Review ESSD-2018-64, Marine Environment Data Sets

Demonstrating minimal knowledge of ocean data, the authors have relied on obsolete sources and a proprietary (ESRI ArcGIS) software to 'produce' a series of unreliable data layers. Their product demonstrates neither quality nor reproducibility, the hallmarks of a good (ESSD) data set. Surprisingly, these authors seem substantially unaware of fundamental deficiencies.

Should we as readers hold these authors accountable for the very poor quality of their source data? Yes, in this case! The authors represent themselves (below) as competent data processors. They demonstrate their confidence by submitting their new compilation of previously un-reviewed data products for review and public comment in a prominent data journal.

These authors imply that based on "recent experience with developing compatible, comprehensive, environmental layers" they have "developed an extensive on-line repository of marine environmental data layers." Assembled, perhaps. Developed, no. Later, these authors admit that all data sets used for GMED "had undergone quality control checks by the primary data collectors and processors". These authors have applied a canned statistical interpolation (from ArcGIS), used some additional ArcGIS tools to check whether their interpolation introduced errors, and then served 60 layers of environmental … what? Reprocessed, improved and consistent data? Or, garbage amplified? Because these authors demonstrate minimal understanding of the quality, temporal extent or spatial coverage of their source data - as evidenced by the appalling array of old and obsolete sources - this reader regards their product as garbage, and old garbage at best. Apologies for strong language but one feels a need to 'wake up' these authors to a world of ocean data outside of their narrow view.

If in fact the ecology community accepts and relies on obsolete deficient ocean data sources then these authors may have done us an unintended favour by exposing the vast gap remaining between what oceanographers use and what ecologists use.

Data Access

The GMED url works, leads one to a useful landing page. The GMED version 2.0 link, at the upper right of that landing page, does not work. The 'DataSets' link does work and presumably leads to the layers discussed in the manuscript.

For this reader the figshare doi did not work in any configuration. ESSD generally does not use figshare for exactly these reliability issues. If the authors have a GMED snapshot they can deposit at figshare, they could easily - and more usefully - deposit it at Zenodo, Pangaea, 4TU, etc? Not reliable as presented, needs a change.

Page 3 starting line 59: I think you mean to say 'Differences in applications of SDM to marine environments compared to terrestrial environments include fewer observation records, extensive spatial-temporal variability of oceanic environments, and complexities in processing marine environmental data for SDM purposes'? But you would need to back up this statement with facts? From a carbon point of view, we might know less about land sources and sinks than we know about ocean sources and sinks? The SDM community takes a different view than the biogeochemical carbon cycle community? The authors seem blissfully ignorant of these larger issues, as perhaps they should, but to publish their data in the same journal that publishes (land and ocean) carbon cycle work they need to assure readers of an awareness of larger context?

Page 3 line 61: the phrase "spatial-thermal" seems strange here (I can find it used in other marine ecology papers) and I doubt the supposed differences marine to terrestrial. Spatial: ocean has surface to deep, land has sea level to alpine. Both land and ocean have strong latitudinal gradients and high variability in immediate coastal zones. If you truly mean thermal rather than temporal, ocean has -1 to 32C, land (even if we exclude ice) has -40 to +45), with ocean temperatures much more consistent than those of land. Ocean proper has very small range of salinity compared to large range of humidity on land. Ocean has no light to full light but so does land. Both have distinct daily and seasonal patterns that strongly impact biology. If you mean to imply range of habitats, we don't know much about 3-D patterns in the full-depth ocean but we

know that land includes a wide array of habitats. You will need to help the reader understand what you mean here because as written you have allowed a too-wide variety of misconceptions.

Page 3 lines 65 to 67: two biggest improvements in ocean monitoring probably come from ocean colour by satellite (e.g. MODIS and follow-ons) and globally-distributed profiling floats (e.g ARGO and, one hopes, soon bio-ARGO). (Some will also argue for the strong impact from GRACE.) This discussion of ocean observations seems to miss these crucial developments?

