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.

We would like to thank Robert Link for his careful reading and his constructive comments.

Please find our replies and the points that will be changed in the revised manuscript below.

On behalf of all the authors,

Peter Nooteboom

1 Major Comments

1. 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.

Author's response

Most of the results in the manuscript do satisfy the constraint $t_i^{test} > t_f^{train}$ above (see figures 8, 9 , 10, 12 of the old manuscript). To satisfy the constraint is convienient in these results, from the intuitive idea that the model is first trained on all data in the past to make a real prediction in the future, as is done in Fig. 12 (which is not a hindcast). It would be more clear if we state here that this condition 'is convenient' in stead of 'must hold.'

However, for the cross-validation method in Fig. 11 (enumeration in the previously submitted version), it is difficult to meet this condition, since the observational time series are too short. As stressed in [1], a cross-validation which only considers a last block such as in figures 9 and 10 (enumeration in the previously submitted version), does not make full use of the data. For the validation method of Fig. 11 we follow Ref. [1] in which it is empirically shown, and justified, that violating the constraint $t_i^{test} > t_f^{train}$ could be acceptable in some cases and lead to an improved performance. Another motivation for this cross-validation method is that asymptotic behavior from theory might behave differently on small test sets. Nevertheless in the rest of our calculations we respect $t_i^{test} > t_f^{train}$.

Changes in manuscript

We will change 'must hold' at page 7, line 6 into 'is convenient.'
We will include reference [1], and we will explain why we chose this type of cross-validation in one of the calculations in the revised manuscript.

2. 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.

Author's response

Thank you for mentioning this point. The cross-validation experiments with different relative sizes are presented to check if the size of the training and test set matters. One might expect that a shorter training set could decrease the prediction skill, simply because there is less data for the model to train. This means that different percentage splits could overlap in time. However, it is true that the manuscript should contain an explanation on why the different relative sizes of training and test sets are considered.

Changes in manuscript

In the revised manuscript it will be explained why the different relative sizes of training and test sets are considered in the cross-validation.

3. 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.

Author's response

From the previous two comments it is clear that we use this type of cross-validation in this particular figure to make full use of the available data, as explained in Ref. [1]. Also, the objective of this figure is to show the stability of the method with different sizes of the training and testing sets.

Changes in manuscript

In the revised manuscript it will be explained why this type of cross-validation method is chosen.

4. 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.

Author's response

The ANN structure is indeed tuned on the data. Therefore, besides the cross validation, Fig. 10 is included to show that this structure can be generalized and more structures lead to a similar result, which is evidence that they converge to a similar function from predictor to predictant.
The order of the ARIMA(p,d,q) model is not tuned. We just present the results where $p = 12$ to consider information up to a year ahead, with which we already obtain good results.
The choice of the predictors was mainly based on the ZC-model results which identify the physical reasons that would lead to a good prediction. This improved the search for attributes which would contain important information for prediction, but remain relatively independent. By choosing them at a specific lag, also their performance, cross-correlation and Wiener-Granger causality with the NINO3.4 index is considered, which could lead to the replacement of physically related variables.

Changes in manuscript

We will follow the suggestion to explicitly name the hyperparameters which have to be tuned for the model in the revised manuscript, and explain how these are tuned. This will done at the end of section 2.4. The hyperparameters which are named are correct and we will give an explanation of the tuning for these different hyperparameters in the revised manuscript.
In the revised manuscript, we will add the spread of hybrid models with different p of the ARIMA order, to show that the predictions do not vary much in this range of ARIMA orders.

5. 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.

Author's response

We show that the prediction is not very sensitive to the hyperparameters which are tuned on the data (the ANN structure and the ARIMA order). The test sets of Fig. 9 and 10 already provide some evidence that the model is not overfitting and the applied cross-validation method shows that the prediction model does not depend on different training and test sets. Nevertheless, we still cannot completely rule out overfitting outside the available data we have. Even if there is a chance the model is overfitting outside the available data we have, we think the proposed approach is still interesting for prediction of ENSO. Note that more studies about El Niño prediction have troubles with the shortness of the available time series [2] and overfitting will always be a possibility.

Changes in manuscript

We will include reference [2] in the discussion of the revised manuscript and explain it is difficult to rule out that the model is overfitting because of the short time series.

6. 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 is possible that they are intended to be averaged or used in some other boosting procedure, but if so, this is not adequately explained.

