Zeng et al. used three machine algorithms (neural network, random forest, and gradient boosting) to estimate ocean pCO2 on a 1x1 grid from 1980-2020. They trained each algorithm to learn SOCAT fCO2 observations using full-coverage fields (SST, SSS, MLD, CHL, LAT, LON, YEAR) as inputs to each algorithm. The output from these algorithms were averaged to create the final product and a bulk parameterization was used to estimate flux. Their flux estimates were lower than the 6 products used in the global carbon budget 2021.

Major comments:
A lot of the methods used in this manuscript are poorly described, leaving the reader wondering whether best practices were used. A validation dataset separate from the training and testing datasets is typically used to determine the model architecture. The authors state on line 110 that data post-1980 was used to train the tree algorithms. There was no mention of whether an independent test-set was withheld or how the architecture of each algorithm was selected. For example, how did the authors decide on the number of layers, number of nodes, optimization algorithm, and weight initialization for the neural network? For the tree-based algorithms (random forest, gradient boosting) how did the authors decide on the number of trees and number of samples used to construct each tree?

When calculating atmospheric pCO2 the authors need to take into account the water vapor correction. The authors use the marine boundary reference product which uses dry-air mole fraction of CO2. Because of this, a correction for the water vapor that was removed needs to be taken into account, see Dickson et al. (2007). This correction is small can become important in the delta-pCO2 when calculating flux.

On line 124 the authors state the gas-transfer velocity and solubility were calculated following Wanninkhof (2014). I would have liked to see more discussion on this since the datasets used to calculated these terms were not clearly stated. The gas-transfer velocity depends on wind speed. However, if you estimate CO2 flux with a monthly resolution then the wind speed variance needs to be taken into account, see Wanninkhof (1992 and 2014). I am concerned the variance in the wind speed was not taken into account. The authors state they used monthly wind from ERA5 to calculate atmospheric pCO2, but there was no mention whether the monthly variance was used to calculate fluxes.

The evaluation of the product needs significant improvements. It is standard practice to test machine learning algorithms against a withheld test-set. This was either not done or not mentioned in the manuscript. I suspect this was not done since the authors state post-1980 observations were used to train the algorithms. There was no comparison to any independent datasets. The six products used in the GCB2021 compare to estimates of pCO2 from HOT and BATS time series and GLODAP, just to name a few. Even though pCO2 at HOT, BATS, and GLODAP is not directly measured, but inferred from the carbonate cycle, I strongly suggest this comparison be done since SOCAT observations are sparse, which the authors mention.

Instead, the authors compare their results against the 6 products in the GCB2021. This will tell you how well the algorithm compares to other products, but is not the same as comparing to

independent observations. Even this comparison to the other products has some flaws. Instead of comparing the pCO2 output the authors compare the flux outputs. They compare the flux from their product to the flux from the others, taking into account the differences in spatial and temporal coverage between the products. My issue with this is they do not re-calculate the flux for each product using a consistent approach. Instead, they compare to the flux calculated by each product's practitioner. The authors do acknowledge that the different wind products used by each practitioner can influence this comparison. I suggest the authors refer to Fay et al. (2021), which addresses this issue. The authors should re-calculate the flux from the products pCO2 using a consistent approach to alleviate this issue. I would also suggest comparing their flux against anthropogenic flux estimates from Denman et al. (2007) and Gruber et al. (2019).

Minor comments:
The authors learn SOCAT fCO2 and then convert to convert to pCO2 after using the method of Weiss (1974). The six pCO2 products used in the GCB2021 that the authors compare to do this conversion prior to training. I am curious if this has an impact on the results and which datasets were used in this conversion. For instance, the conversion requires SST and there was no mention of which dataset was used in this conversion.

References
Denman, K. L., Brasseur, G. P., Chidthaisong, A., Ciais, P., Cox, P. M., Dickinson, R. E., & Steffen, W. (2007). Couplings between changes in the climate system and biogeochemistry. In Climate change 2007: The physical science basis (pp. 499– 588). Cambridge University Press.

Dickson, A. G., Sabine, C. L., & Christian, J. R. (2007). Guide to best practices for ocean CO2 measurements. In PICES Special Publication 3 (pp. 191).

Fay, A. R., et al. (2021). SeaFlux: harmonization of air–sea CO 2 fluxes from surface pCO2 data products using a standardized approach. Earth System Science Data, 13(10), 4693-4710.

Gruber, N., Clement, D., Carter, B. R., Feely, R. A., Van Heuven, S., Hoppema, M., et al. (2019). The oceanic sink for anthropogenic CO2 from 1994 to 2007. Science, 363(6432), 1193– 1199.

Wanninkhof, R. (1992). Relationship between wind speed and gas exchange over the ocean. Journal of Geophysical Research, 97(C5), 7373– 7382.

Wanninkhof, R. (2014). Relationship between wind speed and gas exchange over the ocean revisited. Limnology and Oceanography: Methods, 12(6), 351– 362.