Thank you to our editor, Olivier Talagrand, for handling the review of our manuscript. We are also grateful to all three reviewers who took the time to carefully read the manuscript again and provide further thoughtful critiques. Our responses (in **bold**) to each review are below.

1. Ll. 28- 29, Non-negative definiteness of estimates of B is generally desirable in all DA schemes and is essential in variational schemes. Well, non-negative definiteness of estimates of B is essential in all DA schemes, at least all schemes which, like the Kalman Filter, are intended at minimizing the statistical variance of the estimation error. Even if an algorithm produces results with a B that has negative eigenvalues, these results will be meaningless (as has actually been observed by many experimenters).

   We have replaced this sentence with a new one, following your suggestion: "Positive semidefiniteness of estimates of $\mathbf{P}^{\mathrm{b}}$ is essential for the convergence of variational schemes and interpretability of schemes like the Kalman Filter which are intended at minimizing the statistical variance of the estimation error."

2. Ll. 219-220, Since the minimizer of the 3D-EnVar objective function is the same as the EnKF analysis mean in the case of linear observation (Lorenc, 1986). It is somewhat surprising to refer here to a 1986 paper for a statement concerning 3D-EnVar and EnKF, neither of which did exist at that time. But what you say has nothing to do with ensemble estimation. You refer to the equivalence, in the linear case, between the variational (3D-Var) and the least-variance (Optimal Interpolation, Kalman Filter) algorithms. That equivalence has been known for I don't know how long (but was already known by Kalman in 1960). And I do not think it is even mentioned in (Lorenc, 1986), which deals with the Bayesian character of assimilation in the Gaussian case (and does not mention variational assimilation). Just say that 3D-Var and the analysis step of KF are equivalent in the case of a linear observation operator (for a reference, you might possibly mention Daley, R., 1991, Atmospheric Data Analysis, Cambridge University Press, Cambridge, UK, 457 pp.). (I add that variational and least-variance equivalence does not hold in your case for the time dimension, since the dynamical model you use is non-linear).

   We changed this section following your suggestion and the one from S. Penny. We also moved this discussion to the conclusions.

3. Ll. 156-157, ... the exponent [...] is allowed to vary by process ... What does that mean ?

   Is it clearer if we say "... the exponent [...] can be different for each process while the localization radius *R* is constant across all processes." ?

We thank S. G. Penny for his time in reading our manuscript and providing a thoughtful critique. His comments and our responses (**in bold**) are below.

General Comments:

1. The development of localization schemes for coupled dynamics is an important activity that needs increased attention as operational forecast centers transition to greater reliance on coupled Earth system forecast models. The authors provide a promising advancement to address the localization of cross-domain error correlations. I believe the work should be published after the authors explore a larger parameter space for their experimental results, as described below. In exploring a larger parameter space, it may be sufficient to focus on one or two leading methods (e.g. GC and BW).

   We have explored a larger parameter space, as described in the responses to general comments 2 and 3.

2. One concern is the choice of model, and how well the results can transfer to more realistic scenarios, given the near linear relationship between the slow and fast components in this system (e.g. see S. Rasp note referenced below). Do the authors have confidence that the results can translate in some way to more sophisticated systems? I would be interested to know how the results change as the coupling strength between the slow and fast components is weakened or strengthened from the baseline state used by the authors.

   The note of S. Rasp is in reference to subgrid-scale parameterization with this model, so it is not directly relevant. The coupling between scales is almost always nonlinear, unlike the coupling between fast and slow in the Lorenz-96 model. However, important couplings between atmosphere and ocean can be linear, e.g. their exchange of sensible heat, which is approximately linearly propotional to the temperature difference. We do agree that the two-scale Lorenz-96 model is certainly highly idealized though, and we appreciate the need for further testing in more realistic coupled ocean-atmosphere models and are working on this in a current project. We are testing the performance of multivariate Gaspari-Cohn in q-gcm: http://www.q-gcm.org/

   We have repeated the experiment observing only the "short" Y **process using coupling strengths** $h = 1/2, 1, 4$ **(Figs. 1-3) The coupling strength is** $h = 2$ **in the figure in the paper. The biggest change we saw is that the magnitude of the analysis errors in the unobserved** X **process increased with decreasing** h**. This is not surprising since it just confirms the intuition that the cross-assimilation decreases as the strength of the dynamical coupling decreases. The relative performance of the different localization functions did not change. Multivariate Gaspari-Cohn still led to better performance than any of the other functions and multivariate Wendland led to the worst performance.**

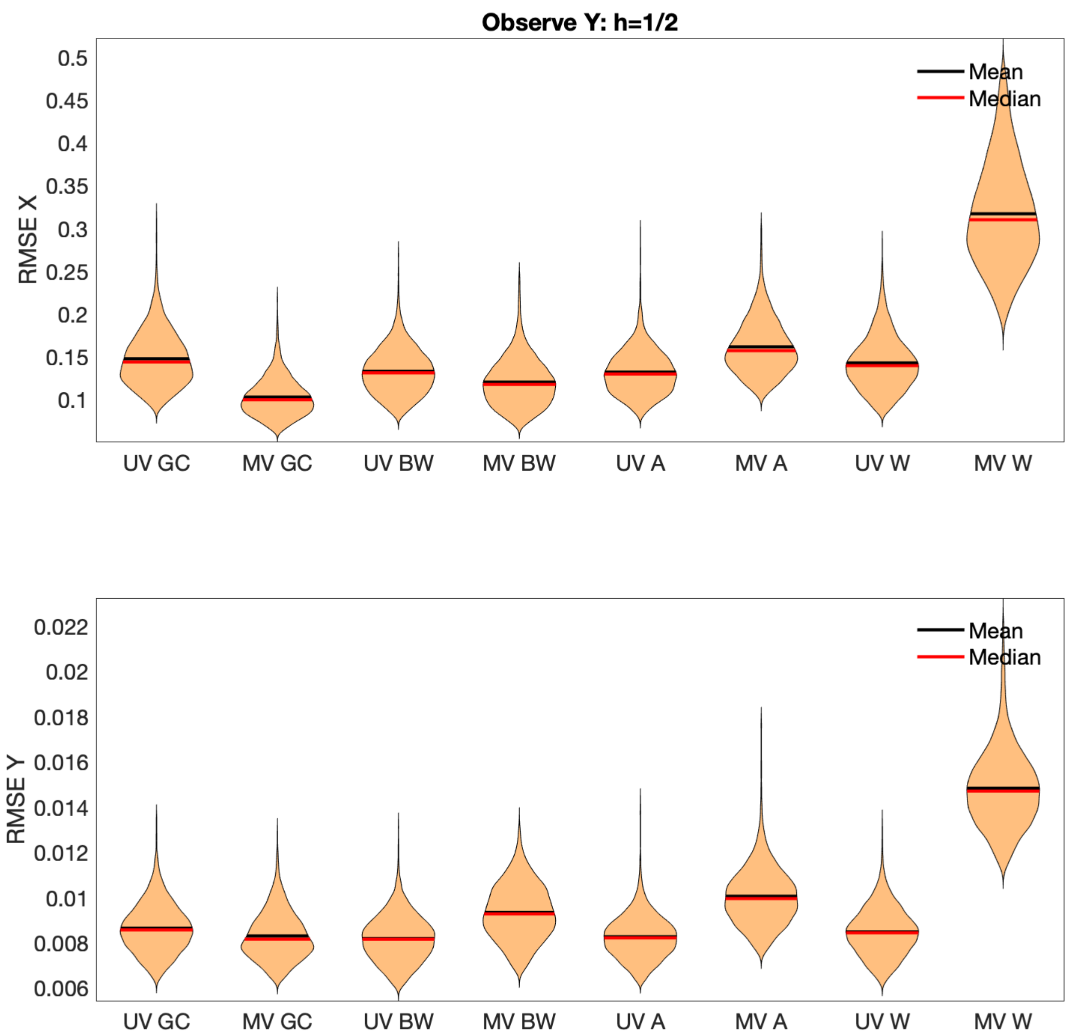

Figure 1: Observe only the Y process with coupling strength $h = 1/2$. UV stands for univariate and MV stands for multivariate.

