Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2019-40-RC1, 2019

Interactive comment on "A machine learning based global sea-surface iodide distribution" *by* Tomás Sherwen et al.

Peer Johannes Nowack (Referee)

p.nowack@imperial.ac.uk

Received and published: 23 April 2019

General comments:

The paper by Sherwen et al. introduces a new dataset for monthly-mean sea-surface iodide concentrations. Their new machine learning approach to create the dataset is both appealing and promising because it can simultaneously account for observationally-constrained relationships between several predictors and iodide in an objective manner while capturing potentially complex functional dependencies. I find their approach interesting not just for the creation of this new dataset (which indeed could be used widely in atmospheric chemistry studies), but also for inference. For example, their work could motivate further research into the physical and biological

drivers of iodide changes at the sea-surface, or measurement campaigns in certain world regions. As such, the study could be of much broader interest than just for the creation of a new dataset.

Overall, the paper is well-written, easy to follow and scientifically thorough. It deserves rapid publication subject to some mostly very minor revisions/suggestions listed below.

The datasets discussed in the paper are indeed accessible through the provided link in the standard netCDF format.

Specific comments:

- In the abstract and main text: for readers less accustomed to global iodide datasets it would be good to explain in somewhat more detail the recommended application context of this dataset. Could it be used to represent iodine emissions in historical, or even future, climate change simulations (where e.g. SSTs are subject to change and have in fact already changed) or ozone hole studies, or is it really applicable only to present-day air quality studies? What are the general assumptions here given that you create a non-transient monthly-mean climatological dataset? Are there any transient effects in the training datasets and what time period have the observations been sampled over? Your citation implies a period from 1967-2018, but it would be good to state this explicitly.
- Abstract: I would say specifically that the sample size has increased by 45% to avoid misunderstandings.
- p.3/4, beginning of section 2: here might be a good place to state the time period and to mention that you make the approximation that the relationships are stationary (?).
- Section 2: in the abstract you mention the use of climatological ancillary fields. You don't specify a time period for the observations either: 'For each iodide ob-

servation, the nearest point in space and time was extracted from the high resolution gridded ancillary data. For the 31 iodide observations where a month was not available (Luther and Cole, 1988; Tsunogai Shizuo and Henmi, 1971; Wong and Cheng, 1998), an arbitrary month was chosen (of March for Northern hemispheric observations and September for Southern hemispheric observations)'. Does this mean that you simply regress iodine observations against the SST etc fields purely based on the seasonal climatology? Why would you not use the temperature at the actual time when the iodide sample was taken? In addition, why would one not just archive the random forest regressor and use this model to predict SST etc consistent iodide concentration interactively in simulations (consistent with the actual state simulated by the model)? Could you discuss these aspects briefly; not necessarily in the main paper but maybe in reply to this comment?

- p. 4 l.26: similar - just for clarification in this review; what is meant by a month was not available? The observations exist, but no corresponding time reference?
- p.4 l.5: it is a question of taste, but I would somehow prefer predictors, regressors, input variables, input features etc over independent variables, which can sometimes be misunderstood (even though not incorrect), see e.g. discussion here: https://stats.stackexchange.com/questions/357745/in-regression-analysis-why-do-we-call-independent-variables-independent [...] I definitely don't feel strongly about this, so this is entirely up to the authors to decide (ie they can leave as is).
- p.5 l.26: a reference for the stratified sampling approach or more detailed description possible?
- p.7: it is not entirely clear to me at this point in the paper if the RMSE improvement after outlier removal is due to (a) the outliers being removed prior to training (are not involved at all), or (b) due to the outliers being removed from the validation/test

data so that the error on these specific predictions is simply not included in the final evaluation (i.e. the algorithm is simply not good at predicting those large value outliers). I guess the last sentence of this section implies (a) is the case here, but maybe good to say explicitly in the same sentence (I later also noticed that you discuss the alternatives below, but better to clarify this aspect here, too).

- p.8 l.3: so this becomes effectively an ensemble of an ensemble method (which random forests are)? Not sure if some people could misunderstand that given that you mention random forests as an ensemble method in Figure caption 3; you might consider using another term than ensemble here? You can leave as is though.
- p.9 section 4.2: Could the features associated with deep bathymetry (see your Discussion on p.11 l20) be down to a non-realistic assumption of the importance of bathymetry in those regions (based on a biased training dataset)? A simple test would be to check the predictions of the best performing models that do not include DEPTH; do those also predict such structure? If not, it could imply that those models actually show better physical generalizability (as far as we know) and could, therefore, be the preferable option. There might simply not be enough measurements in the training set covering grid coordinates along the Atlantic Ridge and as a result, it does not show up as an important error contribution in the training dataset.
- Did you retrain your forest on the entire available observational dataset before making the final predictions using the best performing models during the cross-validation procedure? This might be advantageous because you would take into account all available observations in training your algorithm (while not changing any other tunable parameters).
- p. 10 l.35: Relative (?) uncertainties are largest. . .

- p.12 l.9-13: this could be misunderstood. Do you mean by 'trend' a spatial pattern? It kind of links to my question about the consideration of transient effects and both aspects could be discussed here.

- In our uploaded .nc files, there appears to be no mask over land surfaces even though you only provide iodide data for the sea-surface? Can you explain? How are these values to be interpreted by modellers?

- A1, p. 13 l.13: This could be an interesting feature to explore with other regression models which allow for extrapolation outside the training domain. I guess this 'flat' prediction could be due to the fact that the random forest hasn't seen many inputs representative of this area yet (e.g. in terms of SSTs)? Maybe looking at how predictor-output relationships behave at the boundaries (can it be extrapolated) would be promising? Not necessarily something to be considered for this paper, but for future data updates (i.e. just a thought that may be ignored).

- Figure 7: for consistency, wouldn't it make more sense to plot the average plus standard deviation of the observations as well? Currently, the comparison seems rather unfair towards the parameterisations and emphasizes high values in the observations that deviate much from the predictions.

Technical corrections/typing errors:

- p.2, formula (2): I know this is a unit conversion, but the extra 10e9 multiplication reads like a mistake. Would summarise the two factors into a single multiplication factor.

- p.3 l4: non-coastal

- p.3 l8-10: The choice of parameterisation (Eqn. 2 versus Eqn. 1) results in a difference of 50

- p.3 l16: formulation

- p.4 l23: typo; this is not described in section 2, but in section 3.3.

- p.5 l29: typo

- p.5 l35: revise sentence “All forests...”

- p.6 l10-12: sentence is difficult to read.

- p.6 l29: typo

- p.7 l18: typo

- p.10 l16: typo

- p.10 l32: typo

- p.12 l.10: typo

- Figure 6 caption: revise last sentence

Interactive comment on Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2019-40, 2019.