Line (initial)	Comment	Reviewer	Response (revision or comment)	Revised Line
Section 2.6	For the error propagation described in section 2.6, what is the reasoning for using CO_2 flux? Using CO_2 flux introduces several other potential biases and errors to the assessment:	1	The initial reasoning for inlcuding the CO_2 flux was to show implications of error propogation for various imputations during a common use case for DIC time series. However, both reviewers have raised similar concern about the introduction of multiple sources of error when determining CO_2 flux. The combined uncertainty for the CO_2 flux was initially determined by a Monte Carlo method (n=1000, which was not significantly different than n=10000) and then only the values of imputed DIC and their associeted uncertainty were varied as inputs into the calculation. In this way we attributed the percent difference between imputed time series and observed time series to be related only to the gap-filling method because no other input was varied. Similarly the uncertnainty of the CO_2 flux was determined via this MCM for each method. That said, we understand the concern about multiple sources of error and recognize that this application detracts from the results of just gap-filling the DIC time series. We will remove the CO_2 flux aspect of paper and add a focus on long term trend assessment in its place. This will be more consistent with the intentions of the work, enhance the focus of the paper and address mulitple comments from both reviewers.	Section 2.7 inluces the updated uncertainty budget
	· uncertainty in air pCO_2,	1	removing CO_2 flux aspect	
	· major bias and errors of NCEP winds (see: https://doi.org/10.5194/bg-15-1701-2018, https://doi.org/10.1029/2018GB006047, https://doi.org/10.1002/2017GL073814)	1	removing CO_2 flux aspect	NA
	· uncertainty in the gas transfer velocity coefficient (resulting in total uncertainty in CO_2 flux of ~20%), and	1	removing CO_2 flux aspect	NA
	· uncertainty (~5%) introduced in the calculation of sw pCO_2 from DIC and TA.	1	removing CO_2 flux aspect	NA
	How will those biases and errors complicate your assessment of gap filling error propagation? The relative uncertainty for CO_2 flux at BATS is reported in line 361 as 3.5%. What does this uncertainty take into account? Not items 1 – 4 above, as this value would be much higher. These issues should be addressed in the error propagation, or another parameter should be used for this assessment.	1	removing CO_2 flux aspect	NA
NA	Data used in this study need to be cited properly, which is incredibly important to the programs supporting these time series measurements. Those data should be cited in the methods and/or funders noted in the acknowledgements, depending on what each time series program recommends, not recorded as web addresses in the notes of Table 2. For the moorings, if you are accessing original data files via NCEI, those citations can be found at https://doi.org/10.3334/cdiac/otg.tsm_papa_145w_50n for Papa and https://doi.org/10.3334/cdiac/otg.tsm_keo_145e_32n for KEO. If you are accessing the mooring data from the synthesis product, the citation can be found at https://doi.org/10.7289/V5DB8043. I am not as familiar with the citation requirements of all the ship-based time series, but with a quick search I found this data citation request for HOTS, for example: https://hahana.soest.hawaii.edu/hot/dataaccess.html	1	This was a gross oversight on our part. While the sources for data sets were listed in table 1, they were not properly cited as noted. We will cite these as required.	See Section 2.1 Field Data lines 85-95 and Acknowledgements
NA	Finally, it may be out of the scope to include additional analyses in this paper, but it would be worthwhile discussing future work that can build off these results. For example, what satellite-based products are best suited for the MLR approach? Are there any that can span open ocean and coastal environments, so gap filling methods can be applied consistently across all global ocean and coastal time series? Also, it would be useful to study whether there are discrepancies in calculated trends when using these different gap filling methods (at least the most successful methods) or no gap filling methods at all. Both of these analyses seem like they could have been included in this paper, but I could also understand if those are the next assessments planned using the most promising empirical gap filling methods resulting from this work.	1	These are excellent points some of which we can address in the revision. Firstly, we will include an assessment of impacts on trends in place of the CO_2 flux. Secondly, we have already separately performed a cross shelf assessment of the MLR performance using data from the Munida transect and we can included this appliaction. These aspect taken together will also help address Reviewer 2's comment about focusing the scope of the paper more on presenting this MLR method and comparing it to other gap-filling methods, rather than an extensive comparitive assessment of techniques since we have only selected a few methods from a very large number of possiblities.	spatial extrapolation is not included in the current scope but future developoments are discussed through section 4.2 and 5
