We would like to thank the referee for his careful reading and his/her 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

Overall I think this is valuable and important work, but I think there could be more clarity in the writing. It tends to read as a long sequence of sentences rather than a narrative that walks the reader through the steps of the analysis.

Reply

Thank you for pointing this out. We think that the reason that the current structure could be somewhat confusing, is that a lot of network variables are explained in the beginning, and are not used anymore later in the hybrid model (except for one). We think the results from the network analyses of the ZC model are interesting.

Changes in manuscript

We will move the network variables which are eventually not used in the hybrid model to the appendix. As a result the paper will read more as ‘a narrative that walks the reader through the analyses.‘

At the end, I'm left slightly confused as to (i) How did you use the CZ model; did you actually learn something from that that helped analyze the real world,

Reply

The attributes which are applied in the hybrid model all represent a physical process. The first reason the ZC model is presented, is that it represents these physical processes which are important for prediction (e.g. the atmospheric noise that excites the ENSO mode is a reason to add PC2 in the attribute set). Second, we applied an analysis on the ZC model using Network Theory. This leads to multiple variables, of which one also showed interesting properties in observations and is applied in the attribute set.

Changes in manuscript

In the revised manuscript we will add an additional motivation to use the ZC model in section 2.2.

(ii) How you decided on the specific input variables (rather than what sounds like a jumbled mess of exploring a wide variety of different concepts that might have some relevance)

Reply

First the ZC model is used to investigate which variables could be interesting to apply from a physical point of view and the network analyses is applied on the model to find variables which contain useful information for prediction. Then the cross-correlation and Wiener-Granger causality are calculated at the different lags to see which of the variables we could apply at the different lead times. Finally, those variables are used in the hybrid model to see how good the prediction performed and it is tested if they are also robust at different training and test sets.

Changes in manuscript

In order to avoid the ‘jumbled mess,‘ we will move everything related to ZC model results which are not directly used as input in the ANN to the appendix of the revised manusscript.

(iii) To what extent your improvement in prediction is actually related to ML/ANN versus having identified good predictive variables (e.g., could you have identified a linear model that used those variables and obtained a good prediction? Were the ultimate relationships "learned" by the ANN between inputs and output actually notably nonlinear?)

Reply

ANN is known to be a good tool for prediction in nonlinear systems, such as the ENSO system. The ANN can recognise patterns which are important for prediction, which could be missed by the conventional statistical models. Hence the ANN can recognise nonlinear relations between the input variables and output variables, where a linear model might not.
This is a reason why we hypothesize that the ANN can be more useful for the prediction instead of an arbitrary linear model. However, it would be interesting if a linear model does exists which gives good results in combination with the attributes we applied in the hybrid model. We find there is a significant residual if the linear model ARIMA is applied and it is worth to improve this.

Changes in manuscript

We will write in the discussion of the revised manuscript a reason why the combination of the attributes and machine learning works well. Besides, we note in the discussion that a combination of attributes and a linear prediction model could be interesting.

(iv) It would help to have a single final plot showing rms error vs prediction horizon as compared with the current methods.

Reply

In the original manuscript we decided to only compare with the CFSv2 ensemble. We have thought of making a comparison with other conventional prediction methods such as in [1]. However, this requires that we know the rms error of the other prediction models for the same period or a subset of the period we have predictions for, since comparing the rmse obtained from predictions at different periods could be misleading.

Changes in manuscript

No changes will be made in the revised manuscript regarding this comment.

1. P2, 1st line, not quite sure how to define "intuition and creative thinking", nor (more importantly) why this is relevant here.

Reply

The Machine Learning method which competes with humans in the game GO is different.

Changes in manuscript

We will delete the whole sentence in the revised manuscript.

2. P2, par lines 3-11, this seems a bit awkwardly worded. It isn't a binary choice between many layers and inputs and "simpler", but rather a continuum of choices with an inherent trade-off. Using more layers and input variables means you can rely more on the algorithm to figure out what matters at the expense of needing to train it on more training data, and the fewer variables/layers one uses the less training data might be required but the more that forces the user to make intelligent choices for input variables rather than relying on the algorithm to do so.

