Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2019-46-RC3, 2019
© Author(s) 2019. This work is distributed under
the Creative Commons Attribution 4.0 License.

Geoscientific Model Development Discussions

Interactive comment on "A comparative assessment of the uncertainties of global surface-ocean CO_2 estimates using a machine learning ensemble (CSIR-ML6 version 2019a) – have we hit the wall?" *by* Luke Gregor et al.

Anonymous Referee #3

Received and published: 28 May 2019

The authors present (1) a new algorithm to spatiotemporally interpolate discrete pCO2 measurements into continuous pCO2 field, and (2) present and discuss a comparison between this and existing pCO2 interpolations in the light of several metrics. Concerning (1), though the algorithm is based on the same principles (namely non-linear regression of pCO2 against driving quantities measured more completely) which have also been employed by several existing algorithms for the same purpose, it differs by a formalized selection of how to split the ocean into areas of similar biogeochemical behaviour. In particular, the selection is not done in isolation but involves the regression

step itself. This split into coherent areas is an important element in regression-based pCO2 interpolations. It is therefore an interesting contribution worth to be published. The comparison (2) is a useful part of the evaluation. In the discussion, however, the authors somewhat delve in general statements that have already been discussed in the literature or have been tackled in ongoing research projects (see below). I also think that some statements may be put into perspective (see below). The paper is well written, though at a number of places the text may be revised to become more accessible (see some suggestions below). In summary, I'd like to recommend to publish this study in GMD, after revisions to address the points detailed in the following.

On terminology, there is a problem with the authors' use of "ensemble". Usually, an "ensemble" means a set of several members. At various places (first in L124), however, the word is apparently used for "ensemble mean" (= just one entity, not a set any more). This sometimes distorts the meaning and confused me substantially on first reading. Similarly confusing is the use of "clusters" not only for "points belonging together" but also for "cases having different (or differently many) clusters".

As an interesting feature in Fig 7(a), I notice adjacent bands of strong opposite biases in the eastern Pacific. Could this point to an inappropriately located boundary between the regions? It may help to check if these bands also occur for K21E and BIO23 individually, do they? If so, is there a systematic difference in the location of the region boundary between K21E and BIO23? I imagine that such analysis might give hints on how to improve the interpolation.

The paper also discusses more general aspects of pCO2 interpolation, such as the potential "wall" mentioned in the title, which is definitely an interesting and relevant question. However, I'm a bit surprised by some formulations, such as L677-678 or L578 ("stagnant"), which seem to suggest that "there must be intrinsic limits if not even our method performs better than other methods". Why should we expect your particular method to exhaust all what's achievable? After all, the presented method is based, like several previous pCO2 gap-filling studies, on instantaneous relationships to physical or

biological oceanic quantities. While such relationships have proven to capture a good fraction of pCO2 variability, it is clear that oceanic biogeochemistry is a dynamical system, ie., pCO2 depends not only on the current state but on the past history. Though the need of "new methods" is being mentioned (L629-L638), the discussion remains solely with regressions. For example, it ignores that other approaches like "data assimilation" into process models do exist already (though mostly not yet in a stage to fit the data closely). On the other hand, the discussion likewise ignores that sophisticated methods like regressions against drivers or data assimilation are only needed because large data gaps need to be bridged. In a data-rich world, such as pleaded for by the authors, simpler auto-regressive methods are also sufficient, as indicated e.g. in your Fig 10 by the relatively good agreement of driver-regressions and auto-regressions in the more data-covered areas. In order to make the discussion more interesting in the revised paper, therefore, I feel that it should be done in a wider context of the existing literature and make more concrete statements on how to go forward. An alternative option may be to substantially shorten the discussion and keep ideas for a future research paper (why not in the context of a new SOCOM as the authors propose?).

The same also applies to the discussion of sampling strategies. The dependence of accuracy on data density, the need for denser sampling in many parts of the southern hemisphere, or the use of synthetic data to test sampling strategies, are all not new. While autonomous sampling devices are presented as "a new way", there are papers already, e.g., from the SOCCOM project, which are not even cited in the discussion. These papers already discuss possibilities and limits on a higher level. In my opinion, a discussion of sampling also needs the input by the experimentalists (e.g., I'm not sure how "low-cost" the autonomous platforms really are). I feel the paper would win from a shorter and more concise discussion.

The authors find that the average over their ensemble of regressions perfomed better under several metrics than the individual ensemble members. They present this as a contradiction to warnings against the use of ensemble averages in the literature.

When comparing the presented ensemble of pCO2 interpolations with e.g. the SOCOM ensemble, however, there seem to be distinct differences in how statistically homogeneous these ensembles are: The presented ensemble of regressions against the same explanatory variables likely spreads in the finer details (see the rather similar behaviour in Fig 6) such that averaging may reduce noise, while there are large systematic differences (including members with limited ability to fit the data) in the SOCOM ensemble or other ensembles from the literature. Therefore, I feel the authors should discuss better the conditions under which the average over an ensemble really is a meaningful estimate.

Minor comments:

L47: maybe add "e.g."