Page 3 line 70: many other groups have done a substantially better job at compiling and quality-controlling ocean data sets than Tyberghein and the bio-ORACLE crowd. To understand ocean properties from a biogeochemical or ecological perspective, a modern unbiased user would probably not start from bio-ORACLE? This entire discussion (lines 70 to 75) seems impossibly simple and surprisingly ignorant of other substantial efforts. Large literature exists on ocean SST. Substantial literature exists on ocean Chl A, much of it in ESSD, and extending quite far back before 1997. Have these authors never heard of GLODAP? Remote areas from an ocean data point of view might point to the central Pacific rather than (or as well as) the Arctic? This discussion pretty much convinced this reader that these authors lack working knowledge of current ocean data efforts.

Page 4 lines 78,79 “continuous, global, layers for such variables are predicted from ocean circulation models and by extrapolation of in situ sample data.” A very large and very active community of ocean observationalists and modelers would strongly disagree with this glib uninformed attempt at summary. These authors provide no citations to back their contentions?

Page 4 lines 80, 81: resolution of ocean models represents a hot and active topic, with variety of useful approaches driven by both physics (eddy-resolving, mixing parameterisations) and biology. The authors seem aware of none of that current vital work. Redfern 2006 represents a weak, almost irrelevant, reference.

Page 4 line 85: www.worldclim.org represents a weak, slow, out-of-date data source? Who uses it? The site provides no quality control nor uncertainties, they only repackage data from other unspecified sources. No climate modeller uses these products. Everything in one format but at what cost? Data only goes to year 2000? I wouldn’t touch it. GMED wants to emulate this?

Nice to see (lines 99,100) GBIF and OBIS mentioned but those do not include physical or chemical parameters. Typical of all GEO efforts, GEOBON provides nothing itself but only tries to ride on data sets produced and quality controlled by others.

Based on this introductory discussion a reader comes to doubt whether the authors have any knowledge of ocean observational (in situ, satellite or re-analysis) data sets.

Page 5 lines 115 to 117: basically, land-masking, then interpolation using standard ArcGIS tools, followed by comparisons with data sets already dismissed as deficient. Somehow this represents a useful contribution?

Figure 1 of terrible resolution, basically not even readable. Better figure in Supplement but why should reader need to look in Supplement to find a readable first figure of the paper?

Above, the authors dismissed AquaMAP and KGS Mapper as inadequate due to (apparently) complexity. Here they use the same two sources as their primary data sources? Later they then compare back to AquaMap and KGS Mapper. Can one imagine a more circular process, with less actual quality control contribution?

Page 5 line 128: Jungclaus 2006 - one run of one version of ECHAM5 at T63 downloaded from WDC, now more than 10 years out of date? Run at SRESA1B - an optimistic (and now also out-of-date) emission scenario? Do the authors somehow need to prove their ignorance of global ocean and climate data sets? Kaschner references merely a back-door route back to AquaMAPS.

Page 5 line 128: Kaschner et al 2013 url link does not work. And no wonder: 5 years since last access? Why insult the reader with useless out-of-date links?

In Table 1 authors claim to have used climate projections from CMIP3, one iteration of the UK Met Office model - with no emissions scenario information (as it turns out, already chosen by bioORACLE according to Table S1) - and one from IPSL, configuration unknown but evidently with a full depth green ocean and a coupled ice model, run at A2 (again from Table S1, already selected by and available from AquaMaps). Does reader follow the text (line 128) or the table? No guidance and too little information for anyone to attempt to replicate and confirm. Do the authors even know what they have used? Panels on the GMED landing page, for year 2100, credit UK Met / Hadley and IPSL, so the text at line 128 is wrong! Here we see Hadley run at A1B while IPSL runs at A2. Even if we accept those old CMIP3 scenarios as still valid, those two particular scenarios diverge widely at year 2100. Comparing apples with oranges? Do the authors even know what they present?

Page 5 line 129 and many following instances: for every processing step and every data challenge the authors turn to a standard ArcGIS tool. A reader never finds evaluation of alternatives or citations about how other researchers have confronted and solved these issues. ArcGIS represents an expensive proprietary tool, not suitable for ESSD. Graduate students around the world use open access GIS tools for the simple reason that they can't afford ESRI products. Reviewers and users who likewise prefer R or QGIS will have no ability to test, replicate or confirm procedures and tools used by these authors. Fails completely the open access repeatability expectations of ESSD. Also suggests that the authors lacked skill, interest or motivation to explore other tools. Reads like an ESRI advert.

