Dear Editor,

Please find enclosed our revised manuscript. We have amended the manuscript to address the issues raised by both reviewers.

This document first includes a point-by-point reply to both referees, including the important changes in the revised manuscript. Finally, one can find a version of the manuscript where the differences with the previous version are highlighted (using latexdiff tool).

Thank you for considering our manuscript for publication. We are indebted to the two reviewers for their constructive comments. We appreciate your time and look forward to hearing from you.

Yours faithfully,

Peter Nooteboom
on behalf of all the authors

Response to Robert Link

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 will 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.

Response to Referee #2

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 beginning 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.

Using Network Theory and Machine Learning to predict El Niño

Peter D. Nooteboom[1,3], Qing Yi Feng[1,3], Cristóbal López[2], Emilio Hernández-García[2], and Henk A. Dijkstra[1,3]

[1]Institute for Marine and Atmospheric Research Utrecht (IMAU), Department of Physics, Utrecht University, The Netherlands
[2]Instituto de Física Interdisciplinar y Sistemas Complejos (IFISC, CSIC-UIB), University of the Balearic Islands, Spain
[3]Centre for Complex Systems Studies, Utrecht University, The Netherlands

Correspondence to: Peter Nooteboom (p.d.nooteboom@uu.nl)

Abstract. The skill of current predictions of the warm phase of the El Niño Southern Oscillation (ENSO) reduces significantly beyond a lag of six months. In this paper, we aim to increase this prediction skill at lags up to one year. The new method to do so combines a classical Autoregressive Integrated Moving Average technique with a modern machine learning approach (through an Artificial Neural Network). The attributes in such a neural network are derived from knowledge of physical processes and topological properties of Climate Networks~~and are tested on both~~, and they are tested using a Zebiak–Cane-type model and observations. For predictions up to six months ahead, the results of the hybrid model give a slightly better skill than the CFSv2 ensemble prediction by the National Centers for Environmental Prediction (NCEP). ~~Moreover~~Interestingly, results for a twelve-month lead time prediction have a similar skill as the shorter lead time predictions.

1 Introduction

Approximately every four years, the sea surface temperature (SST) is higher than average in the eastern equatorial Pacific (Philander, 1990). This phenomenon is called an El Niño and is caused by a large-scale ocean-atmosphere interaction between the equatorial Pacific and the global atmosphere (Bjerknes, 1969), referred to as El Niño/Southern Oscillation (ENSO). It is the dominant mode of climate variability at interannual time scales and has teleconnections worldwide. As El Niño events cause enormous damage worldwide, skillful predictions, preferable for lead times up to one year, are highly desired.

So far, both statistical and dynamical models are used to predict ENSO (Chen et al., 2004; Yeh et al., 2009; Fedorov et al., 2003). However, El Niño events are not predicted well enough up to six months ahead due to the existence of the so-called predictability barrier (Goddard et al., 2001). Some theories indicate this is due to the chaotic, yet deterministic, behavior of the coupled atmosphere-ocean system (Jin et al., 1994; Tziperman et al., 1994). Others point out the importance of atmospheric noise, acting as a high frequency forcing sustaining a damped oscillation (Moore and Kleeman, 1999).

Recently, attempts have been made to improve the ENSO prediction skill beyond this spring-predictability boundary, for example by using machine learning (ML) (Wu et al., 2006) methods, also combined with network techniques (Feng et al., 2016). ML has shown to be a promising tool in other branches of physics, outperforming conventional methods (Hush, 2017). ~~In addition, ML did a better job than humans in a game of GO, which is difficult for Artificial Intelligence (AI) since it requires~~

~~intuition and creative thinking (Silver et al., 2016).~~ As the amount of data in the climate sciences is increasing, ML methods such as Artificial Neural Networks (ANN), are becoming more interesting to apply to prediction studies.