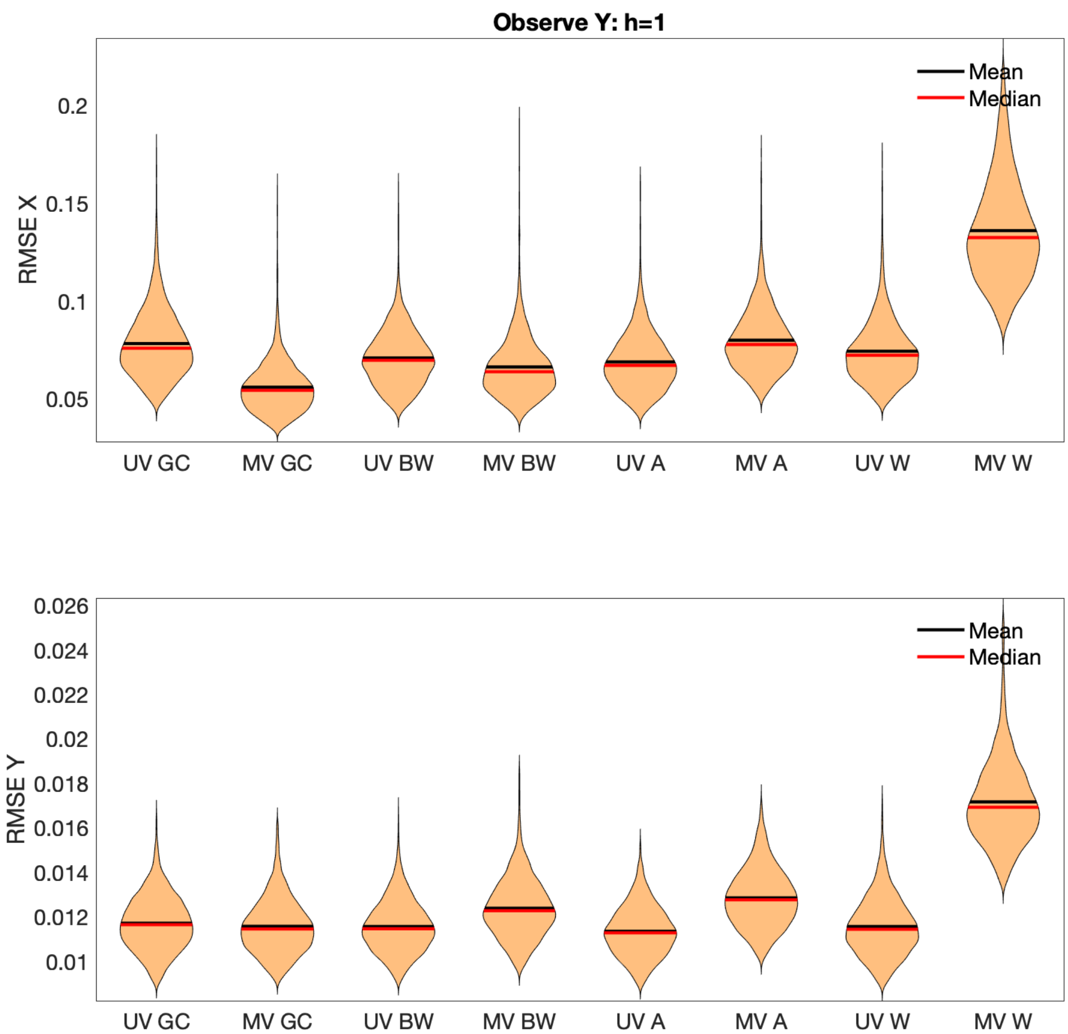

Figure 2: Observe only the Y process with coupling strength $h = 1$. UV stands for univariate and MV stands for multivariate.

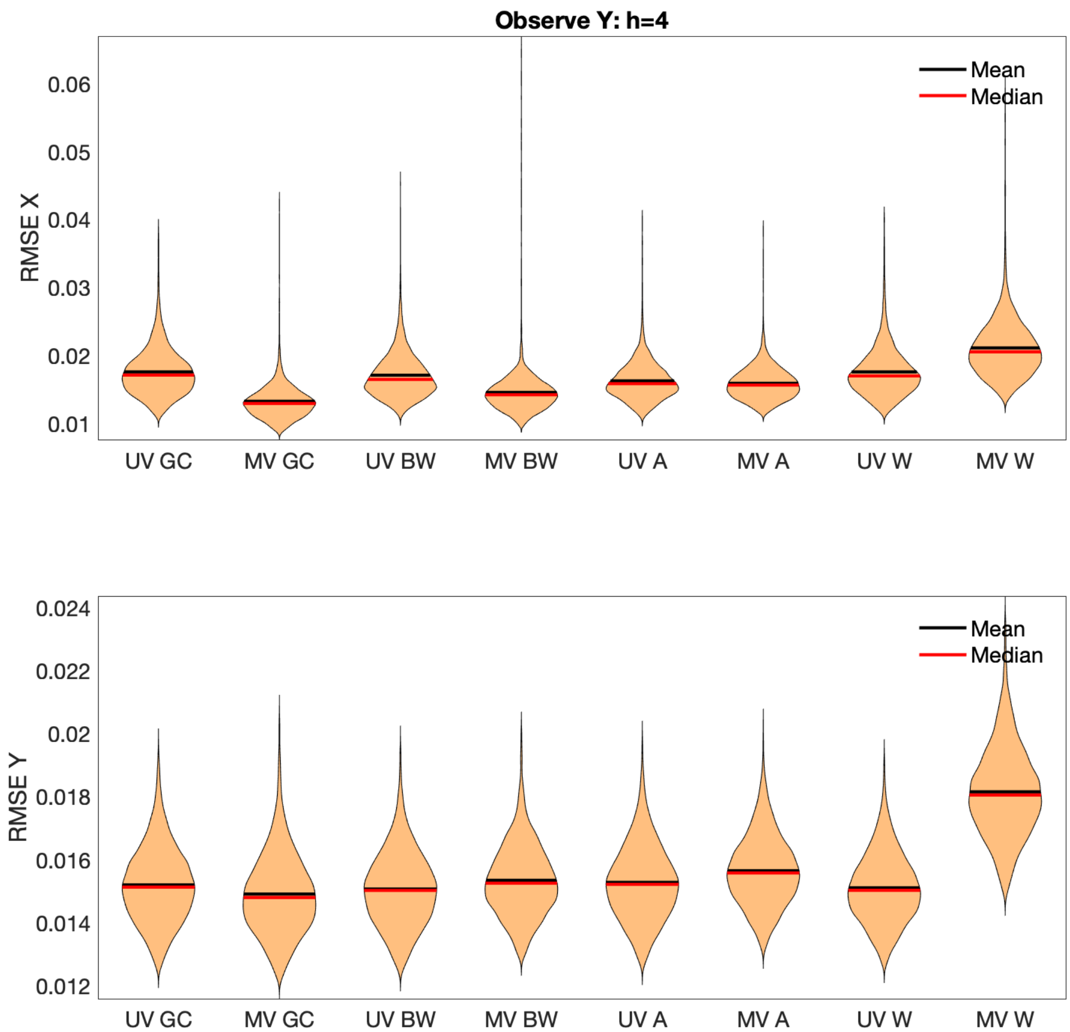

Figure 3: Observe only the Y process with coupling strength $h = 4$. UV stands for univariate and MV stands for multivariate.

3. A second concern is the restriction to observing only the fast dynamics. I would like to see a complete investigation examining the observation of fast-only, slow-only, and the full slow-fast coupled system. I'd like to see Figure 3 repeated for a few different scenarios, including those just mentioned, but also potentially varying parameters of the EnKF, such as the frequency of observations, the density of observations, the amount of observation noise, the length of the analysis cycle, etc. Not all results need to be reported in figures, but some indication that the authors have explored more variations in the problem specification would help to build confidence in the robustness of the final reported results.

   We have included experiments observing only the "long" X process and the full coupled system. When we observed only the long process, all localization functions led to very similar performance (Fig. 9). Note that since weak coupling is stable in this configuration we have included results from weakly coupled runs. In the paper we showed only the performance for univariate, multivariate, and weakly coupled Gaspari-Cohn and stated that all other localization functions performed similarly.

   Observing both processes, at least in our configuration, was quite unstable and often led to filter divergence. About 80% of the trials with weakly coupled localization functions led to catastrophic filter divergence. Trials with univariate and multivariate localization diverged less often, but still diverged about 20% of the time. However, in this configuration the precise shape of the localization function appears to have little impact. We did see differences between univariate, weakly coupled, and multivariate localization functions. Figure 10 shows results from only the trials (out of 50 total) which did not diverge. Weakly coupled localization appears to lead to the best performance, when the filter does not diverge. There is some variation in the results across the different localization functions. In particular, multivariate Askey appears to lead to better performance than weakly coupled Askey, but this may be attributable to the issues with stability. The complicated story with the weakly coupled schemes indicates that, in this configuration, filter performance is highly sensitive to the treatment of cross-domain background error covariances. The small ensemble size combined with small true forecast error cross-correlations can lead to spuriously large estimates of background error cross-covariances. Meanwhile, we have nearly complete observations of both processes, so within-component updates are likely quite good. Thus, zeroing out the cross terms, as in weakly coupled schemes, may improve state estimates. On the other hand, inclusion of some cross-domain terms appears to be important for stability.

   We varied the frequency of observations, length of analysis cycle, and density of observations between the "short"-only, "long"-only, and both "short" and "long" setups. In addition, we varied the observation noise in all setups. For both the "short"-only and the "long"-only we chose the observation error variance to be 5% of the climatological variance for the observed process. In Figs. 6 & 7 we bump up the observation error variance and use 20% of the climatological variance. The magnitude of the analysis errors changes, but the relative performance of the different localization functions is the same.

