systematic bias correction. I generally found this a good and interesting paper. I do have a few mostly minor comments which I'll list below along with the page and line number.

Author's response:

We thank the Reviewers for his/her positive comments. Below we provide answers to each comment.

Comment from Referee#2:

- page 1 title. Should be either "error correction for a limited area ocean model" or "error correction for limited area ocean models"
- page 1 line 23-25. Slightly confusing sentence. What's "hybrid daily estimates" perhaps this would read better as "and a hybrid of the background error covariance ... ensemble derived errors of the day" or similar
- page 1 line 30. present day observations.

Author's response: *We Thank the Reviewer for his/her comments.*

Author's changes in manuscript: *All the suggestions will be implemented in the revised version of the manuscript.*

Comment from Referee#2: page 2 line 10. Divergence happens in chaotic systems even if the model is perfect it's not just a result of model bias.

Author's response: *We Thank the Reviewer for his/her comments.*

Author's changes in manuscript: *We agree with the Reviewers and the sentence will be modified as follows:*

"These approximations affect the model solutions in terms of quality and accuracy and, more importantly, differences between the numerical solution and the true state amplify along time due to the chaotic component of the ocean dynamic."

Comment from Referee#2: page 3 line 26. Write "A promising application..." Also it would be useful to be more specific about the findings of Penny et al.

Author's response: *We Thank the Reviewer for his/her comments.*

Author's changes in manuscript: *In the revised version of the manuscript we will include the following sentence:*

"They compared hybrid, classical 3DVAR and EnKF schemes in an observing system simulation experiment and using also real data, showing that the hybrid scheme reduces errors for all prognostic model variables eliminating growth in biases present in the EnKF and 3DVAR."

Comment from Referee#2: page 3 line 33. I suppose you mean having a long length scale for the bias is more straightforward to implement than implementing multiple length scales for the control variables.

Author's response: *We Thank the Reviewer for his/her comments.*

Author's changes in manuscript: *The sentence will be rephrased as follows:*

"A possible simplification is to assume that systematic errors are characterized by long length scales, as often occurs to some extent (Dee 2005)."

Comment from Referee#2: page 4 line 8. independent on the -> independent of the

Author's response: *Thank, the revised version of the manuscript will be modified accordingly.*

Comment from Referee#2: page 4 equations onwards. Inconsistent bolding of vectors. Check the journal style but I think you should use bold roman for vectors and bold roman capitals for Matrices.

Author's response: *We thank the Reviewers for this comments. According Journal style Matrices are printed in boldface, and vectors in boldface italics.*

Author's changes in manuscript: *we will modify the manuscript accordingly.*

Comment from Referee#2: page 5 line 16. Say a bit more about how this adjustment is made.

Author's response: *We thank the reviewers, the following explanation will be included in the revised version of the manuscript just after eq.6:*

Author's changes in manuscript:

"The relative weighting (α) still requires empirical tuning but in general can be adjusted to the size of the ensemble. Large ensemble size can provide robust estimate of $\mathbf{B}_e$ *and thus 6 can be theoretically implemented with small α values (Menetrier and Auligne', 2015)."*

Comment from Referee#2: page 6 line 4. You don't use the ensemble to estimate the horizontal length scales here. You should say so to avoid confusion

Author's response: *In the experiment with hybrid covariance matrices we use the ensemble to estimate also the horizontal length scales varying daily. This is stated in Section 3 "Experimental set-up" page 9 line 27 and Table 2 of the original version of the manuscript.*

Author's changes in manuscript: *We suggest that the text remains unchanged in the revised version of the manuscript.*

Comment from Referee#2: page 6 line 21. I'm not sure I agree that inaccurate initial conditions produce a systematic error or bias. Initial condition error is likely to average to zero over time and is not therefore systematic.

Author's response: *We thank the Reviewers for this comment. There may be confusion: for initial condition we meant the climatological state we start the experiments from (i.e. the system initialization).We think that the temporal scales considered play a major role. Initial condition error will average to zero if the system is integrated for enough time.*

Author's changes in manuscript: *We will modify the sentence in order to avoid confusion.*

Comment from Referee#2: page 6 line 25. "This idea ... " I don't understand this sentence. Can you rewrite it.

Author's response: *Thank, we will rephrase the sentence.*

Author's changes in manuscript:

"This idea is consistent with the high-resolution model presented in Section 3 and with the experimental setup where the large scale uncertainties (initialization, boundary conditions and surface forcing) are not accounted in the generation of the ensemble members."

Comment from Referee#2:

- page 7 line 4. Should be "It is worth mentioning that the..."
- page 7 line 7. Should be "simulation allows us to retrieve..."

Author's response: *Thank, the revised version of the manuscript will be modified accordingly.*

Comment from Referee#2: page 7 line 21. This is true if there is no bias in the observations.

