Biogeosciences Discussions

Biogeosciences Discuss., referee comment RC2
https://doi.org/10.5194/bg-2021-78-RC2, 2021

Comment on bg-2021-78

Anonymous Referee #2

Referee comment on "An empirical MLR for estimating surface layer DIC and a comparative assessment to other gap-filling techniques for ocean carbon time series" by Jesse M. Vance et al., Biogeosciences Discuss., https://doi.org/10.5194/bg-2021-78-RC2, 2021

The authors present a novel analysis of a number of different methods for the imputation of missing marine carbon data from long-term monitoring stations. These sites are vital for understanding the variability of biogeochemical parameters over a multitude of timescales, but also for the tracking of trends in essential variables that are climatically important / critical to ecosystem health. I'm not aware of a previous study that has looked into these methods specifically for imputing marine carbon timeseries data gaps so this is a necessary and timely piece of work. Their findings that on the most part empirical methods are found to perform better than statistical methods for data imputation will be of interest to many, and hopefully lead to improvements in timeseries analyses within the community. I do have a number of specific comments that will need addressing before publication however:

1 - 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 - 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.

3 - 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?

4 - 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.

Technical comments:

L36 value is singular, so has not have

L38 40% - This is possibly fossil fuel CO_2 emissions? All anthropogenic CO_2 (including land-use change and cement) means the ocean component is probably closer to 25% (Global Carbon Project, Friedlingstein et al., 2020)

L66+ "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

L72 should be principle rather than principal

L75 (and Table 1) - add citation/references for time-series, possibly through additional column in Table

L86 Is there an impact on your analyses of averaging data to monthly means?

L89 would be better to use greek delta notation rather than DT

L90 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?

L122 "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?

L141 is this not an exponential moving average then, rather than a weighted moving average?

L148 cite1 and cite2?

L~150 does this method also input uncertainty into the fitted values used?

L190 as above, why this? Is it not better to calculate pCO2 from bottles at the start, then do imputation on pCO2 data set?

L193 Wanninkhof 2014 recommends to not use Wanninkhof 1992.

L201 why not use Bermuda atmospheric CO2 concentrations?

L215 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?

L223 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.

L227 & 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

L229 "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.

L234 It would be interesting to hear the thoughts behind why PAPA performs so well

L244 put the numbers in the boxes as well - the colour scale is not the most obvious/immediate to show similarity/disparity

L245 - add another line to the bottom of Figure 5 to show mean

L246 Table 5 - change title to Mean model results

L250+ - 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.

L259 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.

L261 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.

L266 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.

L275+ 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.

L280+ 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.

L291 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.

L299 Figure 10 - this plot mght be easier to interpret if it was anomalies from observations rather than actual values side-by-side?
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.

L328 change 'has a dominant effect the carbonate chemistry' to 'has a dominant effect on carbonate chemistry'

L333 need to referencce these different datasets

L335 missing full stop

L353 - I don't think you've shown anything about temporal extrapolation.

L358 either remove the parentheses around the citations, or remove 'in the studies of'

L369 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.

L371-2, 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.

L381 - I'd again suggest that looking at anomaly plots would be more straightforward to interpret than net flux comparisons

L405 - change 'In general' to 'Of the methods we tested'

L408 - 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.

L415 - 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

L426 (and L433)- can be estimated, but to what uncertainty, and is this the same across all times of the year?

L432 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.

L433 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."