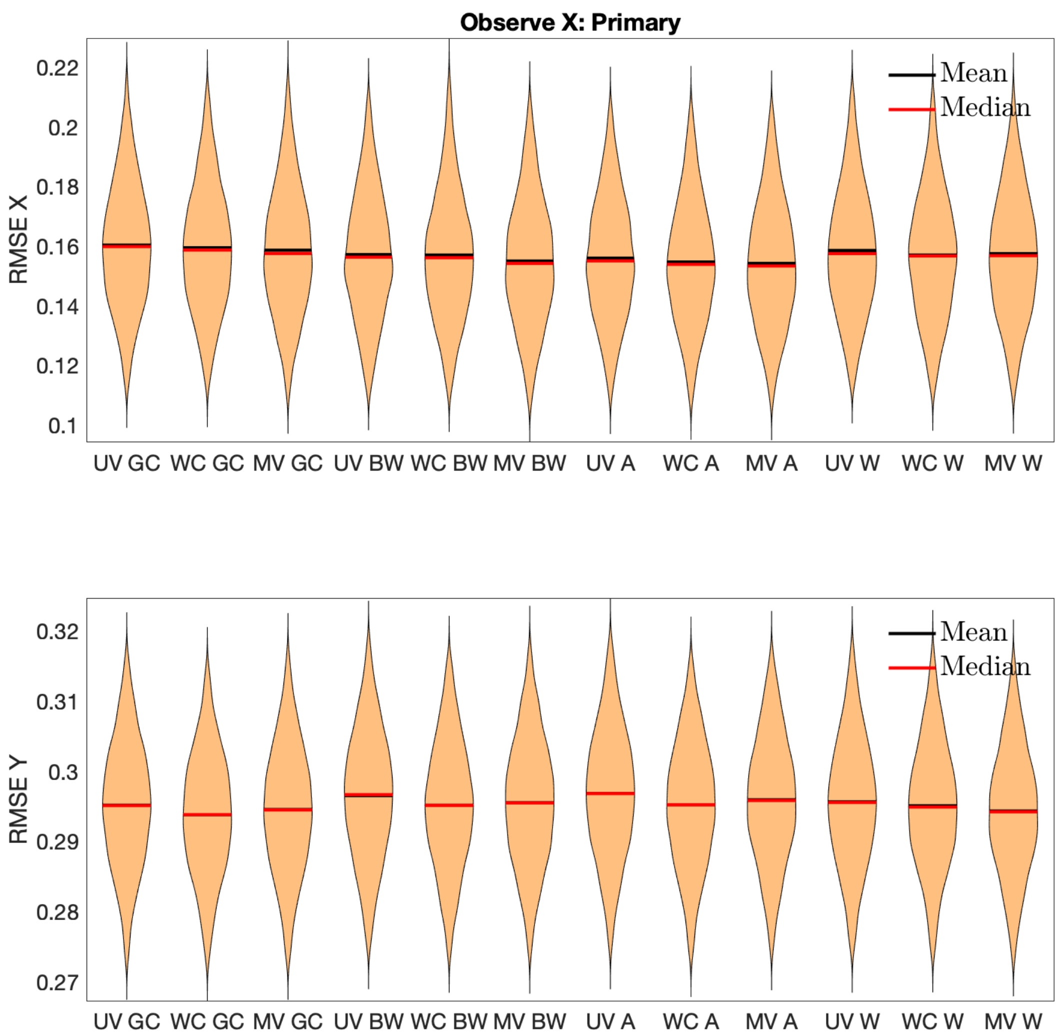

Figure 4: Observe only the X process. UV stands for univariate, WC stands for weakly coupled, MV stands for multivariate.

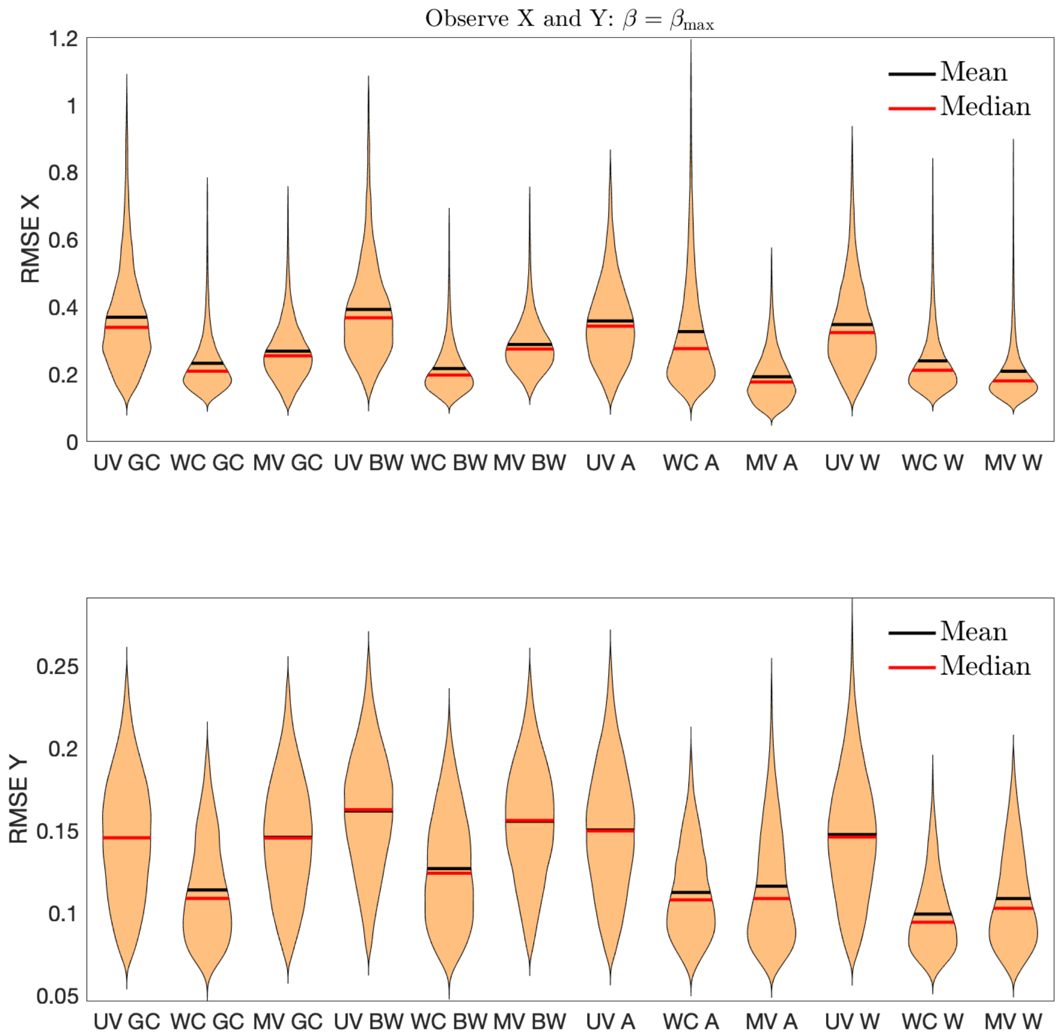

Figure 5: Observe both processes.

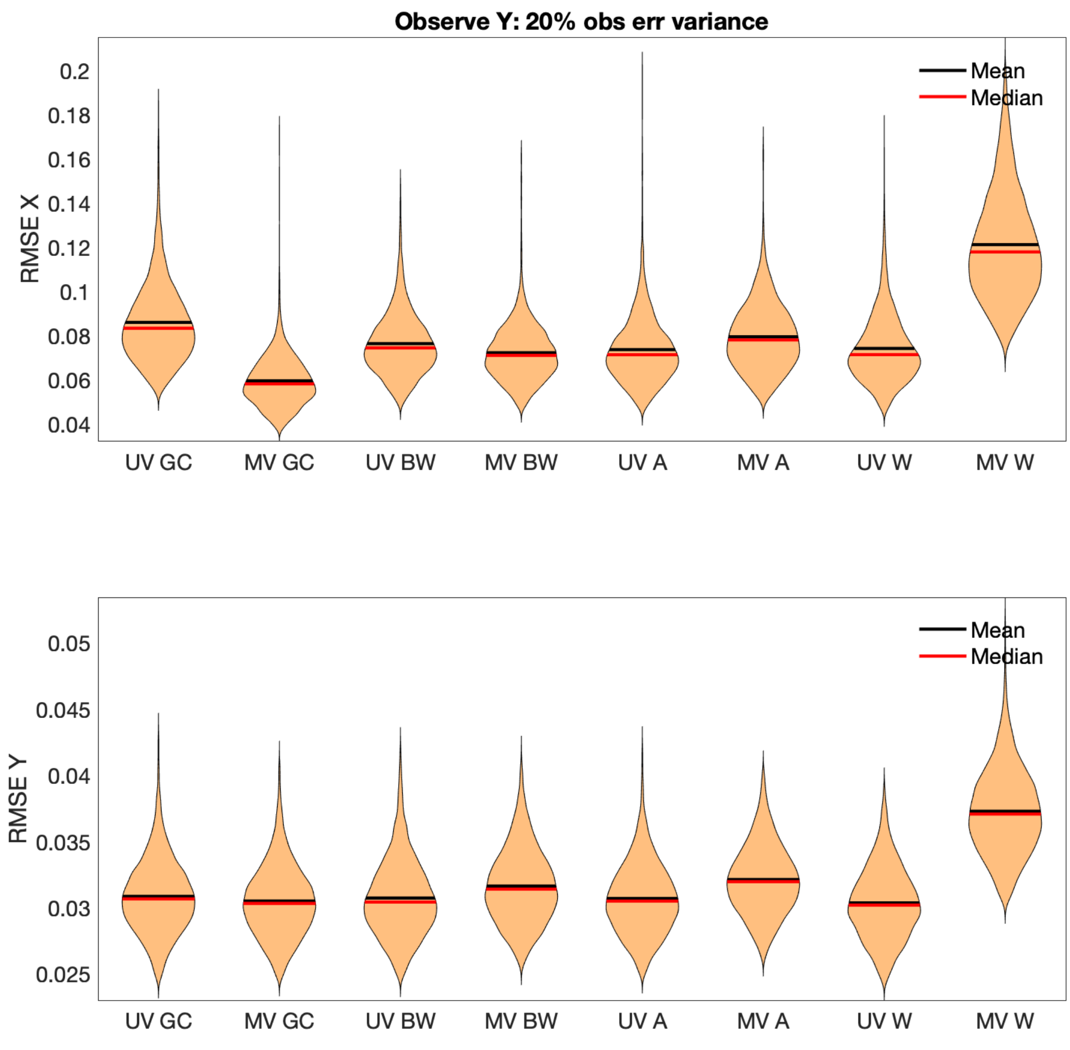

Figure 6: Observe “short” process only with larger observation error variance.

