Response to reviewers

List of Main Actions（LMAs）:

LMA1: Figure 12 was modified to add a group of scatter plot. The numbering of the subplot in Figure 12 was modified.

LMA2: Added description of marine surface products to the abstract. Corresponding numbers in the text have been updated based on the revised subplot numbers of Figure 12.

LMA3: Responded to all of the reviewer's comments and questions regarding this paper in a response letter.

Response to Associate Editor:

Dear xiaohan zhang,

We are pleased to inform you that Henry Bittig posted a new Referee comment in the interactive discussion of the following ESSD preprint:

Title: High-spatiotemporal reconstruction of biogeochemical dynamics in Australia integrating satellites products and in-situ observations (2000–2022)

Author(s): Xiaohan Zhang et al.

MS No.: essd-2024-219

MS type: Data description paper

Please access the discussion at: https://essd.copernicus.org/preprints/essd-2024-219/#discussion

To log in, please use your Copernicus Office user ID 761271.

In case any questions arise, please do not hesitate to contact me.

Response:

Thank you very much for giving us the opportunity to submit the revised manuscript. We have carefully revised the manuscript according the comments and suggestions. The problems pointed by the reviewer have been fixed in the revised manuscript. Responses to the reviewers are given as below.

**Reviewer comments:**

In this work, the authors present a methodology to densify sparse observations of ocean biogeochemistry into a high resolution dataset with the help of machine learning methods. Their starting point is the CANYON-B neural networks, based on GLODAPv2 ocean chemistry sparse sampling, as well as high resolution satellite remote sensing, which are combined by machine learning methods (random forest) to build a high resolution data cube of ocean chemistry in coastal waters around Australia to help ecosystem monitoring and management.

The manuscript is clearly structured, easy to read and well-organized.

The main criticism I have is that the authors want to use open-ocean based parameterizations (CANYON-B) and apply them to highly dynamic coastal systems (Australian EEZ waters), where processes and dynamics do not match together.

Taking the author’s references as indication, the authors have a strong background in remote sensing, less so in sea going biogeochemistry and sample analysis:

The GLODAPv2 data set is a collection of past and recent basin scale, coast-to-coast repeat hydrography cruises, done through different programs such as WOCE or more recently GO-SHIP. The focus is on the deep open ocean. The GO-SHIP target is to cover the world’s ocean with a decadal repeat frequency, i.e., every repeat hydrography section should be covered at least once per 10 years. There are a handful of sections with “higher frequency” (meaning every 1 or 2 years), too, but they are mostly located in the Atlantic basin. In addition, research cruises have a tendency to occur in favourable weather conditions, and in fact there is a seasonal bias of the GLODAPv2 cruises towards summer rather than winter timing. To summarize, the GLODAPv2 dataset is based on multi-decadal observations, with multi-year time resolution between repeats at most. There is no unbiased seasonal sampling yet alone a seasonal resolution. In consequence, with CANYON-B being based entirely on the GLODAPv2 data (and the information captured within these data), CANYON-B’s relations are based on processes acting in the open ocean on multi-annual time scales but neither on seasonal or sub-seasonal processes nor on coastal gradients or complex dynamics.

Like the authors state in lines 40-47:

"Most large-scale ocean chemistry datasets are derived from infrequent ship-based surveys or fixed-point observatories, which are then interpolated to create continuous spatial fields. This interpolation, while necessary, introduces substantial uncertainties, particularly in dynamic regions where biogeochemical properties can vary significantly over short distances and time periods. Traditional interpolation methods [...] may not adequately capture complex gradients or the temporal variability of ocean processes [...]. Such shortcomings can lead to misleading representations of marine biogeochemical environments, potentially skewing our understanding of oceanic processes and their responses to environmental changes." the method used must be fit for the application.

Here, CANYON-B is not fit for seasonal or sub-seasonal application, nor is it fit for coastal waters, i.e., one of the assumptions of this work is invalid, unfortunately. Even if CANYON-B is a machine learning-based method and not a “traditional interpolation method” like the ones listed by the authors, it is a data-driven method that inherits the limits of its parent training dataset. If some information on different orders of magnitude (notably: seasonal dynamics; near-shore coastal processes) is not captured at all by the training data, even a fancy machine learning approach cannot infer that information. In addition, CANYON-B’s predictive skill is high on interior ocean biogeochemistry. It degrades towards the surface and for surface applications, where there is much stronger variability and where the tight biogeochemical coupling between oxygen cycling and CO2 cycling breaks down in waters in contact with the atmosphere (due to different time scales of air sea gas exchange). (I.e., oxygen becomes a less adequate predictor of the chemical species of interest in surface waters than in the ocean interior.)