Line (initial)	Comment	Reviewer	Response (revision or comment)	Revised Line
Section 2.7	For the error propagation described in section 2.6, what is the reasoning for using CO_2 flux? Using CO_2 flux introduces several other potential biases and errors to the assessment:	1	The initial reasoning for inlcuding the CO_2 flux was to show implications of error propogation for various imputations during a common use case for DIC time series. However, both reviewers have raised similar concern about the introduction of multiple sources of error when determining CO_2 flux. The combined uncertainty for the CO_2 flux was initially determined by a Monte Carlo method (n=1000, which was not significantly different than n=10000) and then only the values of imputed DIC and their associeted uncertainty were varied as inputs into the calculation. In this way we attributed the percent difference between imputed time series and observed time series to be related only to the gap-filling method because no other input was varied. Similarly the uncertnainty of the CO_2 flux was determined via this MCM for each method. That said, we understand the concern about multiple sources of error and recognize that this application detracts from the results of just gap-filling the DIC time series. We will remove the CO_2 flux aspect of paper and add a focus on long term trend assessment in its place. This will be more consistent with the intentions of the work, enhance the focus of the paper and address mulitple comments from both reviewers.	Section 2.7 inluces the updated uncertainty budget
	· uncertainty in air pCO_2,	1	removing CO_2 flux aspect	
	· major bias and errors of NCEP winds (see: https://doi.org/10.5194/bg-15-1701-2018, https://doi.org/10.1029/2018GB006047, https://doi.org/10.1002/2017GL073814)	1	removing CO_2 flux aspect	NA
	· uncertainty in the gas transfer velocity coefficient (resulting in total uncertainty in CO_2 flux of ~20%), and	1	removing CO_2 flux aspect	NA
	· uncertainty (~5%) introduced in the calculation of sw pCO_2 from DIC and TA.	1	removing CO_2 flux aspect	NA
	How will those biases and errors complicate your assessment of gap filling error propagation? The relative uncertainty for CO_2 flux at BATS is reported in line 361 as 3.5%. What does this uncertainty take into account? Not items 1 – 4 above, as this value would be much higher. These issues should be addressed in the error propagation, or another parameter should be used for this assessment.	1	removing CO_2 flux aspect	NA

NA	Data used in this study need to be cited properly, which is incredibly important to the programs supporting these time series measurements. Those data should be cited in the methods and/or funders noted in the acknowledgements, depending on what each time series program recommends, not recorded as web addresses in the notes of Table 2. For the moorings, if you are accessing original data files via NCEI, those citations can be found at https://doi.org/10.3334/cdiac/otg.tsm_papa_145w_50n for Papa and https://doi.org/10.3334/cdiac/otg.tsm_keo_145e_32n for KEO. If you are accessing the mooring data from the synthesis product, the citation can be found at https://doi.org/10.7289/V5DB8043. I am not as familiar with the citation requirements of all the ship-based time series, but with a quick search I found this data citation request for HOTS, for example: https://hahana.soest.hawaii.edu/hot/dataaccess.html	1	This was a gross oversight on our part. While the sources for data sets were listed in table 1, they were not properly cited as noted. We will cite these as required.	See Section 2.1 Field Data lines 85-95 and Acknowledgements
NA	Finally, it may be out of the scope to include additional analyses in this paper, but it would be worthwhile discussing future work that can build off these results. For example, what satellite-based products are best suited for the MLR approach? Are there any that can span open ocean and coastal environments, so gap filling methods can be applied consistently across all global ocean and coastal time series? Also, it would be useful to study whether there are discrepancies in calculated trends when using these different gap filling methods (at least the most successful methods) or no gap filling methods at all. Both of these analyses seem like they could have been included in this paper, but I could also understand if those are the next assessments planned using the most promising empirical gap filling methods resulting from this work.	1	These are excellent points some of which we can address in the revision. Firstly, we will include an assessment of impacts on trends in place of the CO_2 flux. Secondly, we have already separately performed a cross shelf assessment of the MLR performance using data from the Munida transect and we can included this appliaction. These aspect taken together will also help address Reviewer 2's comment about focusing the scope of the paper more on presenting this MLR method and comparing it to other gap-filling methods, rather than an extensive comparitive assessment of techniques since we have only selected a few methods from a very large number of possiblities.	spatial extrapolation is not included in the current scope but future developoments are discussed through section 4.2 and 6

