Nat. Hazards Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/nhess-2020-67-RC1, 2020
© Author(s) 2020. This work is distributed under
the Creative Commons Attribution 4.0 License.

Natural Hazards and Earth System Sciences
Discussions

Interactive comment on "Skill of large-scale seasonal drought impact forecasts" *by* Samuel J. Sutanto et al.

Anonymous Referee #1

Received and published: 4 April 2020

General Comments: The manuscript submitted showed the efforts devoted to predict drought impacts with lead-times up to 7 months ahead, using the Logistic Regression and Random Forest machine learning approaches. The idea of relating the drought indices to the drought impacts is relatively new and relevant to the journal's scope of understanding the natural hazards and their consequences. However, the machine learning approaches adopted are relatively old-fashioned. It would be nice if the authors can provide better justification for the selected approaches over other methods available. Besides, there are some queries on some statement made by the authors to be justified. Detailed Comments: 1) Abstract: The authors are advised to include more results in the abstract to provide better overview for the readers. 2) Page 1, Line 3: Kindly revise "with a lead-time of 7 months ahead" to "with lead-times up to

7 months ahead" as the study produces predictions with lead-time of 1-,2-,3-,4-,5-,6- and 7-months ahead, not only 7-month. 3) Page 2, Line 40: Kindly revise "Energy and Industry Pubic Water Supply" to "Energy and Industry Public Water Supply" 4) Page 2, Line 40 – 46: The literature reviews show that Logistic Regression (LR) and Random Forest (RF) are already well studied in different studies for deriving the link between drought hazard and their impact. - May I know why are these two methods selected as the approaches in this study? As there are many other approaches available to be further investigated, such as Artificial Neural Network and etc. - Besides, the methods compared have different nature LR (Linear) and RF (Nonlinear). Shouldn't we test the data's linearity before adopting either of these methods? As it will be unfair to the LR (RF) if the nature of the data is nonlinear (linear)? - Recommended recent paper: Drought forecasting: A review of modelling approaches 2007–2017. Journal of Water and Climate Change. 2019 5) Page 2, Line 58: the symbol "box 1" is confusing, kindly revise as "box i" (similar correction for the caption in Figure 1) 6) Page 3, Line 64: Kindly state the full-form of every abbreviation when it is first used, e.g. SRI-x 7) Page 5, Line 153: It is stated that the RF is able to avoid overfitting. To my best knowledge, this statement is wrong as RF does overfit although the generalization error does not increase when the tree size increases. Kindly justify how do the authors avoid overfitting in the current study? How significance is the difference if the cross-validation was adopted? 8) Discussion: The RF showed better performance and the authors claimed that it was due to the longer memory of RF compared to LR. However, the authors never mention about the linearity of the data. Could it be due to the linear/nonlinear nature of the data? Based on the results available, it seems that nonlinear models are favourable, have the authors compare the performance of RF with other nonlinear models? e.g. ANN, Deep learning, and etc. 9) Supporting information, Figure S2: The y-label of histogram for Log Regression is wrong, kindly revise. Besides, may I know how do the authors summarize the predictor importance of few counties into one histogram?

Interactive comment on Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2020-67, 2020.