

## ***Interactive comment on “ClimateLearn: A machine-learning approach for climate prediction using network measures” by Q. Y. Feng et al.***

### **Anonymous Referee #1**

Received and published: 16 March 2016

This manuscript seeks to provide a toolbox for machine-learning in climate science, and demonstrates its use for a sample application, namely using the machine-learning methods in the toolbox, combined with other tools to calculate climate networks, to come up with better ENSO predictions. My biggest concern is that the manuscript is focused almost entirely on the sample application, and there is very little information on the actual toolbox, which I thought was supposed to be the main topic of the paper. I think you should decide on what you want to promote in your paper. Is it really the general toolbox, or, as it currently seems, the specific application, namely enhancing the ENSO prediction obtained by climate network measures by adding ANNs and GP on top? As it is currently written, it may be stronger on the application, less so on the toolbox.

Please see also the technical questions and comments below.

C1

---

A) QUESTIONS REGARDING THE TOOLBOX The toolbox is supposed to cover three areas (1) Data cleaning; (2) Artificial neural networks (ANN); (3) Symbolic regression with genetic programming (GP).

However, the data cleaning capabilities are never discussed in detail. Furthermore, on lines 345-346 the authors mention that they utilize two open source packages, Weka and ECJ. Since Weka implements ANNs, and ECJ implements GP, I can't help but wonder whether the ClimateLearn package is just a repackaging of those two packages?

My recommendations would be to

- 1) Clearly define the capabilities of the Climate Learn package;
- 2) Clearly state the contributions of the authors, in relationship to the integrated Weka and ECJ packages. What is new? What exactly did the authors contribute?

---

B) QUESTIONS REGARDING THE ENSO APPLICATION:

- 1) Why do you include time,  $t$ , as one of your attributes? Please motivate.
- 2) Why did you select these two particular methods (ANN, GP) for this application (and for the toolbox)? Why are those two methods particularly suitable? Right now this approach seems fairly unmotivated. Would simpler - and more transparent - methods not be able to achieve similar improvements? In my experience, it is usually best to use the simplest method that does the job - and the two proposed methods are neither very simple, nor very transparent.
- 3) I am not completely convinced by the results you present, given that the sample size of predicted events is so small (3-4). Rather than using just one partition for training/testing, why not use several partitions, i.e. once leaving early events, once leaving

C2

late events out of the training set, and see how well the system predicts any of those? Furthermore, I think a comparison to just using simpler regression/fitting/supervised learning methods on top of the CN measures would be in order to justify using the more complex methods.

Lastly, I will leave it up to my colleagues on the climate side to judge the quality of the ENSO results, as I am not an expert in this area.

---

Interactive comment on Geosci. Model Dev. Discuss., doi:10.5194/gmd-2015-273, 2016.