I have therefore great concerns towards the validity of the presented data set. It may show some interesting dynamics, as those data interpolation methods can always provide you with an output number, but I wouldn’t assume them at all to be trustworthy or reliable.

The remainder of their approach, to use in-situ Argo or glider data (temperature, salinity, oxygen, chla, turbidity/POC) and to combine them with ocean colour remote sensing by random forest regression to inform about the biogeochemical conditions (of temperature, salinity, oxygen, chla, turbidity/POC) in coastal waters seems valid. Here, the scale of observations (weekly to monthly timescale, some km resolution) matches the scope of the data product aimed at. But as outlined above, the transfer of Argo/glider data to CANYON-B outputs of nutrients or carbon system parameters is an invalid step in the manuscript's setting in the Australian EEZ waters. The monthly trends of chemical variables cannot be considered as robust or reliable.

Only the validation against independent observations (section 4.2.2) would be somewhat suited to hint towards how strongly the CANYON-B derived quantities match with field observations. I would have expected a style like Figure 10 for the Figure 12 comparison, in which is it hard to make out details. In addition, a presentation of data grouped by month would have been helpful to assess the invalidity of the monthly time scale target (which seems to be discernible from Figure 13, where there are repeated (annual?) oscillations in the percentage error). What I can make out from Figure 12/13 and the provided statistics is that the CANYON-B-based products are able to hit the ballpark value for the actually observed carbon parameters and nutrients (as – at some point – the EEZ waters connect to the open ocean), but that the products fail to get the variability and dynamics.

A way to make their approach work would be to use the independent observations of the IMOS National Reference Stations and establish a regionalized/local “CANYON-B”-like parameterization to connect temperature/salinity/oxygen/... with the target parameters of interest (carbon parameters and nutrients) that is applicable to the system and timescales the authors work in.

Further comment: It remains unclear whether this is a surface data product or one that extends into the water column. I believe it captures the surface only, but this should be stated somewhere (e.g., title, abstract).

**Response to reviewers' comments:**

We are very grateful to the reviewers for their comments on our manuscript. Firstly, it should be clarified that the product produced by our work is indeed a ‘Marine surface product’, as we state in the abstract of the revised manuscript. All added text has been bolded in red, and deleted text has been bolded in blue with strikethrough.

**Review comments 1:** ‘the CANYON-B output for transferring Argo/glider data to nutrient or carbon system parameters is unreliable’.

**Author's Response:**

We thoroughly investigated this issue before proceeding with our work. Our work is mainly based on the following work:

The first work is ‘**Bittig, Henry C., et al. An alternative to static climatologies: Robust estimation of open ocean CO2 variables and nutrient concentrations from T, S, and O2 data using Bayesian neural networks.**’ Frontiers in Marine Science 5 (2018): 328.’ This is the article in which the CANYON-B model is presented, which states that ‘The quantitative analysis of errors between the CANYON-B model simulations and Argo observational values shows that the mean difference for the CANYON-B with the transfer data set is -1 ± 16 µatm.’

The second work is ‘**Mignot, Alexandre, et al. ’Using machine learning and Biogeochemical-Argo (BGC-Argo) floats to assess biogeochemical models and optimise observing system design.’ Biogeosciences 20.7 (2023): 1405-1422.**’ This work uses data from Argo floats as CANYON-B inputs, and uses the estimated data from the output CANYON-B outputs and the Argo observations to construct a comprehensive dataset of biogeochemical profiles. The article states that ‘CANYON-B estimates of NO3 and pH were merged with measurements based on the combination of the RMS error of the CANYON-B estimates (NO3 = 0.7 µ mol kg -1 and pH = 0.013) and the BGC-Argo observation error (NO3 = 0.5 µ mol kg -1and pH = 0.07) are of the same order of magnitude. ’

The third work is ‘**Asselot, Rémy, et al. “Anthropogenic carbon pathways towards the North Atlantic interior revealed by Argo-O2, neural networks and back-calculations.” Nature Communications 15.2024: 1630.**’ The article uses AT and DIC calculations using CANYON-B to reveal anthropogenic carbon pathways towards the North Atlantic interior.