Author's response: *We agree with the reviewer, however this is explicitly mentioned in the original version of the manuscript at line 19: "assuming that also the observational error is unbiased".*

Author's changes in manuscript: *We suggest not to modify the manuscript.*

Comment from Referee#2:

- page 7 line 30. Add brackets something like min(J(dx))
- page 8 line 12 "Sardinia has been conducted" -> "Sardinia was conducted"
- page 8 line 17 "accounting for" -> by. Remove "data"
- page 8 line 19. "remote sensing" -> "remotely sensed"
- page 8 line 28. "Fig.01" -> Fig. 1 and similar elsewhere.
- page 8 line 32. "mean" -> "means"

Author's response: *Thank, the revised version of the manuscript will be modified accordingly.*

Comment from Referee#2: page 9. line 1-10. I found the perturbation of the observations confusing and not well explained. Are you vertically subsampling the profiles? Please clarify the text.

Author's response: *Thank, the revised version of the manuscript will be modified accordingly.*

Author's changes in manuscript: *We will rephrase the explanation as follows:*

"The ensemble members have been generated simultaneously assimilating perturbed observations, varying the corresponding observational error, and assuming different horizontal correlation radii in V_H. For the observation perturbation, either weak or strong criteria for retaining observations are used among the ensemble members, where strong quality check procedure requires both temperature and salinity observations are flagged as good, reducing the total number of assimilated observations. Filters have been applied horizontally and vertically to reduce the higher spatial sampling of observations with respect to the model grid. Within the ensemble members, different vertical cut-off scales have been used in a low pass filter, resulting in differently smoothed profiles. Horizontally, data binning has been applied to the observations falling in 1 or 2 model grid cells while keeping the original vertical resolution. When the filtering or binning procedures are applied, the corresponding full resolution profile standard deviation has been used to as an estimate of the observational error. Similar procedures have been applied to CTD and Gliders data."

Comment from Referee#2: page 10. Line 20. I wonder if you need to perturb things other than the observations and following on from my previous comment are you perturbing the observations enough. See also p 14. Surely it's better to perturb everything even is the effect is limited by the short integration time.

Author's response: *Generating the ensemble members only by perturbing the observations clearly poses the constraint of observations availability. The absence of observations reduces significantly the ensemble spread. As discussed in the manuscript (page 10 lines 22, 23) the methodology can be improved including perturbations in the initialization and in the lateral or surface boundary conditions. However the perturbation of initial, lateral and surface boundary condition could create overlaps between static-climatological background-error covariances, daily-varying ensemble-derived covariances, and large-scale systematic error corrections. As first step we preferred to keep a conservative approach ensuring the separations of the two scales.*

Author's changes in manuscript: *We think that changes in the manuscript due to other comments will better clarify this issue.*

Comment from Referee#2:

- page 11. Line 31. Remove "biases in"
- page 12. Line 16. "while goes" -> "while it goes"
- page 12. Line 18. "that also" -> "also that"
- page 12. Line 25. "Mean" -> "mean"

Author's response: *Thank, the revised version of the manuscript will be modified accordingly.*

Comment from Referee#2: page 13. Are you statistics computed after or before assimilation is it observations compared to analysis or observations compared to model background. Or are the CTD data used for comparison independent? It's not clear.

Author's response: *All the statistics are computed using the model background, i.e. before data are assimilated. We thank the reviewer for pointing this out.*

Author's changes in manuscript: *We will specify the dataset used in the revised version of the manuscript.*

Comment from Referee#2: page 13. It might be worth not abbreviating in all cases as it makes it more difficult to read. For example MB perhaps just write mean bias. Similarly with SS = skill score.

Author's response: *The revised version of the manuscript will be modified accordingly.*

Comment from Referee#2:

- page 13. Line 13-14 "capable of significantly reducing this bias (error)". Perhaps remove the error?
- page 13. Line 32. Remove "can"
- page 15. Line 18. Typo "indicating"
- page 16. Line 2 "of" -> "by"

Author's response: *Thank, the revised version of the manuscript will be modified accordingly.*

Comment from Referee#2: page 16. Line 10. Do you plan to use this Baysesian method?

Author's response: *We think that several aspects of the hybrid approach are still empirical and surely a crucial aspect is to find a robust theory for an objective estimation of the relative weights. Both the methods proposed by Dobricic et al. (2015) or Menetrier and Auligne' (2015) are valid and future investigations will try to address this issue.*

Comment from Referee#2:

- page 16. References - inconsistent "Mon. Wea. Rev."
- page 17. Reference Mirouze has a typo "LOCKLEY" in the author list.

Author's response: *Thank, the revised version of the manuscript will be modified accordingly.*

Comment from Referee#2: Table 3. Why is the 0-50 mean bias worse in the bias corrected runs?

