Earth System Dynamics Discussions

Earth Syst. Dynam. Discuss., referee comment RC1
https://doi.org/10.5194/esd-2021-105-RC1, 2022

Comment on esd-2021-105

Anonymous Referee #1

Referee comment on "Combining machine learning and SMILEs to classify, better understand, and project changes in ENSO events" by Nicola Maher et al., Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2021-105-RC1, 2022

general comments

This paper integrates 18 observational datasets and machine learning algorithms (supervised classification) to classify the CP (Central Pacific), EP (Eastern Pacific), and LN (La Nina) events in the past ~120 years. The trained/tuned model was then applied to SMILEs (single model initial-condition large ensembles) to investigate both the internal variability and forced changes in each ENSO event type. The main findings from this study are 1) machine learning (ML) does a nice job in reconstructing the ENSO events in the past 2) the observed increase in the frequency of CP events after the late 1970s is within the range of internal variability in the SMILEs (thus arguing against climate change as the cause) 3) the ML algorithm doesn't project a change in CP frequency or amplitude in the following decades.

I find this paper well written, bearing important scientific merits, and nicely integrating climate model and machine learning. However, I do have several concerns and I hope the author could address them.

specific comments

ML related.

1) Metrics and scoring. The author used precision as their main metrics to check their model performance. However, as successfully detecting the CP and EP events is the most critical part, I think the author should use the recall rate. Imagine this scenario: if we have 20 total EP events, the ML successfully categorizes 5 of them to EP, and the other 15 are categorized to other types of events, and no more other events are categorized to EP,

then based on the precision formula, the precision for EP will be 5/(5+0) = 1. However, the other 15 EP events are not captured by this. If using recall, then 5/(5+15) = 0.25 and it indicates the model needs to be improved. Although recall also has its own issue, I think at least a thorough explanation of why the authors chose to use precision needs to be there. And I recommend the authors compare the results of using recall compared to precision.

2) The author used several methods to determine/evaluate the ML model (e.g., train and evaluation/test, for the train dataset, use 10-fold cross validation). I think the author also needs to explain how they tune the training model. For example (in Table 2), why they choose 1 in the KNN, why they use the specific hidden layers and max iterations in their NN. More importantly, for the random forest algorithm, the max depth seems to be too big (500). A more detailed description of how they tune (not only evaluate) the models is needed as it will change the final model structure.

3) Can the author explain why they first use HadISST as the test data set? As the author mentioned, this will cover all the events through time and is not the ideal way to evaluate the model performance. I think their second approach is more appropriate (randomly split the events across all augmentation data sets). I suggest the authors delete the HadISST part (unless I miss something…).

4) The authors need to discuss whether the range of the feature values during the training will also cover the ranges for future predictions. One example is the random forest, whose prediction results will be capped by the data used for training. In a future warming world, will the features have values that are out of the scope of the current observational ones?

5) In line 90, for those that don't quite know ML, I suggest the authors add a sentence or two to explain labelled dataset vs. unlabelled data.

6) line 180, "We additionally complete this split 100 times and manually choose 10 data splits that take CP and EP (the classes with the lowest numbers of events) from across the time-series, ensuring that not all events in the split come from the same part of the observational record.". I feel a bit loss here. Does it mean the events in any split needs to cover the whole time period? Needs to be reworded or adding more details.

7) Line 165: We have 14 CP in total (see line 110), why we only have 13 here (12/13)?

Model interpretation related.

1) A very interesting finding (and important!) from this study is that due to the interval

variability of the SMILEs, the assumed change in the frequency and amplitude of the ENSO events can be covered by the models themselves (Instead of required further forcing such as climate change). This is great but I am wondering if this could simply serve as the explanation of the change in the observational trend of ENSO events. For example, in figure 3, for the CP events, the HadISST shows a significant increase in CP frequency. Although the SMILEs cover this increase, it is mainly due to the wide band between minimum and maximum, the trend by SMILE is relatively flat (or slight decrease or increase). The authors need more nuanced explanation.

2) Line 75, The author needs to explain "undersampling internal variability" here.

technical corrections Line 40: change to "is uncertain, sparse, and intermittent"

Line 100: in 5. change to "to use the evaluation set to assess". In 6. Add "better" before performance

Line 130: change "but chose" to "we chose"

Line 330: "too located too far west" reads not ideal

Line 345: delete "are" before needed to evaluate ENSO