Line (initial)	Comment	Reviewer	Response (revision or comment)	Revised Line
Section 2.8	For the error propagation described in section 2.6, what is the reasoning for using CO_2 flux? Using CO_2 flux introduces several other potential biases and errors to the assessment:	1	The initial reasoning for inlcuding the CO_2 flux was to show implications of error propogation for various imputations during a common use case for DIC time series. However, both reviewers have raised similar concern about the introduction of multiple sources of error when determining CO_2 flux. The combined uncertainty for the CO_2 flux was initially determined by a Monte Carlo method (n=1000, which was not significantly different than n=10000) and then only the values of imputed DIC and their associeted uncertainty were varied as inputs into the calculation. In this way we attributed the percent difference between imputed time series and observed time series to be related only to the gap-filling method because no other input was varied. Similarly the uncertnainty of the CO_2 flux was determined via this MCM for each method. That said, we understand the concern about multiple sources of error and recognize that this application detracts from the results of just gap-filling the DIC time series. We will remove the CO_2 flux aspect of paper and add a focus on long term trend assessment in its place. This will be more consistent with the intentions of the work, enhance the focus of the paper and address mulitple comments from both reviewers.	Section 2.7 inluces the updated uncertainty budget
	· uncertainty in air pCO_2,	1	removing CO_2 flux aspect	
	· major bias and errors of NCEP winds (see: https://doi.org/10.5194/bg-15-1701-2018, https://doi.org/10.1029/2018GB006047, https://doi.org/10.1002/2017GL073814)	1	removing CO_2 flux aspect	NA
	· uncertainty in the gas transfer velocity coefficient (resulting in total uncertainty in CO_2 flux of ~20%), and	1	removing CO_2 flux aspect	NA
	· uncertainty (~5%) introduced in the calculation of sw pCO_2 from DIC and TA.	1	removing CO_2 flux aspect	NA
66	On Line 66 is stated "This study aims to identify the optimal gap-filling methods for carbonate time series by establishing which techniques perform with sufficiently low error and bias to assess seasonal and interannual variability of carbonate biogeochemistry and the biological and physical processes that determine it." The manuscript takes the approach that all gap-filling techniques have been explored and that MLR is recommended as the best performing. While the latter is certainly true of the methods compared, I feel it is not currently possible to say the former while one / a number of machine learning (and other) approaches are absent - these have recently been successfully applied in oceanographic research, and so the manuscript is not fulfilling its own aims by omitting them. Clearly it is not feasible to compare all available methodologies, so I would recommend that you either tone down the aims of the paper (by saying that you present a MLR method for DIC time-series data gap imputation and compare it to other common, computationally inexpensive methods) or a selection of additional methods are included e.g. median as well as mean, machine learning (i.e. neural network, regression trees, random forests that you already mention), curve fitting, exponential moving average, k-nearest neighbours etc.	2	This point is well-taken. Given that we have not here (and could not really) assessed all methods, we will shift the stated focus away from optimization of gap-filling and toward presenting the MLR and comparing it against other common approaches as suggested.	Lines 77-82
NA	When comparing methods a lot of focus is on the magnitude of the RMSE. I feel the reader would benefit from some consideration of the structure of the error e.g. are certain times of the year subject to greater uncertainties, do the models reproduce the timing of the seasonal cycle, and the magnitude of the peaks and troughs or are these far worse than those that vary around annual mean values? Equally, is the error of the preferred MLR technique actually normally distributed, as a lot of its power rests on this assumption. The manuscript would certainly benefit from greater examination of the seasonal cycle, and anomalies from this in the imputation methods.	2	This point is also well taken. With removing the CO_2 flux aspect of the paper we can provide more room for showing the distribution of error. As for the structure of the seasonal cycle, we disucss this but had not quantified it. In revision we will provide quanitification of the timing and magnitude of the seasonal cycle and some metric(s) for method performance to make this discussion less qualitative.	This is address through the rivsed results and discussion sections