Reply

We understand that the mentioned paragraph is confusing.

Changes in manuscript

We will rephrase the paragraph in the revised manuscript.

3. The choices in Section 2.3 are not well motivated (that is, why are these the relevant choices to

feed into the ANN, and what else did you try?) This section could benefit from a couple of introductory sentences that describe the goal of the section, and the broad overview of the ideas of the section.

Reply

Section 2.3 includes the methods applied in the network analyses. It resulted in some variables showing interesting properties of climate networks, but only one of them (c_2) is eventually applied in the ANN.

Changes in manuscript

In the revised manuscript all methods which are not applied in the prediction will be moved to the appendix. The new section 2.3 only presents the method to calculate c_2 and it should be clear now from this section why it could be useful for a prediction.

4. Why is it adequate to have all of the memory embedded in the linear part of the model?

Reply

To embed the memory only in the linear part of the model is a choice.
Two methods have been considered to include memory in the ANN. The first is the time delay neural network (TDNN), where also lags of attributes are used as input variable. The second is a recurrent neural network (RNN), where one allows loops in the neural network structure. We decided to stay with the feed-forward ANN, because the other two types of neural networks would only complicate the hyperparameter tuning (i.e. for the TDNN one has to decide which lags to implement and in the RNN the possible different structures increases), and embed all histroy in the linear part of the prediction model.
In future research both TDNN and RNN could be interesting to apply, however we got interesting results with only embedding the memory in the linear part.

Changes in manuscript

No changes will be made in the revised manuscript regarding this comment.

5. For that matter, not entirely obvious to me, since you are using ML to predict the nonlinear terms anyway, whether the ML can also predict the linear (but dynamic) part without any extra effort, or for that matter the nonlinear and dynamic part. Did you try different things and conclude you didn't have enough training data to converge, and kept simplifying, or did you just guess what might work and then it did? I didn't go back and read Hibon and Evgeniou, but it would seem like the question of how to simplify what the ML is actually learning is case dependent rather than absolute. Some more motivation here is required (and at a minimum you should clarify what is meant by "more stable" and provide a few more words of intuition as to why this reduces the risk of a bad prediction.)

Reply

We were looking for an easy method to implement the history in our prediction besides the feed-forward ANN, which resulted in ARIMA as easiest and most straightforward method. Using only the feed-forward ANN did not result in a good prediction.
'More stable' implies here that applying a combination of different types of prediction models, rather than only one type of prediction model, decreases the variability of the prediction skill when both are applied to several arbitrary time series.

Changes in manuscript

In the revised manuscript we will provide the motivation for the model choice. We also clarify what is meant by 'more stable' and why this reduces the risk of a bad prediction.

6. Extra plus sign in eqn 13 and 14. Also, shouldn't the summation on the second term start

at d+1 (otherwise, the j=1 in the second term and the i=1 in the first term are identical, and you have a standard ARMA model rather than an ARIMA model). (Also, don't recall if you said why you were using ARIMA rather than ARMA?)

Reply

Thank you for noticing this error. It is true that the differencing part is not incorporated well in this definition.

Changes in manuscript

We will use the definition similar to the definition in [2] in the revised manuscript, in order prevent any mistakes.

7. P7, L19-20, why would including past El Nino and La Nina information reduce prediction skill?

Reply

We hypothesize that the long-term memory, i.e. of previous La Niña and El Niño events, could contain information that is not relevant for the prediction of the coming year, because this information is not relevant anymore for the outcome in a chaotic system which is forced by high frequency noise.

Changes in manuscript

In the revised manuscript we will change the wording, such that the focus will be on the 'too long ago,' and not on the 'previous La Niña and El Niño events.'

8. P8, L1, I'd have just thought the choice of lead time is like a choice of different variables, that there's nothing wrong with including the same variable at different times as part of the input.

Reply

In this sentence we try to tell that at a specific lead time, one needs an optimal attribute set to optimize the prediction. This does not imply that an attribute cannot be used at several lead times.

