This paper applied the integrated observational dataset to train the classification of the EP El Niño, CP type El Niño, and La Niña with supervised learning and to investigate the ENSO diversity/complexity changes in multi-model large ensembles. Specifically, they found the supervised machine learning can reasonably classify ENSO events/types and the observed increase of CP El Niño events is within the range of internal variability, so does the ENSO amplitude and frequency changes. The research topic is interesting and necessary; however, there are issues in the machine learning setup and the goal/finding is not unique for machine learning. Therefore, this paper should not be accepted in Earth System Dynamics before major revisions.

A few major comments are followings:

ML related

1. The setup of the supervised learning uses the combination of 18 observational datasets However, the combination of 18 observational datasets may overweight a few events and have limited difference. For instance, the events after 1980 are covered for most datasets but the events before are only covered by half of them. The authors should discuss this issue and provide additional analyses in the supplementary. One suggestion is to test with subgroup of the datasets. Another issue for the integrated observational datasets is the lack of differences for the dataset. Even though the reconstructions are all slightly different, the SSTs are still representing the same events. That is, the actual events consider in this study is only 14 CP, 20 EP, and 26 LN. This issue should be mentioned in the manuscript and needs to be tested with a small subgroup (or even extremely just one dataset) of datasets.

2. The setup of the supervised learning uses the features from 5 regions from October to March. However, limited dynamical reasons are provided and other regions and times should be mentioned (or even tested). For example, the authors show results from the smaller regions and times in the supplementary, but not larger regions and times. For instance, the north subtropical region is known to be important for the onset of CP El Niño and recent papers have found an improvement from including it (Tseng et al., 2021). And the summer is related to how specific ENSO type is onset (Yu & Fang 2018). The authors should provide dynamical reasons for the choice of the regions and times, otherwise, the study should examine more regions and times for showing the current choice is an optimal one.

Yu, J. Y., & Fang, S. W. (2018). The distinct contributions of the seasonal footprinting and charged-discharged mechanisms to ENSO complexity. *Geophysical Research Letters*, *45*(13), 6611-6618.

Tseng, Y. H., Huang, J. H., & Chen, H. C. Improving the Predictability of Two Types of ENSO by the Characteristics of Extratropical Precursors. *Geophysical Research Letters*, e2021GL097190.

Writing-related

3. The introduction is a little bit lengthy. It will be easier to read if the authors make the description more succinct. For example, the paragraph for observed CP increased (55-63) should be combined with the EP/CP introduction in the beginning. It will be great if the introduction can be better organized.

4. The ENSO complexity is recently considered with a broader perspective (Timmermann et al. 2018). Besides the EP/CP types of ENSO, the transition, propagation, and duration of ENSO are all parts of the ENSO complexity (Chen et al., 2017; Fang et al., 2020). Although these are not the focus in this paper, the ENSO complexity should be mentioned at least in the discussion section.

Timmermann, A., An, S. I., Kug, J. S., Jin, F. F., Cai, W., Capotondi, A., ... & Zhang, X. (2018). El Niño–southern oscillation complexity. *Nature*, *559*(7715), 535-545.

Fang, S. W., & Yu, J. Y. (2020). Contrasting transition complexity between El Niño and La Niña: observations and CMIP5/6 models. *Geophysical Research Letters*, *47*(16), e2020GL088926.

Chen, C., Cane, M. A., Wittenberg, A. T., & Chen, D. (2017). ENSO in the CMIP5 simulations: Life cycles, diversity, and responses to climate change. *Journal of Climate*, *30*(2), 775-801.

Interpretation-related

5. The study considers the classification of CP El Niño from Pascolini-Campbell et al. (2015) for the past 120 years, which combine various CP classification methods, but no classification is applied in the multi-model large ensembles. That is, the original CP classification is not compared with the supervised learning method in the SMILEs. If the method in Pascolini-Campbell et al. (2015) is too complicated, the authors should at least choose one or two existing method to justify how the existing classification in SMILEs is different with the one from supervised learning.

6. The goal/finding is not unique for machine learning and have been discussed in studies. The authors classify ENSO events and compare the results for SMILEs. However, this can also be done by simply using existing ENSO classification method (Ng et al., 2021). The finding of this study should focus more on the uniqueness of the supervised learning. For example, since the classification method is trained from observational dataset, how each modeled ENSO in SMILEs is different with the observation? Or is machine learning do a better classification than existing methods?

Ng, B., Cai, W., Cowan, T., & Bi, D. (2021). Impacts of low-frequency internal climate variability and greenhouse warming on El Niño–Southern Oscillation. *Journal of Climate*, *34*(6), 2205-2218.

7. The authors compare the changes of SST pattern for the EP and CP El Niño under global warming. The interpretation should be more dynamics, as this change in pattern is seldom mentioned in other studies (maybe due to the difficulty of dynamical interpretation). I will suggest the authors to eliminate this result if no dynamical explanation is provided, as this is only discussed in one paragraph (292-302). Instead, the author can focus on the change in zonal SST gradient in the mean state and compare with the frequency or amplitude.

8. The comparison of the increased CP El Niño frequency to SMILEs should be more precise. The authors use the ensemble spreads in each year to consider as the range of change for the internal variability; however, this is different with the increased CP El Niño frequency over a certain period. The authors should check how large the CP El Niño frequency can change in each ensemble and discuss the spread of the changes for all SMILEs.

Minor comments are provided below:

1. Does the training and classification use the original SST or SST anomalies? Please clearly describe in the text.
2. The calculation of frequency should also be mentioned in the method section, not only in the caption of Figure 3.
3. The Figure 6 is a bit difficult to read as there are many colors and lines.
4. Line 205, ‘to far’
5. Line 48, ‘niños’