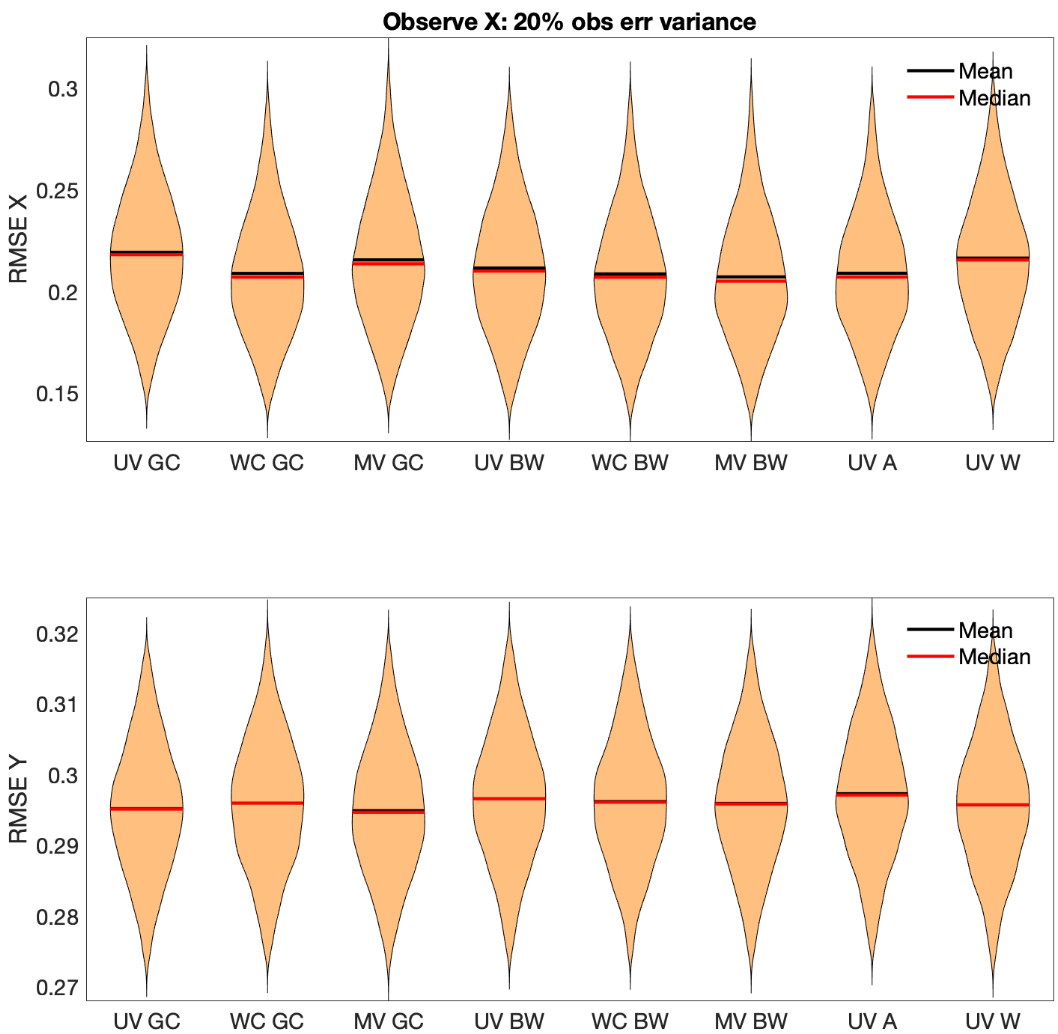

Figure 7: Observe “long” process only with larger observation error variance.

4. Minor issue: There are a few instances where the present tense is used when it should be past tense.

   We have corrected this.

Specific Comments:

1. L 8: "The functions produce non-negative definite localization matrices, which are suitable for use in variational data assimilation schemes." I think the term 'positive semidefinite' is more common, and the one originally used by Gaspari and Cohn (1999). I would suggest changing all instances of this throughout the manuscript.

   We changed non-negative definite to positive semidefinite throughout the manuscript.

2. L 14-16: "The background error covariance statistics stored in B dictate how information from observations propagates through the domain during the assimilation step (Bannister, 2008)" The term 'propagates' seems appropriate for 4D-Var, but perhaps not for all DA methods. More generally, the background error covariance provides a structure function that determines how observed quantities affect the model state variables, which is of particular importance when the state space is not fully observed.

   We changed the discussion of the impact of the background error covariance matrix on the analysis increment following the suggestion.

3. L 25: "Localization is typically incorporated into an ensemble estimate of B through a Schur (or element-wise) product." I would change this to say that localization is typically incorporated into the data assimilation in one of two ways - either through the B matrix using a Shur product, or through the observation error covariance R (e.g. Greybush et al., 2011). You are focusing on the localization applied directly to the B matrix.

   Greybush et al., 2011: Balance and Ensemble Kalman Filter Localization Techniques.

   https://journals.ametsoc.org/view/journals/mwre/139/2/2010mwr3328.1.xml

   We changed the description of B and R matrix localization and added a reference to Greybush et al., 2011.

4. L 32-33: "In Earth system modeling in particular, coupled DA shows improvements over single domain analyses (Penny et al., 2017; Zhang et al., 2020)" Additional sources that determined this point clearly are Sluka et al. (2016) and Penny et al. (2019):

   Sluka, T., S.G. Penny, E. Kalnay, and T. Miyoshi, 2016: Using Strongly Coupled Ensemble Data Assimilation to Assimilate Atmospheric Observations into the Ocean. Geophys. Res. Lett., 43, doi:10.1002/2015GL067238.

   Penny, S.G., E. Bach, K. Bhargava, C-C. Chang, C. Da, L. Sun, T. Yoshida, 2019: Strongly coupled data assimilation in multiscale media: experiments using a quasi-geostrophic coupled model. Journal of Advances in Modeling Earth Systems, 11. https://doi.org/10.1029/2019MS001652

   https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2019MS001652

   We changed the references for this statement.

5. L 35-37: "Schemes that include cross-domain [error] correlations in the B matrix are broadly classified as strongly coupled, which is distinguished from weakly coupled schemes where B does not include any nonzero cross-domain [error] correlations. The inclusion of cross-domain [error] correlations in B offers advantages" To be more precise, the term "cross-domain error correlations" should be used if referred to the error covariance matrix B.

   We changed "cross-domain correlations" to "cross-domain error correlations".

6. L 55: I'll note that Lorenz himself cited this as (Lorenz, 1996). See comment below regarding line 461.

   We changed all relevant references to Lorenz, 1996.

7. L 57-58: "We find that, in our set up, artificially decreasing the magnitude of the cross-domain correlation hinders the assimilation of observations." This is a positive sign for the advancement strongly coupled DA, but I wonder if this could be partly due to the use of the Lorenz system II, which has some highly linear relationships between the small and large scale systems. Some discussion was given, for example, in this blog post by Stephan Rasp:

https://raspstephan.github.io/blog/lorenz-96-is-too-easy/

The linked blog post specifically refers to the need for more realistic models when parameterizing subgrid-scale processes. Coupling between scales is typically nonlinear so the linear coupling between X and Y in the Lorenz-96 model is not a good proxy. But coupling between fast and slow components of a system can still be approximately linear, as in the sensible heat exchange between ocean and atmosphere that is approximately linearly proportional to the temperature difference. It remains to be seen how an intermediate complexity model interacts with the design of methods for data assimilation, like localization. To address the limitations of the bivariate Lorenz system we have verified our results with different coupling strengths. See response to general comment 2.

8. L 61: "localization function[s] from the literature."

   Fixed.

9. L 77-78: "A fundamental difficulty in localization for strongly coupled DA is how to propose a cross-localization function LXY to populate both LXY and LYX" It might be useful to explain at this point which term controls the effect of system X on Y, and Y on X.

   We added an explanation of this in the first paragraph where we introduce $\mathbf{P^b_{XY}}$ and $\mathbf{P^b_{YX}}$.

10. L 102: "we define two processes Zj , j = X,Y" I understood this on the third read through. Perhaps the authors could reword this sentence slightly to make it more clear. For example, "we define two processes Zj, where j can represent either X or Y" Or simply, "we define two processes Zj, with j=X,Y"

    Changed following suggestion.

11. L 105-106: " Thus LXX, LYY ,LXY form a multivariate covariance function, and hence a multivariate, non-negative definite function" Based on the terminology defined so far, I'm not sure how to interpret the triple (LXX,LYY,LXY) forming a single function. Perhaps a line or two could be added to explain this step.

    $\{\mathcal{L}_{XX}, \mathcal{L}_{YY}, \mathcal{L}_{XY}\}$ **are components of a single *multivariate* function $\mathbb{R} \to \mathbb{R}^{2\times2}$. We expanded the introduction of multivariate positive semidefinite functions and put brackets around the collection of component functions to emphasize their interpretation as a single multivariate function.**

12. L 118: The way I am interpreting the notation is that the term $(1 - r/c)_+$ is zero when the term in parentheses is less than or equal to 0, which would occur when $r >= c$. Can the authors explain the comment about the convolution being zero at distances greater than 2c in line 120, it is not immediately obvious.

    The kernel $k(r) = (1 - r/c)_+$ is zero when $r \geq c$. The convolution $[k * k]$ is zero at distances greater than the sum of the kernel radii, i.e. $r \geq 2c$. Two kernels separated by this distance are never overlapping and hence the convolution is 0. We have changed some of the language in the paper to clarify this point.

13. L 139: "who perform[ed] the"

    Changed.

14. L 140: "in never develop[ed] multivariate"

    Changed.

15. L 156: "Porcu et al. (2013) develop[ed] a multivariate version" "et al." is short for the Latin term "et alia," meaning "and others." It is strange to reference the actions of Porcu "and others" in the year 2013 using present tense.

    Changed.

16. L 157: "Roh et al. (2015) [found] that"

    Changed.

17. L 159: "Daley et al. (2015) extend[ed] the work"

    Changed.

18. L 166: “with B the beta function” Could you define this here for clarity.

    We defined the beta function.

19. L 168: “Daley et al. (2015) [gave]”

    Changed.

20. L 184: “This approach leads to a “weakly” coupled scheme, which is not the focus of this work.” I understand this may not be the focus, but it seems that it would be appropriate to compare to this approach given that the weakly coupled DA scheme is the standard approach for current operational forecast systems.

    In the original manuscript, when we observed only the “short” process, weak coupling led to catastrophic filter divergence, so there wasn’t much else to say. We now compare all localization functions to their weakly coupled counterparts in our new experiments. In the “long”-only and both “short” and “long” experiments we present results from weakly coupled schemes that do not blow up.

21. L 184-186: “Additionally, in our setup we observe only one of the two processes and we find that when the assimilation is not allowed to update the unobserved process the result is prone to catastrophic divergence” It might be appropriate to perform a few experiments where both components are observed, and results are compared using weakly and strongly coupled DA.

    Done. See comment directly above.

22. L 200/202/205: “ Lorenz (199[6])” See comment below for line 461.

    Changed.

23. L 203: “using an adaptive fourth-order Runge-Kutta method” Perhaps provide a citation for the method.

    Added citation for the method.