Page 5 line 132: "average value of the 12 surrounding (ocean) cells." If done in 2-D (e.g. on the same depth or pressure surface), immediately adjacent cells would total 4 or perhaps 8. If done in three dimensions, 'adjacent' cells would total 6 or 26. Please can the authors explain and justify this gap filling? Did they simply chose a value number from ArcGIS raster calculator?

Page 5 line 134: the authors had an old CD of GEBCO data files so they used that? Several more recent, more accurate versions exist.

Page 6 lines 141 to 154: whatever tool ArcGIS has, these authors use it with no questions asked. A substantial literature exists on interpolation techniques for environmental data sets, particularly in meteorology and oceanography. These authors apparently look only in the ArcGIS tools manual. if the authors want to understand gridding techniques - and cautions - for ocean data sets they should look at the series of SOCAT products published in ESSD. Or, if they really want nuts and bolts of gridding and quality control for assimilation into global models, they should wade through the observations for model intercomparisons (Obs4MIPS) literature and guidelines. Normally I would not inflict Obs4MIPS but these authors seem in desperate need of a reality check.

Page 6, line 168: Again, the authors apparently pay no attention to unique or difficult properties of the data at hand but simply push the ArcGIS button for 'band statistics'.

Page 6, lines 172, 173: Compare with the same data sets used as sources? In a world of modern data denial and independent validation techniques, this can't be true?

Page 7, line 178: These authors accept all data inputs as already quality controlled? Nonsense; they have no idea what they have included nor why. By their strange choices - or non-choices, as they simply re-assimilated what others had already compiled - they determined final actual quality of their product in a manner that they apparently neither acknowledge nor understand.

Page 7, line 186: "ensure no significant error was introduced with the interpolation process." This is the best they can say about this entire effort? They have no idea of quality of what they put in, they compile this data of unknown quality into new layers at higher interpolated resolution, then assure readers that they introduced no fresh errors by the interpolation itself. Why do they somehow feel that such a weak effort stands up to the quality and reproducibility standards demonstrated by hundreds of valid data processing efforts documented in ESSD?

Page 7, lines 195 to 204: comparisons by looking at ranges of extreme or maximum values? Ludicrous. Do they not understand statistical techniques for data set intercomparisons and validations? They could learn a lot by looking at almost any ESSD paper.

Page 8, line 222: “derived from a more diverse set of sources”. The authors intend this sentence to convey a positive asset of the GMED product. In fact it represents a fatal deficiency because a) they have simply copied what others have already compiled and b) the list of what others have compiled conveys an appalling ignorance of modern ocean data sources which these authors fail to recognise. Obsolete and erroneous source data, no matter how skilfully interpolated, still results in obsolete erroneous data layers.

Page 9, starting from line 249: Finally the reader gains a reality check, of the probable quality of the GMED environmental “surfaces”. Readers find welcome cautions from these authors about quality of source data, although note in one sentence these authors disparage “raw observational data” but a few sentences later assure us - erroneously, as their own tables prove - that they have only used highest quality “Level 3” data. ‘Level 3’ primarily refers to satellite data and misses a large discussion (see matrix often reported by Bates) about maturity and quality of satellite data records. Their citation here covered a very small subset of ocean colour data and is now nearly 20 years of out date.

Page 9, lines 275,276: “verification data indicates that the GMED layers are reliable representations of the source data”. More true than the authors understand! Use unreliable source data, apply proprietary statistical interpolation tools, then “validate” by comparison back to the same deficient source data? Garbage in, leads to same garbage reliably represented in the end products. Do the authors not realise this inevitable consequence? Did they not expose their work to competent oceanographers and modellers?

Page 10, line 287: Unfortunately, readers find no evidence that these authors understand the “difference between a pure statistical and a more mechanistic expert-driven approach in interpolation”. These authors have certainly not shown any indication that they recognise a need for “expert-driven” guidance or advice before embarking on a “pure” but in fact sadly deficient and essentially useless statistical approach.