The fourth work is ‘**Sauzède, Raphaëlle, and Hervé Claustre. ’Ocean Nutrient profiles vertical distribution Product MULTIOBS\_GLO\_BIO\_REP\_015\_006. (2019).**’ The publisher of this document is the European Centre for the Application of Marine and Meteorological Satellites (CMEMS, Copernicus Marine Environment Monitoring Service). This work utilises pressure, temperature, salinity and oxygen concentrations obtained from BGC-Argo data and estimated by the CANYON-B neural network method for nitrate, phosphate and silicate. The accuracy was evaluated as nitrate (NO3): root mean square error (RMSD) of 0.68 µmol/kg; phosphate (PO4) RMSD of 0.051 µmol/kg; and silicate (Si(OH)4): RMSD of 2.3 µmol/kg. The deviations of nitrate ranged from -0.55 to +0.57 µmol/kg with an uncertainty of 0.82 to 1.11 µmol/kg. Phosphate deviations ranged from -0.044 to +0.052 µmol/kg with an uncertainty of 0.059 to 0.080 µmol/kg. Silicate deviations ranged from -1.9 to +2.0 µmol/kg with an uncertainty of 2.4 to 3.9 µmol/kg. The article concludes that ‘BGC-Argo data used as CANYON-B inputs for nutrient concentration estimation at multiple sites around the world show good accuracy and low systematic bias.’

The four existing studies mentioned above proved that Argo data are feasible to be used as input data to CANYON-B, and the estimation results have good accuracy, low systematic bias results, and high reliability of the estimation results. It is based on this that we determined to input Argo data or glider data into CANYON-B to calculate the concentration of the relevant parameters.

**Review comments 2:** ‘CANYON-B is not suitable for seasonal or sub-seasonal applications.’

**Author's Response:**

We acknowledge that CANYON-B has this limitation. The aim of our work is to improve the temporal and spatial resolution of biogeochemical parameters by combining remotely sensed imagery with estimates from CANYON-B. MODIS remotely sensed imagery has the advantage of being time-series, having a high temporal resolution, and being able to cover a large area.

However, inversion of biogeochemical parameter concentrations by remote sensing images requires actual data to establish the relationship between image pixel values and concentrations.

We did obtain measured values through IMOS, and our initial idea was to use these data to construct a random forest model directly with MODIS imagery to invert concentrations. Unfortunately, these measured data were not spatially distributed uniformly, with very little distribution on the northern side of the study area, which would have had some impact on the concentration estimates. In addition, the IMOS measurements are from 2008. As the reviewer stated ‘the accuracy of data-driven methods depends on the parent training data set’, so it is not reasonable to use an inverse model trained on post-2008 data in conjunction with pre-2008 imagery to infer pre-2008 biogeochemical concentrations.

That is why we decided to use Argo data and glider data, which are uniformly distributed in both time and space, to estimate biogeochemical concentrations via CANYON-B's and then combine them with imagery to complete this work.

**Review comments 3:** ‘One way to make their approach work would be to use independent observations from IMOS national reference stations and build a regionalised/localised “CANYON-B” type parameterisation’.

**Author's Response:**

Thank you very much for this comment, and it is very unfortunate that, as answered in 2, problems with the timing of the actual data collection and the non-uniform distribution of the data forced us to abandon the idea of constructing a CANYON-B of parameters specific to the study region for that region.

Of course, there are three main things we did to ensure the quality of the products we produced.

The first one is the exact date-specific relational correspondence mentioned in 2 to reduce the effect of time differences on the accuracy of the product.

The second one is the five-fold cross-validation, which makes full use of the existing observation data for model training and validation by dividing the training set and test set several times to comprehensively use the data to evaluate the model's performance on different data subsets, to effectively measure the model's generalisation ability to reduce the risk of overfitting, and to ensure the accuracy of the constructed random forest inversion model.

