Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2019-236-RC1, 2020

Geoscientific Model Development Discussions

Interactive comment on "Retrieving monthly and interannual pH_T on the East China Sea shelf using an artificial neural network: ANN-pH_T-v1" *by* Xiaoshuang Li et al.

Richard Mills (Referee)

rtmills@anl.gov

Received and published: 5 June 2020

The authors describe the development of an artificial neural network model (ANN) for predicting water column total scale pH (pH_T) in the waters of the East China Sea shelf from observations of pH_T, temperature (T), salinity (S), dissolved oxygen (DO), nitrate (N), phosphate (P) and silicate (Si), together with latitude, longitude, and time of samples. The ANN is trained using data from eleven cruises conducted between 2013-2017, comprising 1854 records, and the model performance is assessed in three different ways: ten-fold cross validation, application to new data from cruises in 2018, and using prognostic variables from the Changjian Biology Finite-Volume Coastal Ocean

Model (FVCOM) as input to the ANN. The ANN appears to perform well in the authors' tests.

This paper is well-written, with excellent English and a logical development of ideas. In its current form, however, I think that the paper is a promising beginning but does not seem complete. I came away from my reading of the paper with the following major questions/concerns which, if addressed, will greatly improve the quality of the paper:

First, since the paper has been submitted to a model development journal, I would like to see more information on how and why the authors arrived at the particular form of the machine-learning model they used, and how this model performed against some other possible model architectures. The authors have used a feed-forward multilayer perceptron network with two hidden layers (with 40 neurons in the first layer and 16 in the second) and full connectivity between the layers. Why did the authors decide on two layers, and how did they choose the number of neurons in each layer? (They do state that they tried varying the number of neurons in each layer, but don't give further details.) How did they choose the activation function? And why did they choose a neural network, instead of another approach such as k-nearest neighbors, random forest regression, or support vector regression? When I first started working in machine learning, around two decades ago, it would not have been expected for authors to try a variety of different types of models, as this would likely involve substantial code development effort, as well as possibly significant computational expense for training models. Today, however, it is easy to try many different models, as code provided in many easily obtained packages such as Scikit-learn or those provided by Matlab (the environment that the authors use for this study), and it is becoming the norm for papers presenting the development of machine-learning models to compare several types to determine the one that performs best for the chosen task. I would like to see some comparison against other models (some of the ones easily constructed using Matlab) to demonstrate that the ANN is the most appropriate choice.

Second, the authors do a good job of citing other papers in which authors have used

similar ANN approaches for similar biogeochemical prediction tasks in marine waters, and compare the RMSE of their model with published values from other models. I think that the paper would be greatly improved if the authors could do a direct comparison. For instance, the authors cite the CANYON neural network model of Sauzede et al., 2017, which has been developed for the global ocean, but note that "coastal seas tend to show greater temporal and spatial variability than open oceans", which I believe is an argument for why they developed the model presented in their paper. I can easily imagine that the model presented here will outperform the CANYON model for prediction on the East China Sea shelf, but I think it would be interesting for the authors to demonstrate this: The CANYON model appears to be freely available online, and it would be interesting to see how much better a model trained speficially for the East China Sea shelf will outperform one developed for the global ocean.

Finally, the authors perform an intersting study in which they use prognostic variables from the Changjian Biology Finite-Volume Coastal Ocean Model (FVCOM) as input to their ANN model in order to recover the pH_T. I am not a marine biogeochemistry modeler, so perhaps I am missing something obvious, but I am guessing that mechanistic models like FVCOM can provide prognostic pH_T. Is this available from the FVCOM runs that were used, or could it be obtained using FVCOM, or ROMS, or another, similar model? If so, how would the prognostic pH_T from FVCOM (or similar) compare to the pH_T from the authors' own ANN model? And what is the motivation for using the ANN? Is it because it can potentially provide a more accurate pH_T, or because it can provide pH_T for situations in which it is not desirable to run a forward simulation or reanalysis to get the pH_T, or some other reason? This may be obvious to an marine biogeochemist, but I and many of the readers of GMD don't have this expertise. The motivation needs to be explained for the general GMD audience.

Detailed comments:

Lines 34-35: The authors state, while comparing ANNs to multiple linear regression, that ANNs have the advantage of not requiring 'an a priori model but rather "learn" the

model from existing data'. I think it would be more precise to say that they are non-parametric models and do not require assuming any underlying statistical distribution.

Lines 75 and 78: The authors say that samples were "poisoned" by addition of HgCl2. I think it may be more idiomatic to say "sterilized".

Line 81: "The final number of data used by the ANN model was 1854". I would say the final number of "observations" or "records", to be precise.

Line 94: The authors talk about a model being "over-matched". I believe that "over-fitted" is the term they mean.

There are problems with low resolution for all of the figures. Figure 1 is not really even readable. Figures need to be re-generated with much higher resolution, or using vector, rather than raster, formats.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2019-236, 2019.