Page 10, line 293. Have the authors provided a useful improvement of the land masks? After reading the dismal state of ocean data they used as sources, one wishes for some tangible improvements, e.g. of land masks. As usual, however, authors provide minimal evidence and no citations.

Overall, a completely unsatisfying presentation of a basically flawed (one hesitates to say ‘incompetent’) effort.

To confirm my impressions about the dismal state of GMED source data, I made a quick scan of ocean data sets available from ESSD. (Authors could do - and arguably should have done - a more careful systematic but fundamentally similar scan.) As these authors apparently recognise (because they attempted to submit their own product) these ESSD-published sources provide up-to-date, well documented, permanently identified, easy access data in standard formats, known and used in the ocean community. My quick scan exposed data sets that cover 60 to 80% of the GMED parameters. These ESSD papers also include many references to other available data. Why the authors did not at least check their AquaMAP and KGS sources against these recent openly-accessible data remains a mystery.

Sources (*with my comments*):

The MAREDAT Special Issue
(especially) doi:10.5194/essd-5-109-2013, 25.3.2013, The MAREDAT global database of high performance liquid chromatography marine pigment measurements *(much better than any chl a product you reference)*

SOCAT *(nearly 15 million data points, 1957 to 2014, the definitive way to compile, grid and quality control ocean data)*
doi:10.5194/essd-5-125-2013, A Uniform, Quality Controlled Surface Ocean CO2 Atlas (SOCAT),
doi:10.5194/essd-5-145-2013, Surface Ocean CO2 Atlas (SOCAT) Gridded Data Products
doi:10.5194/essd-8-383-2016, A multi-decade record of high-quality fCO2 data in version 3 of the Surface Ocean CO2 Atlas (SOCAT)

doi:10.5194/essd-5-295-2013, 12.8.2013, Global database of surface ocean particulate organic carbon export fluxes diagnosed from the 234Th technique *(better POC than anything you have)*

doi:10.5194/essd-7-261-2015, 5.10.2015, Vertical distribution of chlorophyll a concentration and phytoplankton community composition from in situ fluorescence profiles: a first database for the global ocean *(instructive about challenges of compiling global ocean data)*

doi:10.5194/essd-8-15-2016, 1.2.2016, A gridded data set of upper-ocean hydrographic properties in the Weddell Gyre obtained by objective mapping of Argo float measurements *(example of the richness and processing of Argo data)*

doi:10.5194/essd-8-165-2016, 28.4.2016, A long-term record of blended satellite and in situ sea-surface temperature for climate monitoring, modeling and environmental studies *(arguably now the definitive SST data set)*

doi:10.5194/essd-8-235-2016, 3.6.2016, A compilation of global bio-optical in situ data for ocean-colour satellite applications *(interesting quality control, covers all your bio-optical parameters)*

GLODAP *(the definitive ocean data set, with extensive well-documented quality control)*
doi:10.5194/essd-8-297-2016, 15.8.2016, The Global Ocean Data Analysis Project version 2 (GLODAPv2) – an internally consistent data product for the world ocean
doi:10.5194/essd-8-325-2016, 15.8.2016, A new global interior ocean mapped climatology: the 1°× 1° GLODAP version 2

doi:10.5194/essd-8-531-2016, 20.10.2016, Global ocean particulate organic carbon flux merged with satellite parameters *(for ESA CCI, much-used and well-documented)*

doi:10.5194/essd-8-679-2016, 29.11.2016, C-GLORSv5: an improved multipurpose global ocean eddy-permitting physical reanalysis *(interesting re-analysis product, instructive on how they assimilate sparse data)*

https://doi.org/10.5194/essd-10-251-2018, 6.2.2018, Photosynthesis–irradiance parameters of marine phytoplankton: synthesis of a global data set *(another good data assembly example, highly relevant to your intended uses)*

https://doi.org/10.5194/essd-5-311-2013, A long-term and reproducible passive microwave sea ice concentration data record for climate studies and monitoring *(sea ice data at the highest quality level)*