NA	The use of the air-sea CO_2 flux for assessing imputation performance is an interesting choice, as it introduces a whole suite of additional uncertainties (wind-speed, piston velocity, K1/K2 equilibrium constants, how missing alkalinity data is filled etc) that are not considered in your error analysis. These uncertainties would also need to be assessed, or another metric/s chosen for comparison. If the air-sea CO_2 flux is still the preferred metric, is it not better to calculate pCO_2 from DIC/alkalinity first, before imputing missing pCO_2 values?	2	See our response to reviewer 1 comments above regarding our initial methods and reasoning; and note that we will be removing this aspect from the paper.	NA

NA	I appreciate that this may be being considered in a follow up study, but an assessment of the desired sampling frequency necessary to generate a good representation of the seasonal cycle (1, 1.5, 2, 3 month frequency, only summer and winter etc) would be very interesting/useful.	2	We will add this assessment	Updated by inclusion of various gap filters and address through revised results and discussion secitons
36	value is singular, so has not have	2	This has been addressed	Line 46
38	40% - This is possibly fossil fuel CO2 emissions? All anthropogenic CO2 (including land-use change and cement) means the ocean component is probably closer to 25% (Global Carbon Project, Friedlingstein et al., 2020)	2	This has been addressed	Line 49
66	"This study aims to identify the optimal gap-filling methods for carbonate time series by establishing which techniques perform with sufficiently low error and bias to assess seasonal and interannual variability of carbonate biogeochemistry and the biological and physical processes that determine it." - see comment above	2	Response as above	Lines 77-82
72	should be principle rather than principal	2	This has been addressed	Line 82
75	(and Table 1) - add citation/references for time-series, possibly through additional column in Table	2	As per response to Adrinne's comment above, this was an oversight and all dataset citations will be properly added.	Addresed in section 2.1 and Acknowledgements
86	Is there an impact on your analyses of averaging data to monthly means?	2	Uncertainty in monthly values was estimated for both single observations and averaged higher frequency measurements from moorings so they could be properly compared. We will make sure this is clearly communicated in the methdos during revision	Included in uncertainty budget now - see section 2.7
89	would be better to use greek delta notation rather than DT	2	fixed per above as well	line 99
90	What is the uncertainty introduced by the use of estimated DIC values? DIC is only measured at BATS. What do you get if you apply the same techniques to data with DIC, TA and pCO2 e.g. at sea surface?	2	Individual DIC uncertainty budgets were assess by adding the sources (measurement, natural variability (e.g. monthly averaging), and /or propogation from calculating DIC from other carbonate measurements) in quadrature to determine the combined statndard uncertainty for each DIC value in the time series. For DIC calculated from the other variabiles such as pCO2 and TA, the error function in the R package seacarb was used.	See Section 2.7 for updated uncertainty
122	"The primary goal was imputing timeseries at monthly resolution to investigate variability and trends over seasonal, interannual and decadal timescales" - neither trends nor decadal are covered as far as I can see?	2	See our responses above that indicate we will be removing the CO2 flux aspect and adding an assessment of trends and seasonal structure	trends and other seasonal and interannual variability were quantified in the results and discussed in section 4.2
141	is this not an exponential moving average then, rather than a weighted moving average?	2	I suppose it could be stated both ways. It is a weighted moving average, but the weighting is based on an exponential relation to neighbors	It is more cldearly referred to as exponential wma
148	cite1 and cite2?	2	This was some sort of formating typo with Endnote - will fix	Line 173
150	does this method also input uncertainty into the fitted values used?	2	I don't believe this inputs uncertainty - rather values are found through convergence of multiple regressions. Unertainty can be assessed by looking at the spread when the option to have multple outputs for a give value is selected.	NA
190	as above, why this? Is it not better to calculate pCO2 from bottles at the start, then do imputation on pCO2 data set?	2	No imputation of pCO2 data was done. All imputation is on DIC values only. All pCO2 was calucated from the imputed DIC and either measured or estiamted alkalinity	NA
193	Wanninkhof 2014 recommends to not use Wanninkhof 1992.	2	removing CO2 flux aspect	NA
201	why not use Bermuda atmospheric CO2 concentrations?	2	removing CO2 flux aspect	NA
215	what were these uncertainties? It would be good to state them here. pCO2 from DIC and TA at their measurement uncertainty is ~6uatm. What is it when DIC is estimated?	2	We will make uncerainties more explicity during revision	See Section 2.7 for updated uncertainty
223	To give a better feeling of interannual variability it would be useful to have the value for n for each month in Figure 2. For example so that a reader doesn't look at FOT and think there is very little variability in months 1-3, when instead n is only 1-2 for these months.	2	will add this info	Updated Fig. 2