Briefly, ANN is a system of linked neurons that describes, after optimization, a function from one or more input variables (or attributes) to the output variable(s). Generally, ~~two different approaches can be considered when applying the ANN . The first is to use a complicated ANN structure with a lot of layers in the network and many input variables (or attributes). This approach is situated on the deep learning part of the ML spectrum and is believed to~~ one has to choose how large and complicated the ANN structure is. The more complicated an ANN, the more it will filter the important information from the attributes itself~~. The deep learning approach requires a lot of~~ , but it will require more input data and is computationally intensive. Therefore, simpler ANN structures are used in this article. However, techniques will have to be applied in order to reduce the amount of input variables and select the important ones, to make the problem appropriate for the simpler ANN. This reduction and selection problem can be tackled in many ways, which are crucial for the prediction. The main issue in these methods is, however, what attributes to use for ENSO prediction.

Complex networks turn out to be an efficient way to represent spatio-temporal information in climate systems (Tsonis et al., 2006; Steinhaeuser et al., 2012; Fountalis et al., 2015) and can be used as an attribute reduction technique. These Climate Networks ~~(CN)~~ are in general constructed by linking spatio-temporal locations that are significantly correlated with each other according to some measure. It has been demonstrated that relationships exist between topological properties of ~~CNs~~ Climate Networks and nontrivial properties of the underlying dynamical system ~~(Tupikina et al., 2014; Deza et al., 2014; Stolbova et al., 2014)~~(Dez also specifically for ENSO (Gozolchiani et al., 2011, 2008; Wang et al., 2015). ~~CNs~~ Climate Networks already appeared to be a useful tool for more qualitative ENSO prediction, by considering a warning of the onset of El Niño when a certain network property exceeds some critical value ~~(Ludescher et al., 2014; ?; Rodríguez-Méndez et al., 2016)~~(Ludescher et al., 2014; Meng et al., 2017; Rod

In this paper, a hybrid model is introduced for ENSO prediction. The model combines the classical linear statistical method of Autoregressive Integrated Moving Average (ARIMA) and an ANN method. ANN is applied to predict the residual, due to the nonlinear processes, that is left after the ARIMA forecast (Wu et al., 2006). To motivate our choice for attributes in the ANN, we use an intermediate complexity model which can adequately simulate ENSO behavior, the Zebiak-Cane (ZC) model (Zebiak and Cane, 1987). ~~Network variables are chosen as attributes ,~~ The attributes which are used in the prediction model are related to physical processes which are relevant for ENSO prediction. Moreover, network variables are considered as attributes such that they ~~are related to another physical quantity capturing information about the system, but spatial information is conserved.~~ relate to a physical mechanism, but additionally contain spatial information.

Section 2 briefly describes the ZC model, the methods considering both the ~~CNs~~ Climate Networks and ML and the used data from observations. In Sect. 3, the network methods are first applied to the ZC model. Second, the attributes selected for observations are presented. These attributes, among which there is a network variable, are applied in the hybrid prediction model in Sect. 4, which discusses the skill of this model to predict El Niño. The paper concludes with a summary and discussion in Sect. 5.

2 Observational data, models and methods

2.1 Data from observations

As observational data, we use the sea surface height (SSH) from the weekly ORAP5.0 (Ocean ReAnalysis Pilot 5.0) reanalysed dataset of ECMWF from 1979 to 2014 between 140°E to 280°E and 20°S to 20°N.

For recent predictions, the SSALTO/DUACS altimeter products are used for the same spatial domain, since the SSH is available from 1993 up to present in this dataset. The SSALTO/DUACS altimeter products were produced and distributed by the Copernicus Marine and Environment Monitoring Service (~~CMEMS) (~~http://www.marine.copernicus.eu).

In addition, the HadISST dataset of the Hadley center has been used for the SST and the NCEP/NCAR Reanalysis dataset for the wind stress from 1980 to present (Rayner et al., 2003).

To quantify ENSO, the NINO3.4 index is used, i.e., the three-month running mean of the average SST anomaly in the extended reconstructed sea surface temperature ~~(ERSST) ~~dataset between 170°W to 120°W and 5°S and 5°N (Huang et al., 2015).

