Ocean Sci. Discuss.,
doi:10.5194/os-2016-73-RC2, 2017

Ocean Science
Discussions
Open Access

Interactive comment on "Technical note: Evaluation of three machine learning models for surface ocean CO_2 mapping" *by* Jiye Zeng et al.

Anonymous Referee #2

Received and published: 23 January 2017

This "technical note" discusses the formation of global maps of surface ocean CO2 from limited measurements using inferred dependence on (latitude, surface temperature SST, salinity, chlorophyll concentration, mixed-layer depth, difference between monthly- and annual-mean SST). The dependence is inferred by three methods: self-organisation map (SOM), feedforward neural network (FNN) and a new method (support vector machine; SVM). The results of these three methods, "trained" on a fraction of the data, are compared with the remaining data. The correlations are not particularly good for any (best at R2 = 0.715 for SVM) considering there are 6 independent variables aiding the fit. However, the results of all three methods for global air-sea CO2 flux are very close and the CO2 maps are visually similar. This similarity extends to a band of high CO2 concentration in February 2005 extending west from Chile where there are apparently no CO2 measurements. This extrapolation from CO2 observations is

presumably via a similar feature in (at least) one of the 6 independent variables.

There should be more discussion: (i) of the quality of the fit to observed data, especially in relation to the estimates of air-sea flux and the danger that the methods agree with each other more than with reality; (ii) of the extrapolation feature west of Chile (in particular – perhaps also a careful examination for whether there are others) and whether it can be believed in terms of the values of the independent variables – is this set of six values closely approximated somewhere else where there are CO2 measurements constraining the CO2 estimate?

Although the organisation and English are generally good, I think some sections and especially the Appendix are unclear/obscure, mainly due to inconsistent or missing explanations, definitions or notation. Most of the following detailed comments are about this aspect.

(Section 2)

Page 2 lines 12 and 18. "dSST denotes the difference between the monthly and annual means of SST" implies 12 discrete values of dSST; how does this "improve expressing the seasonal variable continuously"?

(Section 4)

Page 3

Line 13. I think you mean ". . to the range (0, 1) for the SVM . . ."

Line 21 (i.e. line after (5)). Why between 0.1 and 0.9 not between 0 and 1? "better" compared with what? Why should scaling the output help?

Page 4

Lines 1-2. "We used Eq. (4) to scale . . SOM". There is no mention of this in Appendix A.1, indeed after (A1) it is stated that the diagonal factors of the scale matrix f are equal to 1.

Lines 2-3. “Based on preliminary studies, we applied a factor of 2 to . . SST and CHL . .”. What preliminary studies? Is this subjective, i.e. why should SST and CHL be emphasised?

Line 7. “prediction for an input” needs explaining. Inputs are supposed to be known, not “predicted”.

Lines 8, 9. “map size”. In normal language the map size is the earth’s surface area. Do you mean resolution, equivalent to the number of CO2 output locations? Please explain / use correct word.

(Section 5)

Page 5

Line 8. “respectively” should be “for SOM”

Lines 11, 15. Please explain “normalized”/“normalization”

Appendix, page 6. To have value, this needs to be understood in its own terms; the reader should not have to refer to cited references to understand the words used and the overall meaning. Too many words are not defined or explained. Also, it is too abstract. This is a manuscript about “output” CO2, depending on “inputs” LAT, SST, SSS, CHL, MLD, dSST. Presumably this applies to A.1, A.2 and A.3 – say so and do not use vague terms like “feature space” – at present the reader has to guess what you mean.

(A.1 . .)

Page 6

Line 23. What is “feature space” in oceanographic terms?

Lines 23-24. “usually represented by grid points in two dimensional space”. Never mind about “usually”; describe in terms of the problem here.

Line 24. “weight vector w”. This name is confusing. On page 7 lines 7-8 weights (weight factors) h are defined by (A3). “w” is the result of applying the weights “h” to combine values of “v” at various locations [presumably to represent “v” at grid locations rather than original locations, but this is not clear to the reader. If this the case, then “w” is “gridded v” or “interpolated v”]. See also the comment on page 7 line 21.

Line 25. Not “a data vector” which might refer to any vector at all, but “an input data vector” (I guess).

Line 30. “best matching cell (BMC)” needs explaining.

Line 30. “minimizing the distance”. What is varied to do this?

Page 7

Line 4. “matched”. Either this is the wrong word or it needs explaining.

(A.2 . .)

Page 7

Line 17. “vector x of input data”. In A.1 the input data were “v”. Use consistent names for variables.

Lines 20-22. You have input data, hidden neurons and output. There should be distinct variable names for each of these, e.g. v, x, y respectively. Here you have y for the hidden neurons and for the output, which is confusing.

Line 21. “w is the weight vector”. Indeed this seems correct for its use in (A4) but that is very different from its use in (A1). Use different terms for different quantities (c.f. comment on page 6 line 24).

Line 22. “The training updates the offset and weight parameters”. What are the starting values before updating? Do you mean “weight vector” as in line 21?

Line 23. What is “e” or is it defined by (A5)? Please make this clear.

Line 24. “modelled . . y” is unclear (especially because you use “y” for hidden neurons and output). Why are two “y” in this line in bold type but not the third or “y” in (A4)?

Line 24. “w includes both . .” This seems to be defining a vector with more components; it should have a new name.

Line 28. “α is the learning rate”. How is its value decided?

Line 30. “derivatives of e by w”. Do you mean “derivatives of e with respect to w”.

(A.3 . .)

Page 8

Lines 6-10. “The SVM . . SVM parameters.” Is this relevant?

Line 14. “independent variables”, “high dimensional space”, “target variable”. Please define these in terms of the oceanographic problem in question.

Line 16. “minimizes” – what is varied to do this?

Lines 18-19. “subjecting to the constraint”. (A11) looks like a definition of “e” and is not a constraint unless “e” is defined in some other way which needs to be stated.

Line 27. Can there be an explicit expression for φ? Where has “b” in (A9) gone to?

Table 1. SOM column half way down. “closest” not “closet”!

Figure 3 caption. Please explain “normalized to 2005”.

Interactive comment on Ocean Sci. Discuss., doi:10.5194/os-2016-73, 2016.