Author's response

The hybrid models at the different lead times are independent of each other. Part of the approach is that we tuned the model at specific lead times, to find which configuration is better for the memory contained in the attributes. That is also why we have different attributes at different lead times. This also means that, if we find more attributes via network analyses in future research which contain different length of memory, these attributes can be applied at the different lead times. This allows us to tune the hybrid model at different lead times.

Changes in manuscript

In the revised manuscript, we make clear that these hybrid models are tuned independently from each other and do not 'start at some $t = t_0$ and work forward step by step' (Sect. 2.4).

7. 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.

Author's response

The reason for the current structure of the paper is that it includes part of the process of how we got to the attributes applied in the hybrid model. We tried to find a physical reason for the variables to be included in the attribute set of the hybrid model, such that it increases the probability of a good prediction. To do this we looked at the dynamics of the ZC model and applied a network analyses to this model. We found some interesting attributes from this network analysis, but most of them were eventually not applied in the prediction model, because they did not behave similar when using observations. We understand this can be of confusion for the reader.

Changes in manuscript

As a solution, the results which are not used in the hybrid model (that is everything in section 2.3 and 3.1 which is not related to the attribute c_2 which is applied in the hybrid model) will be put in an appendix. Hopefully, this will establish a better connection between the results from the ZC model and the part about the hybrid model.

2 Specific comments

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

Author's response

The activation function used is the Sigmoid function.

Changes in manuscript

We will add this information in the revised manuscript.

2. 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?

Author's response

It is true that the hybrid model performs better than the CFSv2 ensemble mean at the shorter lead times, but we do not consider this to be the important result in the the table displayed in the figure. Up to six months ahead, the predictions are known to be quite good nowadays [3]. The most important result we find is that the twelve month lead prediction performs similar or even better than the shorter lead time predictions because of the attributes we chose and hence it is breaking the spring predictability barrier.

Changes in manuscript

In the revised manuscript we will put more emphasis on the important result that the twelve month lead prediction performs similar or even better than the shorter lead time predictions.

3. 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.

Author's response

An important purpose of the ZC-model is to explain the main dynamics which is associated with ENSO. This is used to find good attributes for the hybrid model. That is why the network analyses is first applied to the ZC model, resulting in a network variable c_2, which is eventually used in the hybrid model.
Noise is introduce as a way to model high-frequency atmospheric variability. The effect of adding the noise is explained on p. 4 of the manuscript:

> The effect of the noise on the model behavior depends on whether the model is in the super- or sub-critical regime (i.e whether μ above or below μ_c). If $\mu < \mu_c$, the noise excites the ENSO mode, causing irregular oscillations. In the supercritical regime, a cycle of approximately four years is present, and noise causes a larger amplitude of ENSO variability.

Hence the noise can excite the ENSO variability and can be an important factor for the prediction of ENSO. This leads to the reason for including the second principal component of the residual of the wind stress (PC2) in the attribute set (see p. 12).

Changes in manuscript

We make the purpose of the ZC model more clear in the revised manuscript.

4. 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.

Author's response

The Alpha Go project indeed made use of different type of machine learning.

Changes in manuscript

We will delete this citation and do not mention the project anymore in the revised manuscript.

5. 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.

Author's response

The statistic λ_2 in Appendix A2 is computed from the ZC model in section 2.3.

Changes in manuscript

We will remove Appendix A1 from the old manuscript as suggested.
Part of section 2.3 and 3.1 of the old manuscript is not used in the hybrid model. We will move these parts to the appendix. This means that the part of section 2.3 that will be moved to the appendix will become appendix A1, and the part of section 3.1 that wiil be moved to the appendix becomes appendix A2. As a consequence, the statistic λ_2 is explained in the same section as other Climate Network properties which are applied to the ZC model (but not used in the hybrid model). This makes the purpose of λ_2 more clear. We will make the explanation of how the variable λ_2 is calculated shorter, since it can also be found in [4].

References

[1] Christoph Bergmeir and José M. Benítez. On the use of cross-validation for time series predictor evaluation. *Inf. Sci. (Ny).*, 191:192–213, 2012.

[2] Wasyl Drosdowsky. Statistical prediction of ENSO (Nino 3) using sub-surface temperature data. *Geophys. Res. Lett.*, 33(3):10–13, 2006.

[3] L. Goddard, S. J. Mason, S. E. Zebiak, C. F. Ropelewski, R. Basher, and M. A. Cane. Current approaches to seasonal-to-interannual climate predictions. *Int. J. Climatol.*, 21(9):1111–1152, 2001.

[4] M.E.J. Newman. *Networks: An introduction*, volume 6. Oxford university press, Oxford, 2010.