The warm water volume (WWV), being the integrated volume above the $20°C$ isotherm between $5°N$-$5°S$ and $120°E$-$280°E$, is determined from the temperature analyses of the Bureau National Operations Centre~~(BNOC) (~~ (https://www.pmel.noaa.gov/elnino/upper-ocean-heat-content-and-enso).

2.2 The Zebiak-Cane model

The ZC model (Zebiak and Cane, 1987) represents the coupled ocean-atmosphere system on an equatorial β-plane in the equatorial Pacific (see Fig. 1). This model is used here to infer which processes are important for ENSO prediction and to find the attributes which represent those processes. Also, a network analyses is applied to the ZC model in order to find network variables which could improve prediction, before these network variables are calculated in observations. We use the numerically implicit version of this model (van der Vaart et al., 2000; Von Der Heydt et al., 2011) as in Feng (2015).

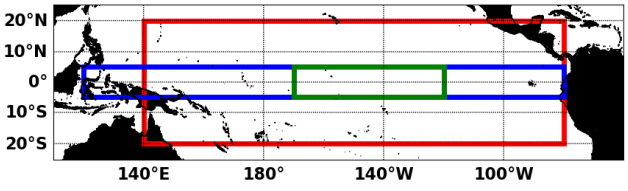


Figure 1. Pacific area (red rectangle) from $140-280°$E and $-20-20°$N , the NINO3.4 area (green rectangle) from $170-120°$W and $-5°$S$-5°$N and the WWV area (blue rectangle) from $120-280°$E and $-5°$S$-5°$N.

In the ZC model, a shallow-water ocean component is coupled to a steady shallow-water Gill ~~(Gill, 1980) atmosphere model~~ atmosphere model (Gill, 1980). The atmosphere is driven by heat fluxes from the ocean, depending linearly on the anomaly of the sea surface temperature T with respect to a radiative equilibrium temperature T_0. The zonal wind stress τ^x is the sum of a

coupled and an external part~~:~~

$\tau^x = \tau^x_{ext} + \tau^x_c$.

~~.~~ The external part ~~τ^x_{ext}~~ is independent of the coupling between the atmosphere and ocean ~~. It~~ and represents a weak easterly wind stress due to the Hadley circulation. ~~It is assumed to be zonally constant and depends on latitude according to:~~

$\tau^x_{ext} = -\tau_0 e^{-\frac{1}{2}\left(\frac{y}{L_a}\right)}$.

~~Here $\tau_0 \sim 0.01\ Pa$, L_a the atmospheric Rossby deformation radius and y is the meridional coordinate.~~ The coupled part of the zonal wind stress ~~τ^x_c~~ is proportional to the zonal wind from the atmospheric model; the meridional component of the wind stress is neglected in this model.

As shown in van der Vaart et al. (2000), the parameter measuring the magnitude of the ocean-atmosphere coupled processes is the coupling strength μ. Without any included noise, a temperature anomaly damps out to a constant value and a stationary state if $\mu < \mu_c$, where μ_c indicates a critical value. However, if the coupling strength exceeds the critical value μ_c, a supercritical Hopf bifurcation occurs. A ~~perturbations~~ perturbation then does not decay, but an oscillation is sustained with a period of approximately four years.

Three positive feedbacks related to the thermocline depth, upwelling and zonal advection, can cause the amplification of SST anomalies (Dijkstra, 2006) while the oscillatory behavior associated with ENSO is caused by negative delayed feedbacks. The 'classical delayed oscillator' paradigm assumes this negative feedback is caused by waves through geostrophic adjustment, controlling the thermocline depth. A complementary, different view is the 'recharge/discharge oscillator' (Jin, 1997), also regarding oceanic waves excited through oceanic adjustment. The waves excited to preserve the Sverdrup balance are responsible for a transport of warm surface water to higher latitudes, discharging the warm water in the tropical Pacific. The thermocline depth is raised, resulting in more cooling of SST. The warm water volume (WWV) is the variable generally used to capture how much the tropical Pacific is 'charged.'

