Review of: “A global monthly field of seawater pH over 3 decades: a machine learning approach” by G. Zhong et al.; submitted to Earth System Science Data

Context and general comment :

The continuous absorption of anthropogenic CO2 by the ocean leads to ocean acidification, which threatens marine ecosystems. While the acidification rate has been extensively documented at the surface, data for deeper waters remain limited. Zhong et al. address this gap by presenting a comprehensive, monthly, four-dimensional, 1°×1° gridded global seawater pH dataset, covering the years 1992 to 2020 and depths from the surface to 2000 meters.

This dataset was developed using machine learning algorithms trained on pH observations from the Global Ocean Data Analysis Project (GLODAP). The methodology employed is a three-step process: 1) self-organizing map neural network for bioregionalization, 2) a stepwise algorithm for predictors selection, and 3) feed-forward neural networks (FFNN) for non-linear regression. The resulting pH product is a valuable resource for studying subsurface ocean acidification and for validating or initializing biogeochemical models. The product is made publicly available through the Marine Science Data Center of the Chinese Academy of Sciences.

Overall, the article is well-written and the figures are clearly presented.

Despite the significance of this new 3D pH product for the scientific community, the article has some notable shortcomings. There is a lack of details in the methodology section, which makes it challenging to fully evaluate the robustness of the method and comprehend the implications involved.

Specific Comments:

Title:

It may be valuable to the reader to add the information that the estimations are depth-resolved, resulting in a 3D product, which is the principal novelty of this methodology.

Abstract:

- *Lines 15-17 :* "Here, we present a monthly four-dimensional 1°×1° gridded product of global seawater pH, derived from a machine learning algorithm trained on pH observations at total scale and in-situ temperature from the Global Ocean Data Analysis Project (GLODAP)." : The role of temperature in the methodology is unclear. Even after reading the entire paper, the specific role of temperature compared to other inputs remains ambiguous.
- *Line 18 :* I suggest rephrasing the method description for clarity. Consider stating: "A three-step machine learning-based algorithm was used..."
- *Line 19 :* The term "stepwise" may not be clear to the readers. Consider elaborating or using a more descriptive term.

Introduction:

- The introduction appears to be missing some crucial references. For example, it would be beneficial to acknowledge that the methodology is inspired by the work of Landschützer et al. (2014) and references following ; i.e. a SOM-FNN approach. Additionally, it is important to mention that this SOM-FNN approach has already been applied to the 3D reconstruction of DIC by Keppler et al., 2020

(https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2020GB006571). These references should be cited to provide a more comprehensive background.
- L*ine 52 :* where existing pH surface products are listed, there is a lack of references and details. It is crucial to include comprehensive citations of existing products (e.g. LSCE-FFNN), especially those used for comparison in Table 4 (+ the missing ones, see my comment below).
- The paper does not explain why the product spans the period 1992-2020. It would be helpful to provide a rationale for this timeframe and discuss why it does not cover a longer period, both in the past and up to the present (i.e., year-1).
- L*ine 55:* The reference to GLODAP by Lauvset et al. (2022) refers to GLODAPv2.2022 and should be cited as such throughout the article (instead of 'GLODAP'). Additionally, if the authors re-run their model during the review process, it is suggested to use the latest version of GLODAP, i.e., GLODAPv2.2023. The updated reference can be found here: https://essd.copernicus.org/articles/16/2047/2024/essd-16-2047-2024-discussion.html

Methods:
Line 60 : It is unclear how temperature is used in the methodology. Additional explanation is needed to clarify its specific role and contribution.
*Lines 69-73 :*The inclusion of other indices, such as the Northern Oscillation Index, should be considered.
Lines 71-73 : The LSCE-FFNN product provides total alkalinity and DIC data monthly from 1985 to year-1: data available from the Copernicus Marine Service (https://data.marine.copernicus.eu/product/MULTIOBS_GLO_BIO_CARBON_SURFACE_REP_015_008/description). If authors refer to 3D products, then it has to be clearly mentioned. For 3D estimations, DIC is available monthly from 2004 to 2019 from MOBO-DIC (Keppler et al., 2020).

Section 2.2 :
- *Lines 103-105 :* The rationale for using these specific parameters to define bioregions needs clarification, especially if they are not significant in the stepwise algorithm for determining important parameters in relation to pH.
- *Line 106 :* The criteria and process for merging provinces with fewer than ten connected grids or less than 100 GLODAP pH measurements should be rephrased and/or detailed because it is unclear.
- *Line 107 :* The need for manual subdivision of provinces separated by continents requires further explanation. Why not using same bioregion even if it is not in the same ocean, it is possible that the uderlying processes are equivalent and so that the FFNN will be performant in both basin ?
- *Lines 112-114 :* The sentence on the division of ocean areas into different layers also requires further details on which drivers are important for each layer as it is stated that drivers differ depending on the layers. Moreover, following this statement, why using the same bioregions for deeper layers ?