24. L 204-205: “The solutions are output with a time interval of 0.005 nondimensional units, or 36 minutes” It seems strange to say there are non-dimensional units and then indicate that it is the same as 36 minutes. Perhaps repeat some of the justification from Lorenz to indicate the relative error growth rates and its relation to more realistic applications that would be approximately equivalent to 36 minutes in operational prediction in the early 1990’s.

    We replaced “nondimensional units” with “model time units” and clarified that Lorenz found 0.005 model time units to be similar to 36 minutes in more realistic settings.

25. Figure 2 caption: “setup” is a noun that means “the way in which something… is organized, planned, or arranged.” This should probably be used in most places where the authors current use two words: “set up“.

    Changed.

26. L 209: “Increasing the coupling strength leads to larger covariances between the forecast errors in processes X and Y , thereby making the effect of cross-localization more pronounced and easier to study.” I believe this is the case. However, I would like to see some sensitivity study of how the benefit of strongly coupled DA paradigm breaks down as the coupling strength between the two components weakens and asymptotes to 0.

    When we vary the coupling strength the magnitude of the analysis errors changes, but the relative performance of the different localization functions does not change. See response to general comment 2.

27. L 213: “We choose to place the variable Xk in the middle” Does the placement of the X variable have any influence on the results of localization? Is there any sensitivity here, or are the results generally the same regardless of how the placement of the X and Y variables are interpreted?

    Placing X_k at the beginning of the sector, as in Roh et al. (2015), means that half of the nearby Y variables are nearly uncorrelated with X_k. The analysis errors are larger when X_k is placed at the beginning of the sector rather than the middle of the sector. However, the relative performance of all the localization functions is the same in both cases. Figure 11 shows the distribution of analysis errors when we observe the “short” process with X_k placed at the beginning of the sector.

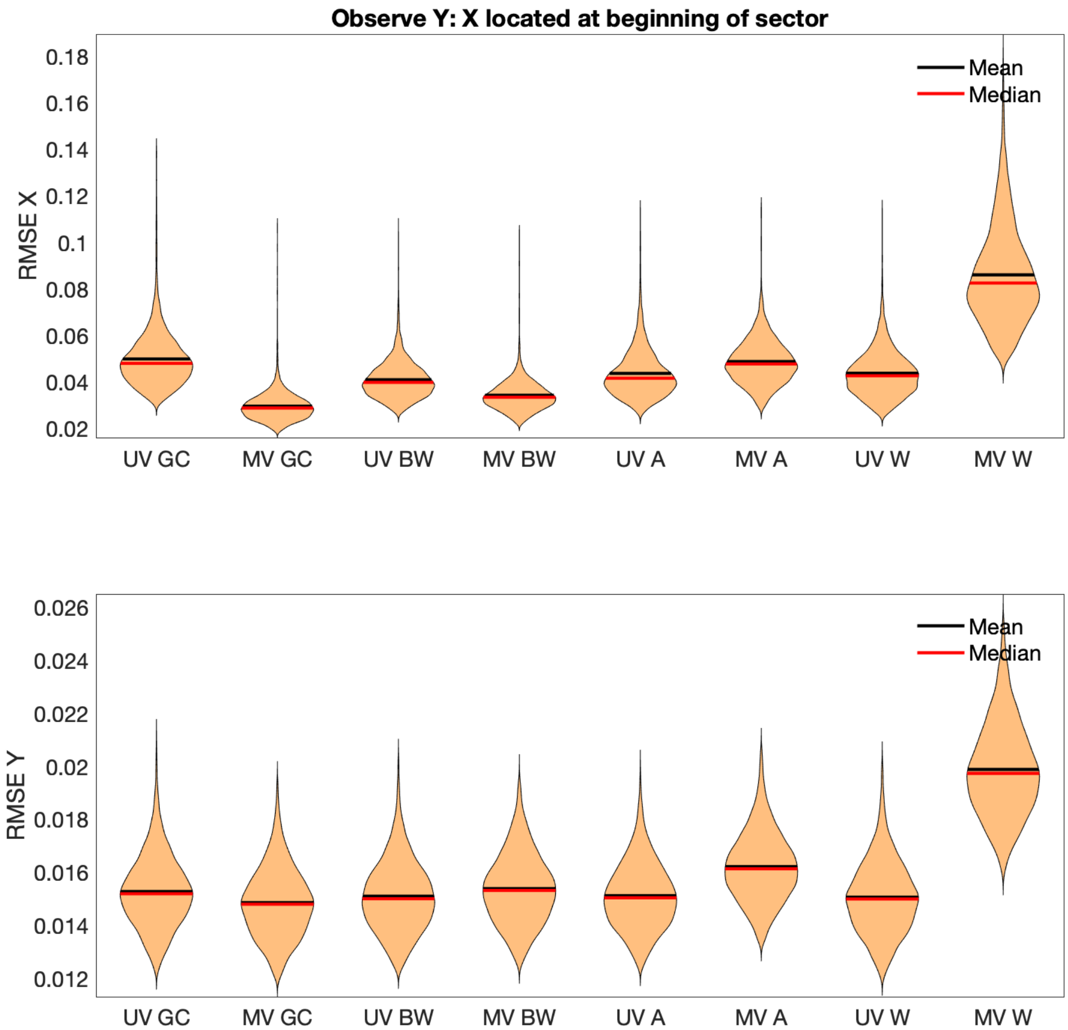

Figure 8: Move the location of X_k to the beginning of the sector, instead of the middle of the sector.

28. L 218: “We develop localization functions for EnVar schemes where non-negative definiteness of the localization matrix is essential to ensure convergence of the numerical optimization. Since the minimizer of the 3D-EnVar objective function is the same as the EnKF analysis mean in the case of linear observation (Lorenc, 1986), in this experiment we make use of the EnKF rather than implement an ensemble of 3D-EnVar assimilation scheme (Evensen, 1994; Houtekamer and Mitchell, 1998; Burgers et al., 1998)” This discussion is a bit confusing. I think it could be cleaned up with a little reorganization, e.g.

    We develop localization functions for data assimilation schemes that rely on Schur product modification of the background error covariance matrix B. In our experiments we use the stochastic EnKF (Evensen, 1994; Houtekamer and Mitchell, 1998; Burgers et al., 1998). However, because the minimizer of the 3D-EnVar objective function is the same as the EnKF analysis mean in the case of linear observations (Lorenc, 1986), our results translate to EnVar schemes as well. The positive semi-definiteness of the localization matrix is essential to ensure convergence of the numerical optimization methods used to implement EnVar [cite].

    The section on equivalence between EnKF analysis mean and the 3D-EnVar minimizer in the case of linear observations has been rewritten following the suggestion and has been moved to the conclusions.

29. L 230-231: “In this experiment we use the adaptive inflation scheme of El Gharamti (2018) and apply the inflation to the prior estimate.” Can this be added to the EnKF equations above for clarity?

    We expanded the description of the inflated ensemble.

30. L 233-234: “We run each DA scheme for 3,000 time steps, discarding the first 1,000 time steps and reporting statistics from the remaining 2,000 time steps.” Is this referring to model time steps, or the number of analysis cycles?

    This is referring to the number of analysis cycles. We changed “time steps” to “analysis cycles” in the manuscript to clarify this.

31. L 235-237: “The observation operator H is such that all of the Y variables are observed, and none of the X variables are observed. In this way we can isolate the effect of the localization on the performance of the filter for the X variable.” This means you are observing the fast dynamics and using this to update the slow dynamics through the error covariance statistics. This has been shown effective in a number of studies exploring strongly coupled DA. Penny et al. (2019) showed that the reverse was also possible, particularly if the size of the analysis window is decreased (or the frequency of observation updates is increased).

    We have included an experiment that observes only the “long” process. See response to general comment 3.

32. L 256: “, so that we hypothesize that GC allows” Change to: “, so we hypothesize that GC allows”

    Changed.

33. L 266-267: “ By contrast, the BW and Askey functions show virtually no difference between the multivariate and univariate versions” The BW method looks slightly improved, and the Askey method slightly degraded.

    This is true, however the difference is not statistically significant. This sentence has been updated to reflect this distinction.

34. L 280: It is up to the authors, but generally the conclusions reads more clearly if this is now written in past tense. E.g. “In this work, we develop[ed]. . . ” “We compare[ed] multivariate GC to three. . . ” “We [found] that, in a toy model. . . ” “In this work we investigate[d]. . . ” “We [found] that this. . . ”

    Changed to past tense in the conclusions.

35. L 284: “the localized estimate of the background [error] covariance matrix”

    Fixed.

36. L 294: “A natural application of this work is localization in a coupled atmosphere-ocean model. Multivariate GC allows for within component covariances to be localized with GC exactly as they would be in an uncoupled setting, using the optimal localization length scale for each component Ying et al. (2018). In this work we discuss the importance of the cross-localization radius in determining performance. However, this work does

not address the question of optimal cross-localization radius selection, which is an important area for future research"

This is certainly of interest - are there any conjectures that can be made about the applicability of the results here extending to an application like a coupled atmosphere-ocean model?

We do not have any conjectures at this moment, but we are actively working on testing the multivariate Gaspari-Cohn and Bolin-Wallin functions in a coupled atmosphere-ocean model.

While the interpretation of localization is clear in the within-component covariances, how would you interpret localization on the cross-component covariances?

Could you clarify your interpretation of within-component localization? In our minds, within- and cross-component localization have very similar interpretations. They are both tapering a covariance between variables based on distance, with the only difference being that in cross-component localization the two variables are in different model components.

Could a situation in which it might be desirable for an atmospheric observation to have an influence on the ocean state but not vice versa create difficulties with the symmetry relied on above in forming LXX, LYY, and LXY as a triple, and the need for maintaining positive semi-definiteness?

A setup where the atmosphere influences the ocean state, but not vice versa would necessarily be associated with a background error covariance matrix which is not symmetric (which is not possible). Variational methods require symmetric B matrices (in any case, the objective function ignores the antisymmetric component of B) and hence this setup would not be possible with a variational scheme. This kind of asymmetric setup is possible by using asymmetric localization of the Kalman gain in EnKF-type methods though.