Apart from the coupled ocean-atmosphere processes, ENSO is also affected by fast processes in the atmosphere, which are considered as noise in the ZC model. An important example of atmospheric noise are the so-called westerly wind bursts (WWB). These are related to the Madden-Julian oscillation (Madden and Julian, 1994). The WWB is a strong westerly anomaly in the zonal wind field, occurring every forty to fifty days and lasting approximately a week. 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.

~~To represent the~~ The atmospheric noise in the model ~~, we obtained~~ is represented by obtaining a residual of the wind stress from observations ~~. We used the ERSST over the Pacific for the period 1978-2004 and the Florida State University pseudo-wind-stress data (Legler and Brien, 1988) for the same period. First the part of the wind stress anomalies linearly related to the SST anomalies are subtracted from the wind stress anomalies. Then the residual is projected on its Empirical Orthogonal Functions (EOFs). The six EOFs that explain most variability were considered to describe the spatial patterns of~~

the noise and their Principal Components (PC) were used to construct the time series (Feng and Dijkstra, 2016). These time series are modeled as an independent first order Auto Regressive (AR(1)) process by: as in Feng and Dijkstra (2016).

$$x_{t+1} = ax_t + \sigma\epsilon_t,$$

where a is the lag-1 autocorrelation of each PC and $\sigma\epsilon_t$ is the white noise with variance σ. Since weekly data are considered, every discrete time step in the model is one week.

2.3 Network variables

Different methods can be used to construct CNs. Here only the Pearson correlation (appendix **??**) will be considered for Here we explain the methods to calculate a property of a Climate Network which is tested in the ZC model and observations and will be used in the hybrid model. From the network analysis we found several Climate Network quantities with interesting properties for prediction, but which are not used in the hybrid model of the construction of two different types of CN. In the first method an next section. The methods to calculate these properties can be found in Appendix A1.

An undirected and unweighted network is constructed . making use of the Pearson correlation of climate variables related to ENSO (e.g. SST, thermocline depth or zonal wind stress). Network nodes are model or observation grid positions i and the links are stored in a symmetric adjacency matrix A, where $A_{ij} = 1$ if node i is connected to node j and $A_{ij} = 0$ otherwise. A_{ij} is defined by:

$$A_{ij} = \Theta\left(|R_{ij}| - \epsilon\right) - \delta_{ij}. \tag{1}$$

Here R_{ij} is the Pearson correlation between node i and j, ϵ is the threshold value and Θ denotes the Heaviside function. Hence, if the Pearson correlation exceeds the threshold ϵ, the two nodes will be linked. The δ_{ij} is the Kronecker delta function, implemented to prevent connection of nodes with themselves.

The second method creates a weighted, undirected network and will be applied for only one network variable later. The cross-correlation $C_{ij}(\Delta t)$ at lag Δt, i.e. the Pearson correlation between the variables $p_i(t)$ and $p_j(t + \Delta t)$ is considered. Then the weights between the nodes are calculated by:

$$W_{ij} = \frac{\max_{\Delta t}(C_{ij}) - \text{mean}(C_{ij})}{\text{std}(C_{ij})}.$$

Here $\max_{\Delta t}$ denotes the maximum, std the standard deviation and mean the mean value over all time steps that are considered.

From the CNs, we construct several properties. From the unweighted network we compute the local degree d_i of node i in the CN as,

$$d_i = \sum_j A_{ij},$$

~~i.e. degree is equal to the amount of nodes that are connected to node i. The spatial symmetry of the degree distribution is of interest, since it informs where most links of the network are located. More specifically, our interest will be in the symmetry in the zonal direction in a network. Therefore, the skewness of the meridional mean of the degree in the network is calculated. This defines the zonal skewness of the degree distribution in a network.~~

