Earth Syst. Dynam. Discuss.,
https://doi.org/10.5194/esd-2018-13-RC1, 2018

Earth System Dynamics
Discussions

Interactive comment on "Using Network Theory and Machine Learning to predict El Niño" *by* Peter D. Nooteboom et al.

R. Link (Referee)

robert.link@pnnl.gov

Received and published: 28 March 2018

General comments

In "Using Network Theory and Machine Learning to predict El Niño" the authors develop a predictive model for the NINO3.4 index of El Niño strength. The model uses network theory to select a set of predictors to use in the regression. The predictions are generated by summing an ARIMA function with the output of a neural network with the predictors as inputs. This design can be thought of as an autoregressive extrapolation of trends in the time series, modified by modified by shocks forecast by the predictors. This model design is an interesting and innovative approach to the problem. However, the paper suffers from several major flaws that call the results into question.

Printer-friendly version

Discussion paper

The first is the unusual design of the cross-validation calculation. The initial description on p. 7 of the separation into training and testing sets is standard, and the authors make an important point:

> Note that, since we are predicting time series, for any training set $[t_i^{train}, t_f^{train}]$ and test set $[t_i^{test}, t_f^{test}]$, $t_i^{test} > t_f^{train}$ must hold...

This is entirely correct, but on p. 16 the authors acknowledge that they violate this condition in their cross-validation experiment. Additionally, in that same section they appear to treat cross-validation calculations with different relative sizes of testing to training sets, run on the *same* dataset as independent cross-validation experiments, which they definitely are not. Together, these factors render the entire cross-validation exercise highly questionable, particularly where the results depicted in Figure 11, and any conclusions derived from them, are concerned. In particular, it seems likely that the peaks in Figure 11 are a reflection of the fact that many of the testing sets used in the result overlap with the training set, and not a realistic estimate of the model's likely performance out of sample.

A related problem is the paper's treatment of hyperparameter tuning. The authors do not provide a list of the hyperparameters used in the model, but certainly the p, q, and d parameters of the ARIMA model qualify, as do the number and sizes of the neural network layers. Possibly the choice of predictors and their lead times are another set of hyperparameters, although possibly not, if they were chosen exclusively based on the Z-C model results. The paper is vague on this point, but several passages, such as this one:

> Deciding which of the variables to use is not a straightforward problem, yet crucial for the eventual prediction. Sometimes a pair of two variables can be compatible in the prediction, but perform poorly when applied alone...

suggest that the predictor choice was tuned using the data. Indeed, the entire subject of how the hyperparameters were tuned is not discussed at all. This, combined with the problems with the cross-validation, suggests that the tuning of hyperparameters is likely to have caused substantial overfitting in the model.

I also found it rather difficult to understand the intended operation of the model. One might expect that the model is meant to be applied starting at some $t = t_0$ and working forward step by step, presumably with the model fidelity degrading the further the forecast is pushed into the future. However, the paper presents a family of three models tuned for different lead times, each with different model structures, and in one case different predictors. Since each model can make a forecast at any future time by either extending the forecast (for the short lead time models) or by using the intermediate steps from equation (14) (for the long lead time models), it is not clear how these variants on the model are meant to be reconciled. It's possible that they are intended to be averaged or used in some other boosting procedure, but if so, this is not adequately explained.

Finally, the paper's confusing structure makes it very difficult for readers to work out the exact details of the modeling and validation procedures. Much seemingly irrelevant information is included, some important information is left out, and detailed explanations are often deferred until later in the paper, well past when the topics they pertain to are introduced. A major contributor to this confusion is the bottom-up organization of the paper. Calculations are introduced early in the discussion without context (and sometimes, as in §2.3, without even a clear indication of what variables the calculations are being applied to). Later on, these calculations are assembled into a final product, but in the meantime readers are left with little guide as to why the constituent calculations are being done a certain way, which calculations are significant and which are merely asides, how the pieces being described will eventually fit together, and so on. The paper would be a lot clearer if it provided more context early in the discussion, so that readers can more easily understand what role each of these calculations will eventually

play in the final model.

Specific comments

At no point are we ever told what activation function was used for the neural networks.

In Figure 9 on p. 14 the NRMSE loss function for the three variants of the model compared to the corresponding figures for the CFSv2 ensemble mean. The loss values quoted in the figure are:

Lead time	CFSv2 loss	Hybrid model loss
3–4 mo.	0.17	0.16
6 mo.	0.21	0.18
12 mo.	N/A	0.17

Is the reported difference between the Hybrid model and the CFSv2 a substantial improvement? The performance of the 3–4 month models looks nearly equivalent, and even the 0.03 difference in NRMSE for the 6 month model looks likely to be within the range of variation in the models' performance over different datasets. What argument can the authors make to support the idea that this model will produce materially better ENSO predictions than existing models?

Section 2.2, covering the Zebiak-Cane model goes into a lot of detail that doesn't seem strictly germane to how the Z-C model will be used in the construction of the predictive model. On the other hand, the single most important detail, namely, the outputs of the Z-C model that will be used in the construction of the predictive model, is omitted. This section also gives a lengthy discussion of a procedure for adding noise to the Z-C results, but the purpose of adding this noise is not explained.

In the introduction there is a reference to the Alpha Go project. This isn't really relevant to the topic of this paper. First of all, the neural networks used in Alpha Go are much

more complex than the ones used here. Second, the tasks they are being asked to perform are quite different from the task described here. Therefore, the success of the neural networks in that project doesn't tell us much about what kind of success we might expect in this application.

Appendix A seems a little extraneous. A.1 is a restatement of the equation for the Pearson correlation coefficient. This statistic is well-known, and its definition need not be repeated here. The statistic in A.2, on the other hand, does merit description, but it is not clear what it is actually used for in the analysis. It seems to be mentioned at the end of section 2.3 and then not used again.

Interactive comment on Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2018-13, 2018.

Printer-friendly version

Discussion paper