37. L 461: The full citation for Lorenz-96 is not given. It should be Lorenz (1996): Lorenz, E.N., 1996: Predictability—A problem partly solved.Proc. Seminar on Predictability,Vol. 1, Reading, Berkshire, UnitedKingdom, ECMWF, 1–18. Note that Lorenz cited it himself this way in:

    Lorenz, E.N., 2005. Designing chaotic models. J. of the Atmos. Sci. 62, 1574–1587. DOI:10.1175/JAS3430.1.

    Full citation is now correctly given.

We thank Reviewer #2 for their time in reading our manuscript and providing a thoughtful critique. The reviewer's comments and our responses are below.

The paper by Stanley et al. develops multi-scale extensions to the traditional families of the parameterized localization functions such as the Gaspari-Cohn 5th order polynomial. A key contribution of the paper is to note that each of these polynomial expressions can be represented as a product of the square-root kernels that can be cross-multiplied to achieve a positive-definite cross-scale localization. Authors, correctly, draw relevance of these new techniques to the problem of cross scale localization encountered in coupled data assimilation. Authors correctly choose the right type of the test problem that is useful-enough to test the mathematics of the developed extensions yet is not too complex to obscure the interpretation of the results. I agree with authors that there is no need to over-interpret the results of this simple experiment with regard to its relevance in a more complex ocean-atmosphere problems.

I found this article very relevant, well-written, with adequate experimental plan, and appropriate interpretation of the experimental results. I congratulate the authors on a nice contribution to the literature and suggest this paper for publication after minor revision.

Summary of suggested changes:

1. This contribution parallels the work of Mark Buehner on multi-scale localization. I suggest that authors draw this parallel by referencing some of his work such as [https://doi.org/10.3402/tellusa.v67.28027]. By including these references, authors can then discuss how their work can also be related to the problem of multi-scale localization (e.g. such as assimilation in convective-resolving models).

   We added a paragraph to Section 2.1 discussing the similarity between our work and the work of Buehner and Shlyaeva.

2. I made some minor suggestion to how authors might consider changing or extending other references in the introduction section (see pdf attached).

   See responses to minor suggestions below.

3. I tried to read the paper from the perspective of someone who might want to implement some of the localization formulas discussed by the authors. I made some suggestions on clarifications. I highly appreciate that authors published the source code for their work. I suggest that authors mention that in the main body of the paper.

   We now mention that code is published in main body of paper.

Minor suggestions from attached pdf:

1. L32: "Penny et al., 2017" this is just a workshop report i dont think it has any results to support your statement. Maybe reference to paper by Sluka et al. 2016 or Frolov et al. 2016 will be more appropriate.

   Changed to suggested references.

2. L34: "Cross-domain correaltions ... (Penny et al., 2017)" Is this the same as flow-dependent? If yes, then some reference to works by Polly Smith will be approbate. You can also reference the following paper https://doi.org/10.1175/MWR-D-20-0352.1.

   Changed reference to Smith et al. 2017 and Frolov et al. 2021

3. Sec 2.2: I realize that you came to this idea independently. However, similar ideas are also common in the field of multi-scale localization. See papers by Mark Buehner. It would be good to mention this parallel.

   Buehner and Shlyaeva (2015) indeed construct localization matrices that have similar form to the localization matrices we construct, i.e.

   $$\mathbf{L} = \begin{bmatrix} \mathbf{L_{XX}} & \mathbf{L_{XY}} \\ \mathbf{L_{YX}} & \mathbf{L_{YY}} \end{bmatrix}. \tag{1}$$

   However their method differs from ours in that they do not construct analytic formulas. Rather, the construct the off-diagonal blocks of the L matrix through $\mathbf{L_{XY}} = (\mathbf{L_{XX}})^{1/2} (\mathbf{L_{YY}})^{T/2}$. This formulation can be related to a discrete approximation to kernel convolution (a perspective that we are developing), but this was not explored by Buehner and Shlyaeva (2015). Constructing the off-diagonal blocks through matrix square roots is appropriate for scale-dependent localization where $\mathbf{L_{XX}}$ and $\mathbf{L_{YY}}$ are of the same

size. It is not immediately obvious how to extend this to multiple spatial domains, as we have in strongly coupled data assimilation. We have mentioned the parallel to multi-scale localization and cited Buehner and Shlyaeva (2015) in Section 2.1 where we introduce multivariate localization.

4. Sec 2.3: You can point out the similarity and differences of your results from Frolov et.al. 2016. Similar to Frolov et.al. 2016, the localization scale for cross-scale localization is the average of the scales in the coupled.

   Added reference to Frolov et al. (2016) and commented on the cross-localization radius they used.

5. Fig 1: please add a comment to the figure why Lx,x seem to drop to zero faster for Wendland than for any other function. Otherwise it looks like you miss-specified the localization radius for Wendland.

   Wendland uses the same $R_{XX} = 45$ value as the other functions. You are correct in observing that it drops to near 0 more quickly than the others, however it does not reach 0 until $d = 45$. We added a sentence to the caption stating the localization radii for all functions.

6. L145: what does κ^{-3} stands for? I am not sure i know how to compute that.

   Added sentence clarifying that $\kappa^2 = \frac{\max\{R_{XX}, R_{YY}\}}{\min\{R_{XX}, R_{YY}\}}$ is a ratio of length scales.

7. Eq. 9: seems like something went wrong with typesetting here.

   Fixed the issue with the '.' appearing in awkwardly at the end of the equation.

8. Eq. 13: I am really confronted by the complexity of this formula and what would it take to implement it in practice. Not sure I follow how to specify ν or γ. Maybe this is something you can explicitly mention after Eq. 13.

   The parameters ν and $\{\gamma_{ij}\}$ are related to the shape of the localization functions (Eq. 10 & 11), and are necessary to guarantee positive semidefiniteness in a given spatial dimension. In general it makes sense to take ν as small as possible while still guaranteeing positive semidefiniteness in the space (for us the space is $\mathbb{R}^3$) and then allow γ_{ij} to vary so that each component localization function can have a different shape. Once ν and $\{\gamma_{ij}\}$ are chosen, Eq. (13) gives the formula for the maximum cross-localization weight factor (i.e. height of the cross-localization function at distance $d = 0$) to guarantee positive semidefiniteness. We added a sentence to section 2.5 describing the parameters ν and γ_{ij}.

We thank Reviewer #3 for their time in reading our manuscript and providing a thoughtful critique. The reviewer's comments and our responses are below.

General Comments:

This manuscript presents a new multivariate extension of the standard Gaspari-Cohn localization function which is compared with 3 other multivariate functions, plus their univariate versions. These techniques are extremely relevant to problem of cross-domain localization in strongly coupled data assimilation and this work is an encouraging step towards developing appropriate methods for such systems. The localization techniques are illustrated using the bivariate Lorenz 96 system.

Whilst it would be nice to have an example illustrating how the new multivariate GC method translates to a more realistic system I appreciate that it is always important to test new ideas like this in a relatively simple system, where results can be more easily interpreted. I hope that the authors have the opportunity to extend this work to a more complex coupled model system in the future.

It is always good to see new work addressing issues related to the application of coupled DA. The article is timely, highly relevant and clearly written; it will make a nice contribution to the coupled DA literature. I suggest it is published after minor revisions.

We are actively working on extending this work to a coupled atmosphere-ocean model.

Specific comments:

1. It is a shame that results were only shown for case where the fast (Y) component is fully observed, and further that the performance of each method was only measured/ reported in terms of the RMSE of the X (slow, unobserved) component. I would like to see some results from experiments where only the X (slow) component is observed, and also where both the X and Y components are observed, both fully and partially. I appreciate that this would potentially increase the number of figures/length of the manuscript, but it may not be necessary to explicitly show all the results. A brief discussion of the results in order to confirm that the general conclusions still hold under different observing scenarios would give the reader greater confidence in the performance of the new GC method.

   We have included experiments observing only the "long" X process and the full coupled system. When we observed only the long process, all localization functions led to very similar performance (Fig. 9). Note that since weak coupling is stable in this configuration we have included results from weakly coupled runs. In the paper we showed only the performance for univariate, multivariate, and weakly coupled Gaspari-Cohn and stated that all other localization functions performed similarly.

   Observing both processes, at least in our configuration, was quite unstable and often led to filter divergence. About 80% of the trials with weakly coupled localization functions led to catastrophic filter divergence. Trials with univariate and multivariate localization diverged less often, but still diverged about 20% of the time. However, in this configuration the precise shape of the localization function appears to have little impact. We did see differences between univariate, weakly coupled, and multivariate localization functions. Figure 10 shows results from only the trials (out of 50 total) which did not diverge. Weakly coupled localization appears to lead to the best performance, when the filter does not diverge. There is some variation in the results across the different localization functions. In particular, multivariate Askey appears to lead to better performance than weakly coupled Askey, but this may be attributable to the issues with stability. The complicated story with the weakly coupled schemes indicates that, in this configuration, filter performance is highly sensitive to the treatment of cross-domain background error covariances. The small ensemble size combined with small true forecast error cross-correlations can lead to spuriously large estimates of background error cross-covariances. Meanwhile, we have nearly complete observations of both processes, so within-component updates are likely quite good. Thus, zeroing out the cross terms, as in weakly coupled schemes, may improve state estimates. On the other hand, inclusion of some cross-domain terms appears to be important for stability.