227	& Fig 3. Is this a single MLR encompassing all data from all sites? Or the results of individual MLRs plotted and pooled? I'm don't think this is clear in the text	2	This is pooled results. MLRs must be fit using site specific observations and have unique coefficients. Will update language to clarify	Updated to Fig caption (Now Fig 4)
229	"worked well"? A RMSE of 12 is beyond the 'weather' goal of measurement quality to assess spatial and short-term variability. I'm not sure stating this metric is useful as it obscures the capability of the method in (primarily) oceanic sites. Instead it might be better to simply focus on individual monitoring station results.	2	This is a good point and we will address during revision	updated these points in results and in discussion section 4.2
234	It would be interesting to hear the thoughts behind why PAPA performs so well	2	The MLR appears to perform best at sites where there is a high correlation to temperature and a large seasonal cycle. The performance of the results follows the trends shown in Figure 5 where there is selective ommission of predictor variables. We will investigate and elaborate on this further during revision	Line 290
244	put the numbers in the boxes as well - the colour scale is not the most obvious/immediate to show similarity/disparity	2	We will add this info	updated tile figures now Figs 6 & 8
245	add another line to the bottom of Figure 5 to show mean	2	We will add this info	updated figure 6
246	Table 5 - change title to Mean model results	2	This has been addressed	Line 932
250	Figure 6 - might be better showing as well / instead the residual (y) versus the measured (x)? - this may better highlight the better performing models, with the distribution of the residual ideally normal about 0.	2	We will explore this suggestion and other possiblities for expressing error distribution across both observed DIC ranges and sites	Updated Figure to show residuals as kernel density plots as this seems to provide the best representation for the point that was being asked here. Basic residual plots were visually messy and did not provide additional clarity.
259	I struggle somewhat with this plot (Fig 7) too. The colour scale is not the most obvious/immediate to show similarity/disparity, and seems to be the opposite to Figure 5 where light colours indicate better performance - here they indicate worse performance.	2	We will look to increase the contrast and make these figure gradients consistent for clarity	updated tile figures for consistency
261	I think that showing the performance of the models in recreating the seasonal cycle would be very useful. Whether they get the amplitude and timing correct is important for potential end users of these methods. Showing the anomaly from the observed seasonal cycle may also be useful.	2	This is a great point. As indicated in response above, this was qualified in the discussion but was lacking quantification and we will add that during revision	Created new analyses and metrics to quantify this and included in results and discussion sections
266	Fig 8A I like this plot, but i think it is making false equivalences by using different y scales for the 7 different methods for each monitoring station. It might be worth having this as a standalone figure to give more space to what is an enormous amount of information.	2	With the removal of the CO2 flux aspect and associate figure we will have space to break out this figure. The y scales were held consistent across sites so that methods could be compared. If the scales are held constant for all sites and all methods it will loose significant detail for visual interpretation.	The y scales are consistent for sites. This figure was also updated to make it slightly cleaner for visibility
275	Assessing error on seasonality and annual sums - not sure these numbers capture this. As mentioned above I'd be interested in seeing the performance of individual methods of capturing the seasonal cycle / amplitude and annual mean, and how they compare to the data, both using the full timeseries, and when there are artifical data gaps. It would certainly be useful to know how critical it is to sample seasonal maxima/minima (or not) in correctly formulating a seasonal cycle, and getting lowering the uncertainty with respect to annual budgets.	2	The aspect of sampling optimization is a good point that is missing here. As noted in responses above, we will quantify the performance of retaining seasonal structure and we will explore assessing sampling optimization.	Created new analyses and metrics to quantify this and included in results and discussion sections

280	and Figure 9A. While these plots are interesting it might be better represented by adding/replacing wih anomaly timeseries. Also, I was wondering whether you could comment on how there appears to be a positive bias for the bimonthly and 3 month data gaps towards higher concentrations? Is the reason there are no red dots at the lowest concentrations (particularly in the 3 month timescale) simply the result of random data gaps, or something else? For the 6 month gaps I'd be interested in the performance of the models when only summer data is available, or perhaps completely missing winter data, as this would be a situtaion facing other time series sites.	2	We can explore representing anomalies here for clarity. As for the gap placement in the 3-month gap series, yes this is just due to randomization. We could explore artificially removing particular seasons and assessing impacts on annual cycles.	NA