Changes in manuscript

To prevent any misunderstanding the sentence will be rephrased to: 'Moreover, at every lead time an optimal attribute must be selected.'

9. P8, L17, "generally" as in, "in this paper", or "generally" as in "in most research"?

Reply

"Generally" as "in this paper" applies here.

Changes in manuscript

We will replace "Generally" by "in this paper" in the revised manuscript.

10. Section 3.1, any reason why you only used 45 years of ZC output? Why not use a few thousand years of output? (I ran it for that long quite a long time ago, so I know it isn't a computational challenge to do.)

Reply

We used only 45 years of data, because this comprises more then 10 ENSO cycles and this should be enough for the analyses we applied to the model. We recall that our main interest is to make predictions from the observational data, and in the observations we do not have much longer time series.'

Changes in manuscript

No changes will be made in the revised manuscript regarding this comment.

11. Also, section 3.1, you might want to say up front a bit more about motivation -are you trying to learn from ZC which variables are best to use, or ultimately comparing predictive capability on ZC vs the real world, or get a good initial estimate of ANN weights from ZC so that you don't have to converge as much when you apply to the real world? These are all possible goals, but other than the second one, may be problematic if the physics in ZC doesn't match the real world physics (and while with their original parameter choices the equilibrium point in ZC is unstable with a chaotic self-sustained response, I think the general consensus now is that the real world isn't exhibiting chaos but rather stochastically forced response of a damped stable system). This is similar to the comment on Section 2.3; it would be helpful to have a few additional sentences that talk about where you're going with a section, why is it here, what are you hoping to learn, and what the structure of the section is. (I note subsequently that you never actually look at the predictability of CZ model, improvement thereof with ANN, and you also don't use the same variables in the real world analysis. . . can you be clear as to why this section is here and what you learned? Is it here just because you spent a lot of time on it and figure that should be documented somewhere, or is it essential to motivate the analysis of the real world?)

Reply

It is true that the physics of the ZC model does not completely match the physics in the real world. However, we found a network variable in the ZC model which showed the same behaviour as in the observations (i.e. c_2).

Changes in manuscript

In the revised manuscript we will explain in the end of section 2.2 how the ZC model helped us to get to the finally applied attributes (as is explained in the beginning of this reply at comment (i)). Section 3.1 will change, since all network variables which are not applied in the prediction will be moved to the appendix, and it will be made made clear why the network variable that was used can be important for prediction.

12. P10, L2, I think what you mean here is something like "when the ENSO index changes from increasing to decreasing (peak El Nino) or from decreasing to increasing (peak La Nina)"? (The wording is a bit unclear to me.) Similarly line 7, refer to the derivative of the ENSO index, rather than the derivative of ENSO. . . (to me, "ENSO" refers to the overall dynamic phenomenon, which isn't a thing that has a sign or a derivative).

Reply

We understand the confusion.

Changes in manuscript

We will change the wording in the revised manuscript.

13. Section 4, rather than just focusing on a few things like 2010 (which is cherry-picked as a year where the default scheme does badly), and a few prediction horizons, one thing that would help evaluate this method would be a single plot of rms prediction error versus time for the two methods (that is, for any month once you have sufficient past data, do the N-month prediction for every N up to a year or more using both methods, and then over this big set of month N predictions, what's the rms error?) This would also be a great way to compare your ARIMA alone with ARIMA + ANN.

Reply

This figure was presented to show that the hybrid model can improve the other models drastically in a specific case. We agree that a figure showing the rmse at different lead time predictions could be nice to compare results.

However, this will require additional tuning at the different lead times. Besides, we will need the rmse of the other models in the same time interval at all these lead times, which we do not have.
The ARIMA prediction alone still had a very significant residual after the prediction.

Changes in manuscript

To compare the ARIMA only and ARIMA + ANN prediction, we will mention in the revised manuscript that this residual is very significant, which is a reason to add the ANN part in Sect. 4.

14. P14, L11, what do you mean by "best-performing"? What metric? Does that mean that adding more neurons made it worse? Or do you just mean that adding more neurons didn't make it better?