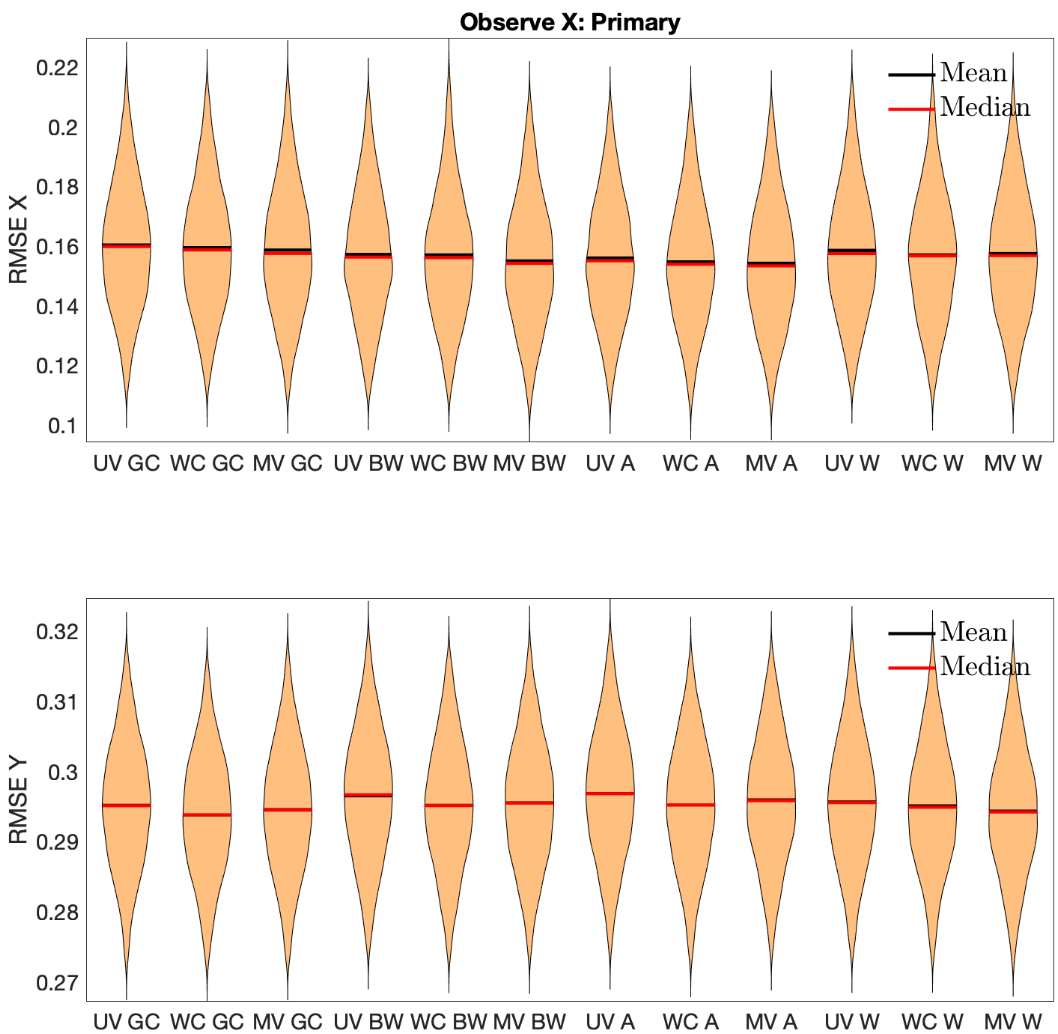

Figure 9: Observe only the X process. UV stands for univariate, WC stands for weakly coupled, MV stands for multivariate.

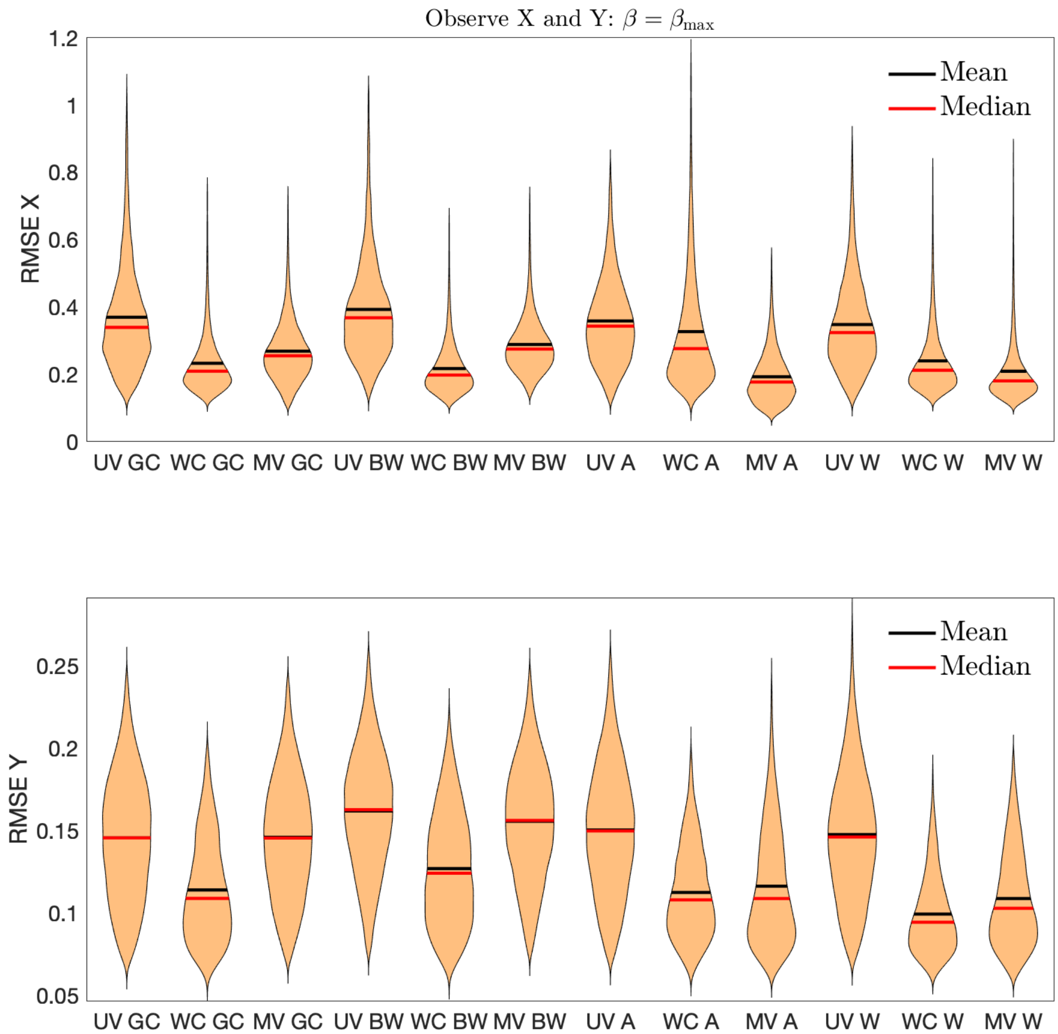

Figure 10: Observe both processes.

2. I am not entirely clear on how the univariate localization functions were implemented. Lines 181-182 state: ”We compare the four multivariate localization functions in Sect. 2 to a simple approach to localization in coupled DA, which is to use the same localization function for all model components. We call this approach univariate localization.” I think this means that each block of the localization matrix L uses the same localization function and radius for all blocks, rather than a different radius for the X and Y blocks and a different function and radius for the cross X,Y block, is this correct? I think what is confusing is that you are calling it univariate localization but you are actually localizing the cross XY blocks of the matrix B. Perhaps this needs to be stated more explicitly somewhere. In systems with very different error correlation scales this type of univariate localization function could be not really be expected to perform well.

   Your interpretation is correct. We have changed the introduction to univariate localization (first paragraph of section 3) so that it is now explicitly stated that the same function is used to localize all matrix blocks.

Minor comments:

1. The references are a bit strange – there are multiple web links for a lot of the papers; the https://doi.org/xxx link will be sufficient in most cases.

   References have been fixed and web links have been removed.

2. Further minor comments and technical corrections are marked in the attached pdf.

   See list below for changes made based on suggestions in pdf.

Minor comments from attached pdf:

1. L27: “non-negative definite” positive semi-definite is the more common terminology

   Changed to positive semidefinite.

2. L35: add reference to Smith et al 2017, https://doi.org/10.1175/MWR-D-16-0284.1

   Added reference.

3. L53: “EnVar” Ensemble-Variational - You need to introduce this type of abbreviation the first time you use it.

   Changed.

4. L107: “$R_{XY} = \frac{1}{2}(R_{XX} + R_{YY})$” is this choice required for the matrix L to be positive semi-definite?

   For Gaspari-Cohn, Bolin-Wallin, and other functions created through kernel convolution, there R_{XY} is completely determined by R_{XX} and R_{YY} and is equal to $R_{XY} = \frac{1}{2}(R_{XX} + R_{YY})$. We have changed our discussion in Sec. 2.2 to clarify this point.

5. Fig 1: Does Wendland use the same R_x value? Maybe add values of R_x, R_Y and R_{XY} in the caption

   Wendland does use the same $R_{XX} = 45$ value as the other functions. You are correct in observing that it drops to near 0 more quickly than the others, however it does not reach 0 until $d = 45$. We added a sentence to the caption stating the localiztaion radii for all functions.

6. Eq. 10: what are ν, $\gamma_{i,j}$? How are they chosen/ defined?

   The parameters ν and $\{\gamma_{ij}\}$ are related to the shape of the localization functions, and are necessary to guarantee positive definiteness in a given dimension. A sentence has been added to describing these parameters.

7. L166: “B the beta function” This needs to be defined or an appropriate reference added. Also, this isn’t the greatest choice of notation given that the background error covariance matrix is denoted B

   We defined the beta function and changed the notation for the background error covariance matrix from B to $\mathbf{P^b}$.

8. L174: "we choose $R_{XY} = \min\{R_{XX}, R_{YY}\}$" has this value been experimented with?

   To maintain positive semidefiniteness we require $R_{XY} \leq \min\{R_{XX}, R_{YY}\}$. From our understanding of the true forecast error correlations we estimate that for optimal performance R_{XY} should be at least as large as R_{YY} (which is the smaller of the two within-component localization radii). For this reason, we have not experimented with smaller values for R_{XY}.

9. L181: grammar?

   This section has been reworded.