Author's response: *We think that at these depths the assumptions done to compute the bias correction are probably not adequate as explained in the manuscript (page 13 line 15).*

Author's changes in manuscript: *In the revised version of the manuscript we will expand the statements as follows:*

"However, the systematic error correction increases the temperature mean bias between 20 and 70m depth, meaning that scales (both spatial and temporal), procedure or observation sampling used are probably not adequate at these depths. On the other hand, at similar depths,

the systematic error correction reduces the salinity mean bias (Fig.9 B). We argue that temperature and salinity systematic errors in these layers have different length scales."

Comment from Referee#2: Table 3. Not enough significant figures to see anything useful. I think it would more useful to use MB and SDE rather than the squares. Either that or add a significant figure.

Author's response: *We thanks the Reviewers for his/her comments. We substituted the values in Table 3 with MB and SDE not squared and not normalized. We think that together with the new figs.8 and 9 this better illustrate our results.*

Author's changes in manuscript: *New Table 3, new Figures 8 and 9 will be used in the revised version of the manuscript.*

Comment from Referee#2: Figure 1. Could say the date range over which the observations are plotted.

Author's response: *Thanks the info will be Included in the figure caption.*

Comment from Referee#2: Figure 2. Would be good to add a legend to this plot and some axis labels.

Author's response: *We thank the Reviewer.*

Author's changes in manuscript: *New Figure 2 with legend will be used.*

Comment from Referee#2: Figure 3. Don't understand the explanation of the middle panel. Label x axis

Author's response: *Xlabel axis will be included in the new version of the manuscript and the following caption with a better explanation of the middle panel:*

Author's changes in manuscript: *New figure 3 Caption as follows:*

"Figure 3: Example of perturbed CTD vertical profile with different quality check procedure and filtering applied. The solid black line indicates the full resolution CTD profile while horizontal lines are the associated observational error. The other colours indicate the perturbed profile. In the middle panel the 3 tested couples of horizontal correlation length scales ($L^{\varepsilon}_{x,y}$) are shown. The circles indicate the distance where the horizontal correlation of a single observation is zero. The green circle length scales are $L^{\varepsilon}_{x} = 12$ and $L^{\varepsilon}_{y} = 21$ km, this

set has been used also in the reference experiment. The red circle radii are $L_x^{\varepsilon} = 6$ *and* $L_y^{\varepsilon} = 12$ *km. The blues circle radii are* $L_x^{\varepsilon} = 3$ *and* $L_y^{\varepsilon} = 6$ *km."*

Comment from Referee#2: Figure 5. Spread == standard deviation of ensemble?
Author's response: *Thanks, the figure caption will be modified accordingly.*

Comment from Referee#2: Figure 6. Are both matrices are for the same location?
Author's response: *The climatological* **B** *is homogeneous and thus the same vertical correlations are applied in all the locations.*
Author's changes in manuscript: *We think that changes in the manuscript due to other comments will better clarify this issue.*

Comment from Referee#2: Figure 7. Give the depths of the horizontal slices. Typos in the caption.
Author's response: *Thanks we corrected the figure and figure caption accordingly.*
Author's changes in manuscript: *New Figure 7 and new caption will be included in the revised version of the manuscript.*

Comment from Referee#2: Figure 8 (also Figure 9). Consider plotting MB and SDE rather than the squares it will make it easier to distinguish the lines particularly where the errors are lower. A legend would be useful on this plot too.
Author's response: *We thank the Reviewer for his/her suggestion. We have re-drawn the Figs.8 and 9 using MB and SDE instead of their squares and we think Figures are improved, we also included a legend.*
Author's changes in manuscript: *New Figures 8 and 9 with new caption will substitute the original ones.*

Comment from Referee#2: Figure 8 (also Figure 9). I notice that the non-hybrid results are quite good sometimes better than the hybrid results why might this be? It may be a case of needing to do more tuning perhaps and that not making it worse is quite a good result perhaps. It might be worth saying a bit more about this in the text.
Author's response: *The Reviewer is right, and we think that this is due to some assumptions done in our hybrid formulation. In particular the relative* ***B*** *weights and the ensemble size can*

be significantly improved, introducing temporal and spatial dependencies in the weight or increasing the ensemble size. However the quality of the results obtained with the hybrid scheme is, in average, better than the corresponding static formulation.

Author's changes in manuscript: *We will include the following statement in the conclusion section:*

"The vertical dependency of the hybrid systems performances suggests that the empirical methods used in the estimates of the ensemble size and the relative weights of static and hybrid **B** *require a more objective and formal approach. The two quantities are clearly correlated. In this study we used a small ensemble size (14 members) and constant (both spatially and temporally) weights obtaining, however, encouraging results."*