~~Second, percolation theory is~~ Percolation theory is then considered, describing the connectivity of different clusters in a network. It has been found that the connectivity of some ~~CN increases when approaching~~ Climate Networks increases just before an El Niño and decreases afterwards (Rodríguez-Méndez et al., 2016), as local correlations between points increase and decrease. At such a percolation-like transition, the addition of only a few links can cause a considerable part of the network to become connected. Before the percolation transition, clusters of small sizes will form. Therefore the variable c_s will warn for the transition:

$$c_s = \frac{sn_s}{N}. \tag{2}$$

Here n_s is the amount of clusters of size s and N the size (i.e. the total amount of nodes) of the network. Thus c_s is the fraction of nodes that are part of a cluster of (generally small) size s. ~~For another variable related to the percolation-like transitions, links are added to a network one by one, adding the link with the largest weight first (Eq. (A3)). At every step T that a link is added, the size of the largest cluster $S_1(T)$ is calculated. At the point of the percolation transition, $S_1(T)$ increases rapidly. The size of this jump is Δ:~~

$$\Delta = \max\left[S_1(2) - S_1(1), \ldots, S_1(T+1) - S_1(T), \ldots\right].$$

~~The quantity Δ can be used to capture the percolation-like transition (**?**).~~

~~The final two CN properties are derived from a so-called NetOfNet approach. This is a network constructed with the same methods as previously, but using multiple variables at each grid point (as specified later in the results). This gives a network consisting of the networks from the different variables interacting with each other. Only NetOfNet of two different variables are considered. First, the cross clustering contains information about the interaction between two unweighted networks. The local cross clustering of a node is the probability that two connected nodes in the other network are also connected to each other. The global cross clustering C_{vw} is the average over all nodes in subnetwork G_v of the cross clustering between G_v and G_w:~~

$$C_{vw} = \frac{1}{N_v} \sum_r \frac{1}{k_r\,(k_r - 1)} \sum_{p \neq q} A_{rp} A_{pq} A_{qr}.$$

~~Here r is a node in subnetwork G_v of size N_v, p and q are the nodes in the other subnetwork G_w and k_r denotes the cross degree of node r (i.e. amount of cross links node r has with the other subnetwork). In addition to the clustering coefficient, also the algebraic connectivity (λ_2) is considered (appendix **??**) within a NetOfNet. This variable includes the diffusion of information within one or multiple networks.~~

2.4 Hybrid prediction model

A hybrid model (Valenzuela et al., 2008) will be applied to predict ENSO, in which the observation Z_t at time t is represented by

$$Z_t = Y_t + N_t. \tag{3}$$

Here Y_t is modelled by a linear process and N_t by a ML type technique. Let $\tilde{Y}_t$ be the prediction of the part Y_t using ARIMA, then $Z_t - \tilde{Y}_t$ is the residual with respect to the observed value. This residual will be predicted by the feedforward ANN:

$$\tilde{N}_t = f\left(x_1(t), \cdots, x_N(t)\right). \tag{4}$$

Here f is a nonlinear function of the N attributes ~~$x_1(t), \cdots, x_n(t)$~~ $x_1(t), \cdots, x_N(t)$ and $\tilde{N}_t$ the prediction of residual $Z_t - \tilde{Y}_t$ at time t. Notice the nonlinear function f does not depend on history, whereas the ARIMA part $\tilde{Y}$ does. The final prediction $\tilde{Z}_t$ of the hybrid model:

$$\tilde{Z}_t = \tilde{Y}_t + \tilde{N}_t. \tag{5}$$

Previous work showed that the results of a hybrid model are in general more stable and reduce the risk of a bad prediction, compared to a single prediction method (Hibon and Evgeniou, 2005). 'More stable' means that a hybrid model has a lower variability of prediction skill for different arbitrary time series. Besides, ARIMA is a simple method to include information about the history in the prediction model, which is not in the feed-forward ANN.

This scheme describes a 'supervised' model, implying that the predictant is 'known.' This known quantity is the NINO3.4 index. The standard procedure for supervised learning is to optimize the ML method on a 'training set' to define an optimal model, which predicts ENSO with a certain time ahead. This function will then be tested on a test set. Here a training set of 80 % and a test set of 20 % of the total time series is used. The data set can be represented by a $T \times N$ matrix, where T represents the length of the time series and each time $t = 1, \ldots, T$ has a set of N attributes $x_1(t), \ldots, x_N(t)$. 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~~ is convenient (where $t_i^{train}, t_f^{train}, t_i^{test}, t_f^{test} \in [1, T]$). In the following, we describe more in detail the different parts of this hybrid prediction method.