291	Fig 9b - would it be possible to have the legend across a single row, to aid in identifying models? Or indeed numbering the different box plots.	2	We will address clarifying the identification of methods in this boxplot	NA
299	Figure 10 - this plot mght be easier to interpret if it was anomalies from observations rather than actual values side-by-side?	2	This figure will be removed along with the CO2 flux aspect of the paper	NA
	The uncertainty bars also seem particularly low - has the uncertainty from the imputed data been propagated through the calculation? Even a DIC RMSE of 6 umol/kg would have an impact of 10-25uatm of pCO2 depending on temperature. I imagine if there are missing DIC observations, there will also be missing alkalinity observations as well. It will likely be too much to include an estimate from these values as well, but I think you should comment on the fact that the error estimates relating to air-sea CO2 fluxes presented here will be an underestimate, as there will also be addiional uncertainties associated with imputing alkalinity.	2	The uncertainty budget was assessed using a MCM method as noted above in response to other comments, however we will be removing this aspect regardless in place of more focus on assessing trends and seasonal strcuture	NA
328	change 'has a dominant effect the carbonate chemistry' to 'has a dominant effect on carbonate chemistry'	2	This has been addressed	Line 456
333	need to referencce these different datasets	2	This will be address as noted above	See Section 2.1 Field Data lines 85-95 and Acknowledgements
335	missing full stop	2	This has been addressed	Line 464
353	- I don't think you've shown anything about temporal extrapolation.	2	This has been addressed	Line 484
358	either remove the parentheses around the citations, or remove 'in the studies of'	2	This has been addressed - this was an Endnote formatting typo	Line 498
369	This may be so but I don't think the figures you have presented make this obvious. A figure showing the mean seasonal cyle from the full data set compared to those imputed for different percentages of missing data would be necessary to show this.	2	Our quantification of seasonal structure during revision will address this	See Fig 14 with results and discussion revisions
371	it's not clear visually, as you're missing a figure showing it. Figure 9 suggests it's only really obvious for the 6 month gap, while Figure 12 suggests that the mean approach has some of the highest uncertainties for the bi-monthly data gaps.	2	Our quantification of seasonal structure during revision will address this	See Fig 14 with results and discussion revisions
381	I'd again suggest that looking at anomaly plots would be more straightforward to interpret than net flux comparisons	2	Point well taken, and we will explore this for clarity	NA
405	change 'In general' to 'Of the methods we tested'	2	This has been addressed	Line 573
408	May and possibly are really not strong enough - the artifice of the mean imputation method introduces bias, and actively removes any trend from the input data.	2	Good point, we will revise language here	Line 575
415	- MLR certainly has the lowest error, but this doesn't necessarily tell the whole story. Showing the residuals of the predicted values will help - would you like to comment on the tendency of MLR methods to revert to the mean, where higher values are typically predicted lower, and lower values are predicted higher. This will have an impact on estimating maxima/minima. And I'd hesitate to recommend best practice until MLR is compared against a fuller suite of gap-filling methods, including machine learning	2	As also noted above in responses to a similar comment, we will revise the focus of the paper to dial back the language for establishing best practices and shift to scoping it as a presentation of this MLR compared to some selected common methods. The expanded seasonal structure assessment will help the discussion about max/min biasing	Lines 585-591
426	(and L433)- can be estimated, but to what uncertainty, and is this the same across all times of the year?	2	We will assess and present seasonal error distribution to futher support this; however uncertainty must be assess on an individual site/data set basis. We will make revisions to the language here to make sure that point is clear	Lines 604 - 612
432	I sound like a broken record but I think plots of seaonal cycles/anoamlies of seasonal cycles/internannual anaomlies are really what are needed to help determine this.	2	Noted and will address	See Fig 14 with results and discussion revisions
433	Change "the most robust option for imputing gaps over a variety of data gap scenarios." to "the most robust option from those we compared for imputing gaps over a variety of data gap scenarios."		This has been addressed	Lines 604 - 612