Section 2.3 and Table 1 :
- The choice of a single-layer FFNN instead of a multi-layer network should be justified. Has this been tested ?
- The use of sin(Lat) as a predictor is questionable since latitude is not circular.
- Clarify how depth is used as a predictor and whether it corresponds to the depth of retrieval of the output or if the FFNN estimates X values for X depth levels.

- The choice of the ECCO2cube92 model should be discussed and better supported by citations in the text.
- MEI should be defined as the Multivariate ENSO Index.
- Adding a column to Table 1 to indicate which process each variable is associated with would be informative.
- Consider using merged satellite ocean color data products like OC-CCI or GlobColour for longer time series would help for future usability and sustainability.
- Provide details on how the most informative parameters were chosen and how hyperparameters (architecture, number of neurons) were handled in this stepwise process.
- Clarify how co-correlation among selected predictors was removed in the stepwise FFNN selection procedure.
- The sentence regarding additional FFNNs trained with predictors in Table S1 for polar areas and periods before August 2002 needs clarification.
- Discuss Tables 2 and 3 scientifically in the Results section to highlight important processes driving pH variability.
- Figure 3 is extremely difficult to understand and should be clarified or redesigned.
More generally, the section 2.3 is currently unclear and needs to be rewritten with more detailed explanations.

Section 2.4:
Line 191: The paragraph is unclear. The statement, “Therefore, the uncertainty of our pH product was directly estimated from the FFNN pH predicting errors, instead of synthesizing the inherent uncertainty of each used predictor product,” needs further clarification. How was this done?

Section 3.1:
- *Line 214 :* This interpretation might be overstated. The broader value range likely contributes to a better model fit, and pH values exhibit less variability at depth.
- *Figure 4 :* Authors might add the slopes of the linear regression to the statistics.
- *Line 235 :* The impact of the Oxygen Minimum Zone (OMZ) on the product should be discussed more in details.
- *Figure 5 :* This figure requires re-arrangement. The map should be larger, and pH differences against depth should be plotted with depth as the y-axis, as is more common for reading profiles. Additionally, including seasonal variability for each major basin along with yearly variability would be beneficial.
- In the validation section, it would be valuable to compare the global scale trend with the Copernicus Marine Service data: https://marine.copernicus.eu/access-data/ocean-monitoring-indicators/global-ocean-acidification-mean-sea-water-ph-time-series.
Moreover, it would be interesting to add comparison against qualified pH data from BGC-Argo dataset.
- *Figure 6 + text:* Comparing to other available pH time series would be interesting. These are listed in the recent ESSD paper by Lange et al. (2024): https://essd.copernicus.org/articles/16/1901/2024/. For instance, the Mediterranean Sea, where data from GLODAP V2 are very scarce, could be validated against the Dyfamed pH time series.
- *Figure 6 :* Discuss the extreme values not reconstructed by the FFNN in the text.

- *Line 254 :* Chau et al. (2022) may not be the best reference, as they are also model (ML)-based.
- *Line 261* : Describe in what specific ways the product differs from other products.
- *Table 4:* More products could be compared, such as Jiang et al. (2022): Remote Sensing of Global Sea Surface pH Based on Massive Underway Data and Machine Learning (https://doi.org/10.3390/rs14102366). Additionally, some products compared here have not been previously cited in the article (refer to the comment on the introduction). The effect of the different time ranges of the different products on the computation of trends should also be analyzed and discussed.

Section 3.1.2 and Figure 7: Not sure whether this paragraph and figure are necessary.

Section 3.2:
- *Lines 301-304 :* This issue is problematic and should be discussed in more details for the user. Additionally, the significant differences between the GLODAP climatology and this product at 1000 m in the Southern Ocean should be discussed/addressed.
- *Figure 9 :* The longitude of the zonal average should be specified in the caption and/or the text.
- *Section 3.2.2 :* The discontinuity problem requires more discussion, both methodologically (explaining why this issue occurs despite the use of the cross-boundary method) and in terms of implications for users. If local uncertainties are available, they should be included in the NetCDFs.

Section 5:
Authors should provide more concrete examples of applications for their product in the Conclusion.

Typos :
- *Line 78 :* was converted to a 1°×1° resolution by averaging ~~**16**~~ 0.25° grids into one 1° grid
- *Line 84 :* ~~**(**~~*
- *Line 116 :* Therefor**e**

Data:
I encountered an error when attempting to open the NetCDF file using R. The error message was as follows:

```
Dans nc_open("/home/user/Data/2012.nc") :
   WARNING file /home/user/Data/2012.nc is not compliant netCDF; variable pH  is numeric but has a character-type missing value! This is an error!  Compensating, but you should fix the file!
```

Although I didn't receive any warnings when using xarray with Python, this issue should be addressed to ensure compatibility with other tools. Additionally, when opening the dataset with xarray and attempting to plot it using the library's functions, I noticed that the longitude and latitude are reversed (not in the name), and the longitude is plotted on the y-axis. To enhance user-friendliness when using Python tools, it would be beneficial to adjust the format accordingly.

Regarding the availability of MATLAB code, I am not a MATLAB programmer, so I am unable to provide feedback on its use.