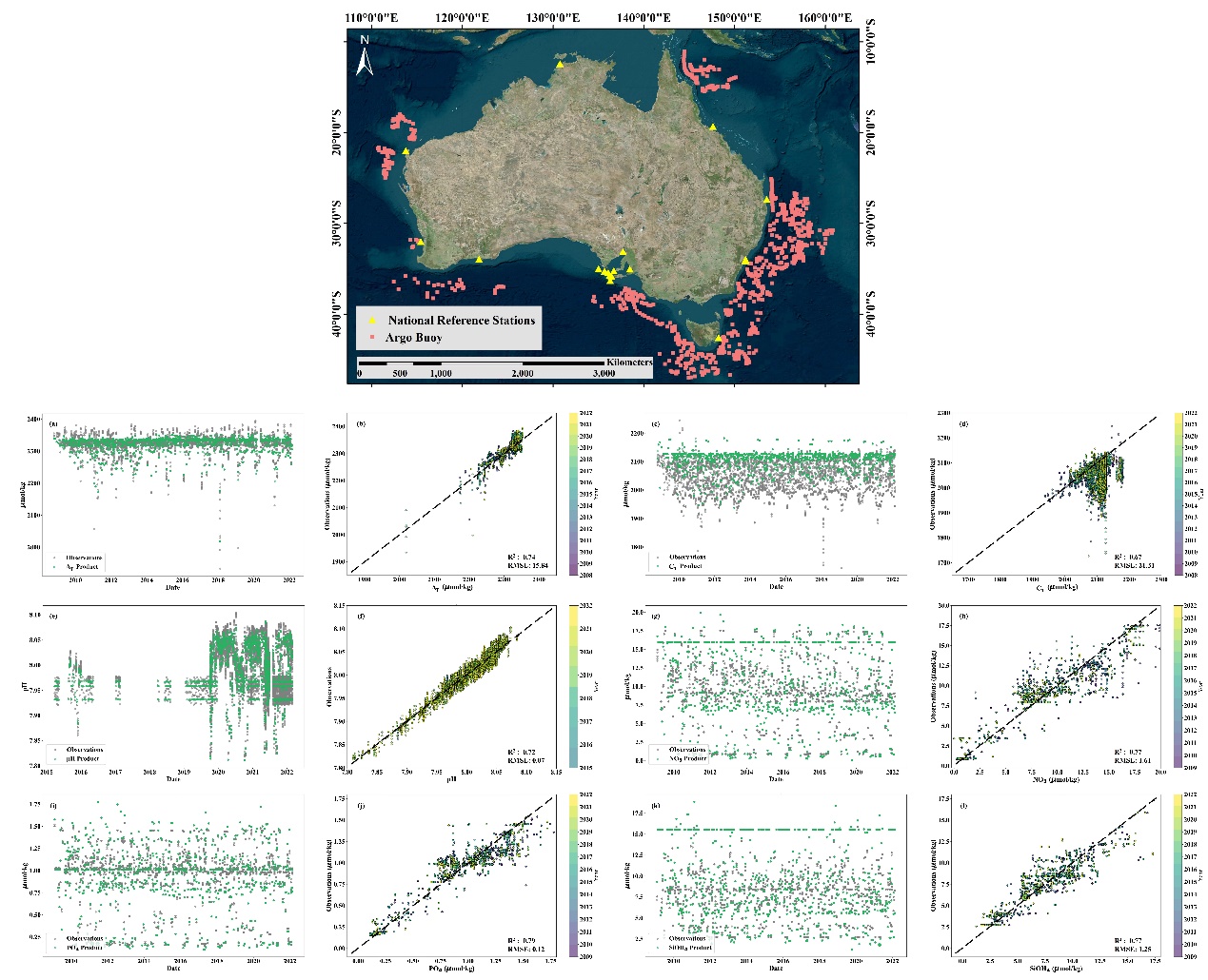
Finally, an observation dataset independent of the training data is used to provide a truer picture of the model performance in real-world applications. The use of independent observation data, such as IMOS national reference stations and Argo buoys, effectively avoids spurious accuracy caused by overfitting of the model to the training data, which directly proves the reliability of the product.

**Review comments 4:** I would have expected a style like Figure 10 for the Figure 12 comparison, in which is it hard to make out details.

**Author's Response:**

We modified Figure 12 to add the same scatterplot as Figure 10. And we modified the textual description of the manuscript regarding Fig. 12.

Figure 12 modified:



Description of the figure 12 (Line 307-316):

“……This verification includes six components, labelled (a) to **~~(f)~~(l)**, each demonstrating the validation of reconstructed ocean parameters against an ensemble of independent observations collected over time.”

“In Figure 12 (a) **and (b)**, ……. Figure 12 **~~(b)~~(c) and (d)** …….Proceeding to Figure 12 **~~(c)~~(e) and (f)** , …….

Figures 12 **~~(d), (e), and (f)~~** **(g) and (h) , (i) and (j), (k) and (l)** …….”

**Further comment:** It remains unclear whether this is a surface data product or one that extends into the water column. I believe it captures the surface only, but this should be stated somewhere (e.g., title, abstract).

**Author's Response:**

Abstract modified (Line 9):

“……. This work reconstructs the concentrations of these **marine surface** biogeochemicals at the sea surface within Australia's Exclusive Economic Zone over the period from 2000 to 2022 on a 1-kilometre scale. …….”