Responses to Reviewer #1: PyESPER

Overall great. The first iteration of this in Matlab was already sound in my opinion so this translation requires less scrutiny. I have not run the code myself, and although it would be intensive, I believe the accessibility would improve significantly if there is a possibility for a computer scientist to create a simple UI for either packages.

We thank you for your helpful and supportive review. We hope to implement a UI in the Puget Sound in the near future that is ESPER-inspired. This could serve as a template for a more global version. We will also investigate UI solutions that can be quickly implemented as a part of this product (if time allows during this review process) or next ESPER updates, such as Voila, Mercury, Panel, or JupyterDash.

40 - Should add note of the potential high error when using a model to estimate a variable then used to calculate carbonate chemistry parameter without nutrient information too

We understood this comment to imply that ESPERs offer an alternative to these high-error model estimates and added a sentence about this following the sentence in L40 that introduces ESPERs:

L43-44. This method offers an alternative to using models to estimate variables for carbonate chemistry calculations when nutrient information is unavailable, which potentially has high error values.

50 - I would argue that it may not be considered entirely findable for many scientists who are not coding competent and even those who are, are likely unaware of the Zenodo and GitHub repositories though I recognize that is not entirely your responsibility

Yes, this is a difficult barrier. We hope to develop a simple UI that is similar to ESPERs for the Puget Sound. If successful, this could be expanded in the future for the entirety of ESPERs. We have added a sentence addressing this possibility at the end of this section for now (see below) and are testing options for easy to implement UI's to add onto this version.

L55-56. Future updates may include even more accessible features such as a user interface.

68 - If all models perform comparably then why is there a need for all three why not just use the mixed as an ensemble prediction

We have added the following information to the bottom of section 2.1 regarding this valid question:

L. 72-87. There are a couple of reasons to maintain the separate ESPER LIR, NN, or Mixed options, from an end-user perspective, and these reasons are also true for PyESPERs.

1. **ESPER_LIRs predate the ESPER_NNs and have been used as a standalone data product for various research purposes (see Carter et al., 2016, doi: 10.1002/lom3.10087; Carter et al., 2018, doi: 10.1002/lom3.10232). Long-term users of these LIRs have previously expressed desire for consistency between versions (e.g., when depth was taken out as predictor for pH_T), and some of them already use CANYON-B (Bittig et al., 2019) as a neural net option for comparison. Therefore, these users who desire consistency would most likely prefer to use ESPER_LIR.**
2. **ESPER_LIRs are more transparent than ESPER_NN, as it is simple to parse apart coefficients at the gridded locations and easier to see how the equations are a result of these. ESPER_LIRs also rely on a grid, which may appeal to some users.**
3. **ESPER_NNs work a bit better on average than ESPER_LIRs, and work more like a mapping product in that 3D coordinates are predictors, which may alternately appeal to some users.**
4. **Although the ESPER_Mixed estimates perform better on average than LIRs or NNs do independently, there are cases where they have greater bias and RMSE than LIRs and/or NNs (e.g., when using equations 1-3 for phosphate or nitrate at all depths; Carter et al., 2021). Users may want to assess each scenario independently and choose which method is most appropriate according to their needs.**
5. **The NNs are more closely reproduced between the MATLAB and Python ESPER implementations.**

100 – if there's inadequate data number and the area size is doubled, does the output indicate this? Has it been checked if this correlates with an increase in error? Why is it jumping straight to double instead of small increase intervals?

We have added the following text to help explain the rationale of the windows:

> **L. 122-126. In LIRv2, windows were iteratively scaled by a factor of the iteration number until at least 100 measurements are selected to train each regression. For ESPER_LIRs (LIRv3), it is argued that increasing window size has the following benefits: (1) includes more data for regression fits, (2) introduced more modes of oceanographic variability into fitting data, and (3) reduced multicollinearity. However, the risk of increasing window size is that they will be less appropriate locally. The weighting term helps account for this (Carter et al., 2021, doi: 10.1002/lom3.10461).**
>
> **Here is the weighting term used:**

$$W = \max\left(5, \left(\frac{10(\Delta z)}{100+z}\right)^2 + (\cos(\text{lat})(\Delta\text{lon}))^2 + 4(\Delta\text{lat})^2\right)^{-2}.$$

There do remain instances where the windows need to be doubled, but these amount to 5 data points out of ~50,000 (for DIC in one previous version of ESPER; no triplings of windows occurred). A previous version of ESPERs did include the data needed to determine how many doublings were required with the release for each grid cell, but we did not provide means to interpolate that information to an arbitrary location and we found that these portions of the files were rarely used. In next version updates (where we have more freedom to change the overall methods rather than replicating past ESPER methods), we hope to investigate whether doubling of window sizes has an effect on error and, if so, to modify our methods to iteratively increase window sizes instead.

160/172 - Should add a caveat that in addition to not predicting past 2030 they should not be used in areas with abnormal atmospheric CO_2 absorption or profiles ie. upwelling, coastal areas, high freshwater outflow mentioned in 261 and may seem obvious to some but not others

Good point. We added the following statement:

L. 199-200. Likewise, these methods are not adequate for making reliable projections beyond the year 2030, or perhaps sooner in coastal or other areas where the underlying global open-ocean anthropogenic carbon estimations have greater uncertainties.