L48: "maximise" does not seem to be the right word here

L57: "consolidated" probably means "collated" or similar

L65: as far as I see, there is not actually any weighting in that paper

L67: from my knowledge of the literature, most studies analysing existing pCO2 interpolations actually use several of these, rather than one "most widely used method"

L106: maybe add "separately" after "applied"

L107: "K-means clustering" should be briefly explained (either here or later, e.g. around L232)

L115: "described by" may better be "denoted as" (same in L120)

L117: you probably mean "a range of 11 to 25 clusters"

L123: spurious "and"

Fig 1, section "DATA": Table "XX"

Fig 1 Sect 4 (and many other places in text, tables, and figures): "HOTS" should be

"HOT"

L140: maybe add "a" before "predictive"

L155: Explain or spell out "GHRSST"

L172: Is the use of "random noise" a standard technique to fill incomplete input data? Isn't there a chance that this creates instability to the regression? A brief explanation or a reference would be useful here

L188: Maybe add "separate" or "individual" before "regressions"

L198: "palate" is probably misspelled, what about "shown by separate colors"

L214: Reference to Fig 5 would break the order of figures, but could easily be removed here

L216: not sure "a.k.a." is a suitable abbreviation

L217: "80:20" seems to contradict "75:25" in Fig 1

L224-235: I found this paragraph difficult to understand. Can you say more exlicitly which "hyper-parameters" you mean? It may also help to better link this paragraph and the previous one.

L244: add "below" after "Eqn 3 and 4"

L247-257: I did not find this paragraph very clear. Does it mean that you repeat the previous steps with other selections of years in the test-training split?

L300: The Rˆ IAV metric was first mentioned in the SOCOM paper (ie. Rödenbeck et al., 2015)

L303-304: According to Rödenbeck et al. (2015), their benchmark has no interannual variability, but it does have seasonal variability.

L306-307: Also here, the metric used by the authors (or at least the description of it) is

not the same as that presented in Rödenbeck et al. (2015): SOCOM used the standard deviation over yearly averaged pCO2 mismatches, not standard deviations over the full data in individual years. If the metric used in this study has indeed been calculated in the way it is described, it should be sensitive to within-year variations (probably dominated by the match to the seasonal cycle) but be insensitive to interannual variations.

L309-310: A benchmark constructed this way still contains the interannual variations of pCO2 (as the atmospheric pCO2 has very little IAV compared to seawater pCO2, except for the rising trend). Also here, this is opposite to what has been described by Rödenbeck et al. (2015) which removed IAV from the benchmark.

L313: I guess you mean "in contrast to the standard deviation which is sensitive to outliers."

L314: It is not clear to me what "interannual resampling" means, please be more explicit.

L319: What do you mean by "second part of the experiment"? Is it the "regression step of the algorithm"?

Fig 5: Why can e.g. D ever be worse than C, given that D has more degrees of freedom than C?

L326-327: If Fig 5 is showing the averages of the scores from the 4 methods, I was wondering whether these 4 methods show the same general behaviour (ie. whether the better/worse scores occur in similar rows/columns)? I'm asking because only if yes, Fig 5 would give a meaningful picture about which number of clusters and which features are best. A statement on that should be added.

L343: This has been said before and should be omitted here.

L344: Contradictory use of the term "ensemble", see remark above

L347: add "average" after "ensemble"

Tab 3 (also Tab4): I think the 1st column should better termed "clustering"

L355: I guess you mean "for each number of clusters"?

L366: I was wondering whether the occurrence of lower biases in the test years may actually be systematic (ie. not by chance). Are the same years (as listed in Sect 2.4) used in all the regressions? If so, couldn't it be that they implicitly lead to low biases through the model selection?

L380: You probably mean "sampling density"?

L397 and 399: The "||" around the unit is a rather sloppy notation. Better be explicit by writing "|bias| < 5uatm, RSME < 10uatm".

L410: Duplicate "bias"

L411: Add comma after "ERT", otherwise difficult to read.

L426: Are the statistics shown in the Taylor diagrams calculated over all individual data points? That is, do they reflect both spatial and temporal features?

Fig 8: Consider to use the same radial axis limits for all 6 Taylor diagrams. For example, the estimates seem to lie more apart for HOT, but that's only because the variability at HOT is smaller than in others of the independent data sets.

L436: Spurious "that"

L469: "2002" seems to contradict "2000" in L475 and 480.

L470: Shouldn't "regions" be "region groups"?

L479: It seems to me that "reflect" would better fit than "highlight"

L523-525: You could nicely link this back to Fig 2

L526: Add "trend" after "NH-ST", as it is mainly the trend which is reflected by IQRˆIA (if I understood correctly)

L544: replace "ensembles" by "ensemble averages" (see remark on terminology above)

L593, Fig 12: It remains unclear to me how to interpret the "seasonal cycle reproducibility". Doesn't it get smaller with stronger IAV? A short explanation would be helpful.

L611: duplicate "first"

L628: According to the paper, the resolution is daily, not 6-hourly.

L623: What do you mean by "procedural architectures"?

L633: The method by Denvil-Sommer et al. (2018) is named as an example of a "fundamentally new method". In fact, however, the "CARBONES-NN" contribution to the SOCOM ensemble also employed a climatological and an interannual step, but did not outperform other methods there. To clarify this interesting question, I'd suggest to include the Denvil-Sommer et al. (2018) results into your comparison Sect 3.3-3.5, as this would allow a clean comparison.

L645: If I understood correctly, the IQRˆIA metric is specifically sensitive to the trend. Why do you particularly propose this metric in the sampling context?

L695: Is "spatial coherence" really the right word here?

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2019-46, 2019.