10. L184: "Additionally in our setup we observe only one of the two processes and we find that when the assimilation is not allowed to update the unobserved process the result is prone to catastrophic divergence." This is interesting ... it would be good to understand why this occurs. Does it depend on the temporal/ spatial frequency of the obs?

    We have yet to find a stable configuration where we observe only one process and do not allow any updates to the unobserved process. It is possible that we could increase the coupling strength enough that the unobserved process would synchronize with the observed process, but short of that it is difficult to imagine a stable configuration when one process is entirely detached from observations.

11. L200: "so there are 10 times as many long variables as short variables" 10 times more short variables than long variables

    Changed.

12. L204: how does .005 translate to 36 minutes?

    We rewrote this sentence to clarify that this is a result from Lorenz (1996). The reasoning used in Lorenz (1996) is based on relative error growth rates. A full discussion of the reasoning from Lorenz (1996) is superfluous to the discussion here and is hence not included.

13. L213: "We choose to place the variable X_k in the middle of the sector" is this choice important to the performance of the localisation?

    Placing X_k at the beginning of the sector, as in Roh et al. (2015), means that half of the nearby Y variables are nearly uncorrelated with X_k. The analysis errors are larger when X_k is placed at the beginning of the sector rather than the middle of the sector. However, the relative performance of all the localization functions is the same in both cases. Figure 11 shows the distribution of analysis errors when we observe the "short" process with X_k placed at the beginning of the sector.

14. L217: This paragraph is clunky - needs revising

    This paragraph has been rewritten following this suggestion and a suggestion from S. Penny.

15. L220: observation operator?

    The changed paragraph now says "linear observations".

16. L221: delete "ensemble of"

    Deleted in rewrite of paragraph.

17. L234: How frequent in time are the obs?

    In the original experiment in the first submission, we observe every 0.005 model time units.

18. L260: "Figure 3 shows RMSE for process X" what about Y? How is RMSE computed? vs a 'truth'?

    Yes that is correct. RMSE is computed vs. a 'truth', which is also used to generate the 'observations'. We added a sentence clarifying this.

19. Eq. A15 ?

    Could the reviewer clarify what about this expression is confusing?

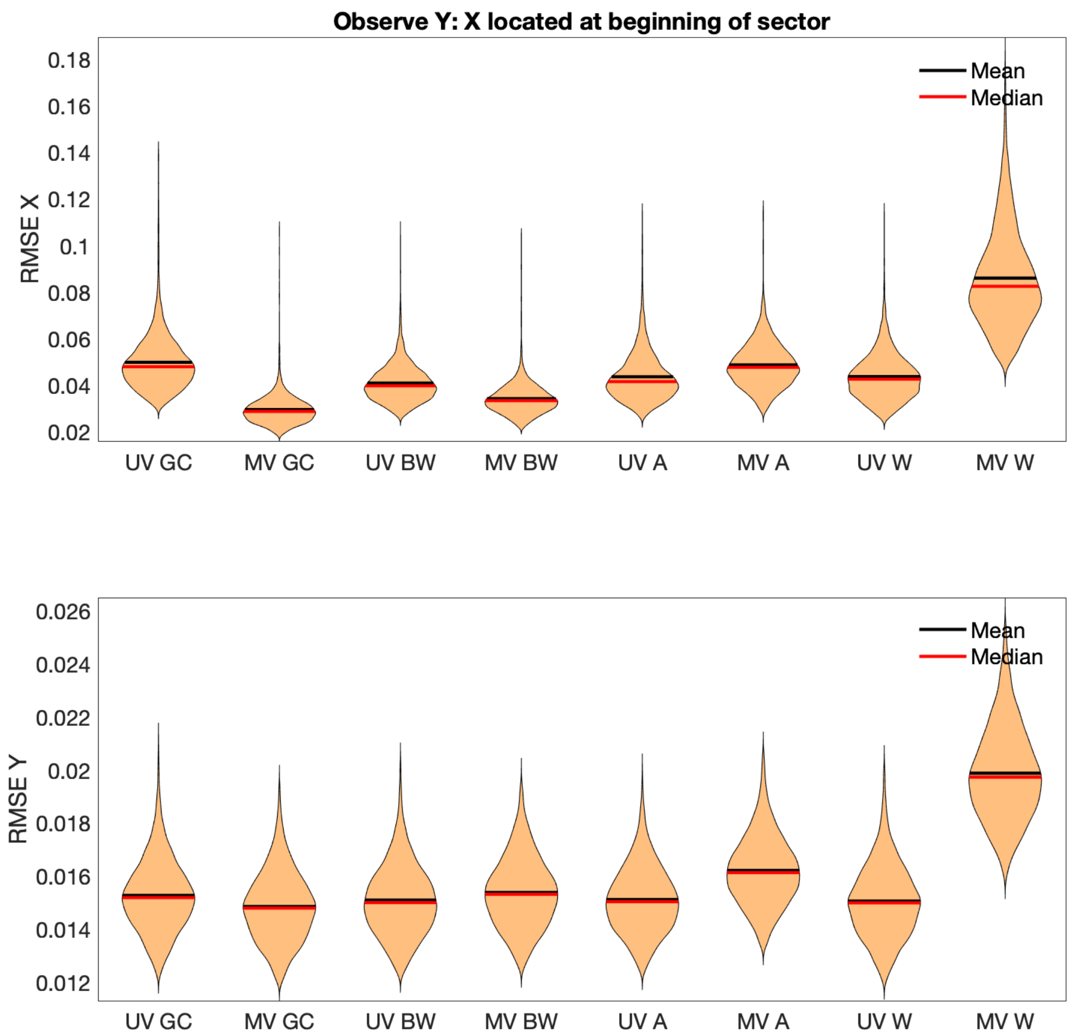

Figure 11: Move the location of X_k to the beginning of the sector, instead of the middle of the sector.

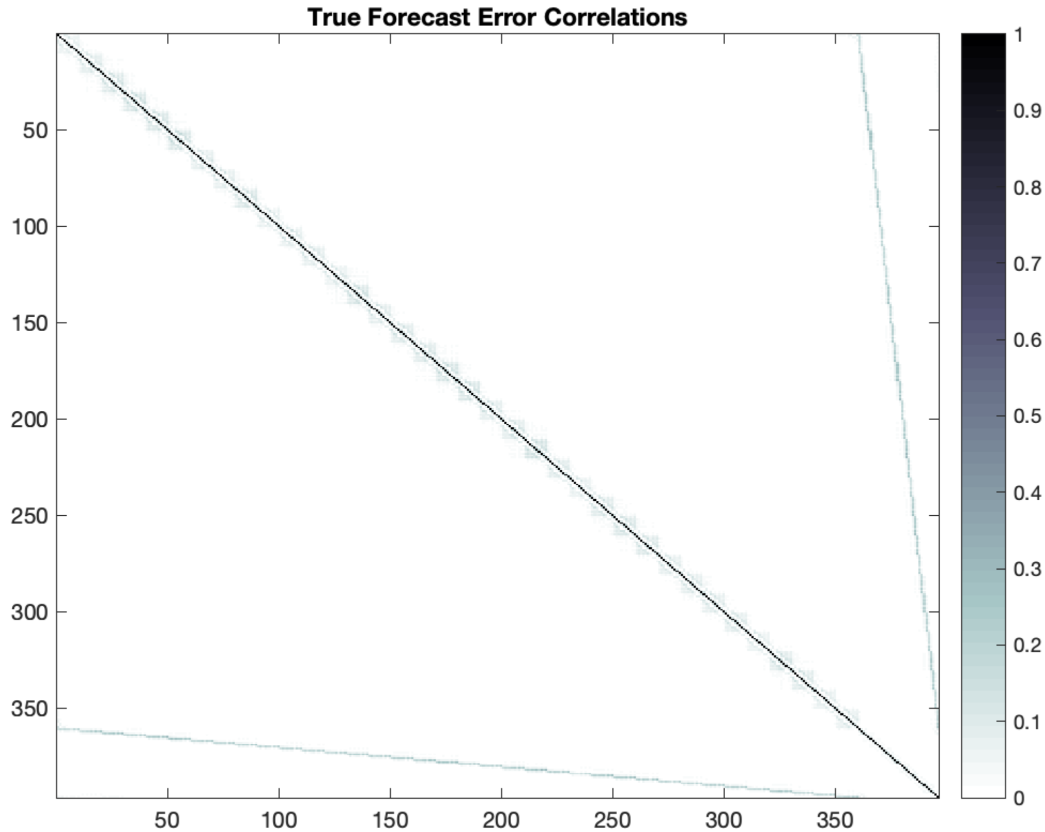

Figure 12: True forecast error correlation matrix. Upper left block shows correlations between the "short" Y variables. Lower right blocks shows correlations between the "long" X variables.

20. L352: how is κ defined for this case?

    κ is defined as $\kappa^2 = \frac{\max\{R_{ii}, R_{jj}\}}{\min\{R_{ii}, R_{jj}\}}$. We added this clarification to the manuscript.

21. Fig B1: Have you plotted the whole correlation matrix? Or even the individual blocks? It may be easier to see the correlation structures, particularly for the cross-correlations

    We have plotted the whole correlation matrix, see Fig. 12. We find it difficult to estimate the correlation length scales from the full matrix because the "short" and "long" variables are on different grids and these grids do not show up clearly in the matrix. For example, looking just at this matrix one might reasonably conclude that the correlation length scale is longer for the Y process (upper left block) than the X process (lower right block), when in fact the opposite is true.

22. L390: "$\gamma_{XX}, \gamma_{YY}, \gamma_{XY}$, and ν" I don't understand what these parameters represent.

    See response to #6. The parameters ν and $\{\gamma_{ij}\}$ are related to the shape of the localization functions, and are necessary to guarantee positive definiteness in a given dimension. A sentence has been added to section 2.5 describing these parameters.

23. Fig B3: Are these showing the RMSE for X?

    Yes, however the section describing the estimation of the localization parameters has been rewritten and this figure has been removed. We felt that this figure was unclear and causing confusion.