Reply

Here it means this ANN structure resulted in the lowest NRMSE from the ensemble of different ANN structures.

Changes in manuscript

This will be stated in the revised manuscript.

15. P15, L4, why compare the two methods at different lags instead of the same lag?

Reply

We compared two different lags here, because we only have the three month lead prediction instead of the four month lead prediction of the CFSv2 ensemble. In the hybrid model on the other hand, the attribute set resulted in a better result at the four month lead prediction compared to the three month lead hybrid model prediction, because the attribute set apparently contains better information four months ahead.

Changes in manuscript

No changes will be made in the revised manuscript regarding this comment.

16. P15, L7, doesn't this contradict the abstract?

Reply

We do not think this sentence contradicts the abstract.

Changes in manuscript

To be sure we are clear, however, we will change the wording of this sentence.

17. P15, L14, I'm confused by this sentence -you do a better job at predicting things 1 year in advance than 6 months?

Reply

That is true. This is possible because the ANN is trained at a specific lead time, say n months ahead. The ANN hence gives a function from the attributes to the output n months later. It is possible that the ANN trains better with the attributes at longer lead times. In this case, the attribute set also changed (i.e. the WWV is replacd by c_2), such that c_2 has this longer memory to improve the predictions longer ahead.

Changes in manuscript

No changes will be made in the revised manuscript regarding this comment.

18. Also, I must have missed something; I thought you'd already picked the set of input variables, and now it sounds like you are only using a subset, and a different subset for each prediction

horizon. Overall, this sounds incredibly fragile. You do a lot of work to pick a few really good input variables, and any time you change the time horizon you might need to change those, and change the number of neurons. . . I thought the whole point of ANN was the ability to be lazy and let the algorithm do all the work for you!

Reply

As explained in Sect. 3.2, we use cross-correlation and Wiener-Granger causality to determine the information of attributes at different lead times. Since the attribute set at the shorter lead times does not work well at larger lead times, we replace the WWV by c_2, which are physically related to each other.
In our method we cannot be just lazy. As explained in the introduction, we are not applying deep learning. In deep learning, where the ANNs are large in size, more attribute selection/reduction is done by the algorithm itself. The smaller the ANNs that are used, the more effort has to be done in the attribute selection. We apply this method instead of deep learning because the observational time series are relatively short. Deep learning is only known to work well if a lot of data is available.

Changes in manuscript

No changes will be made in the revised manuscript regarding this comment.

19. P15, L15-16, again, I'm a bit confused. . . why do we need to maintain a whole ensemble of different ANN structures? This doesn't converge to something with enough neurons? Also, Figure 11, am I interpreting this right that you found a bunch of possible ANN structures that outperform the ones in Figure 9? (Sorry, I'm totally lost at this point so this might be off-base and simply imply some insufficient description.) Why not go back and redo Fig 9 with the better ANN structure? This entire section reads a bit as a collection of odds and ends of results rather than as a post-facto summary.

Reply

The purpose of considering an ensemble of different ANN structures in Fig. 10 is that it shows the outcome is not sensitive to the specific ANN choice. This implies that the prediction shown in Fig. 9 was not a lucky shot, but more ANN structures converge to similar predictions.
In Fig. 11, we perform the cross-validation for the models of Fig. 9 for different training and test sets. This means we apply exactly the same attributes and ANN structures as in Fig. 9. We find that the models from Fig. 9 can perform better if different parts of the total time series are used as test and training set.
The section is meant to give the final prediction results, followed by a generalisation and validation of these results.

Changes in manuscript

In the revised manuscript, we will make sure the purpose of this section is made clear in the beginning of section 4. Besides, all comments/questions regarding this section will be addressed.

References

[1] Anthony G. Barnston, Michael K. Tippett, Michelle L. L'Heureux, Shuhua Li, and David G. Dewitt. Skill of real-time seasonal ENSO model predictions during 2002-11: Is our capability increasing? *Bull. Am. Meteorol. Soc.*, 93(5), 2012.

[2] Yi Shian Lee and Lee Ing Tong. Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming. *Knowledge-Based Syst.*, 24(1):66–72, 2011.