First, the training set is used to optimize an ARIMA(p,d,q) process for the NINO3.4 time series. The standard method maximizing the log likelihood function is used to fit $\alpha_1, \cdots, \alpha_p, \beta_1, \cdots, \beta_q$, such that $\sum_t \varepsilon_t^2$ is minimized for time series Z_t with t in months:

$$\left(1 - \alpha_1 B - \ldots - \alpha_p B^p\right)\left(1 - B\right)^d Z_t = \sum_{i=1}^{d} Z_{t-i} + \left(1 + \sum_{j=1}^{p} \beta_1 B + \alpha_j Z_{t-j} \ldots + \sum_{k=1}^{q} \beta_{kq} B^q\right) \varepsilon_{t-k} + \varepsilon_t, \tag{6}$$

where ε_t is the residual, differencing order d determines the amount of differencing terms, p the amount of ~~AR~~ autoregressive terms and q the amount of ~~MA~~ moving average terms on the right hand side, B ($BZ_t = Z_{t-1}$) is the lag operator. Finding the

most optimal ARIMA order (p, d, q) is not trivial (Zhang, 2003; Aladag et al., 2009). General methods include the Akaike's information criterion (Akaike, 1974) or minimum description length (Rissanen, 1978). However, these methods are often not satisfactory and additional methods have been proposed to determine the order (Al-Smadi and Al-Zaben, 2005). In this article we mainly present results obtained with orders $p = 12$, $d = 1$ and $q = 0$ or $q = 1$, which ~~resulted in good prediction results. Besides, this order avoids including information of past El Niño and La Niña events, which could possibly reduce the prediction skill~~gave good prediction skill and it can be argued that in such a chaotic system, information from too long ago is not important anymore.

The eventual ARIMA ~~prediction $\hat{Y}_t$~~ equation results in a prediction $\hat{Y}_t(Z_{t-1}, \ldots, Z_{t-p}, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q})$ of $\tau = 1$ months ahead~~is~~:

$$\hat{Y}_t = \sum_{i=1}^{d} Z_{t-i} + \sum_{j=1}^{p} + \alpha_j Z_{t-j} + \sum_{k=1}^{q} \beta_k \varepsilon_{t-k}.$$

Here $\varepsilon_{t-1} = Z_{t-1} - \hat{Y}_{t-1}$. Let $\tilde{Y}_t$ be the ARIMA prediction of $\tau > 0$ months ahead, by ~~applying Eq. ??~~ calculating $\hat{Y}_t$ for τ times in the future and replacing any observation Z_t with the consecutive calculated $\hat{Y}_t$, ~~if~~ where t is in the future and Z_t therefore unknown. Similarly, if $q = 1$ and $\tau > 1$ months, the residual is calculated by $\epsilon_{t-1} = \tilde{Y}_{t-1} - \hat{Y}_{t-1}$, since the observed value Z_{t-1} is in the future. Hence the ARIMA prediction $\tilde{Y}_t$ will be a time extrapolation with the optimized ARIMA model.

After $\tilde{Y}_t$ is predicted by the ARIMA model, the ANN will be used for the prediction $\tilde{N}_t$, making use of more variables than the NINO3.4 index alone. Deciding which of the variables to use is not a straightforward problem, yet crucial for the eventual prediction. ~~Sometimes~~ Generally in an ANN, a pair of two variables can be compatible in the prediction, but perform poor when applied alone. Other pairs can be redundant and cover important information when used alone, but solely noise is included when used together (Guyon and Elisseeff, 2003). Adding a variable to the attribute set and see if it improves prediction, can only conclude whether it improves prediction with respect to the old attribute set, not whether the variable is predictive in itself. To determine the attribute set, we consider which variables represent a certain physical mechanism that is important for the ENSO prediction. This helps to find attributes which are not related to each other, but include important information on their own. Besides, it is tested whether the prediction skill is reduced if a variable is dropped out of the attribute set.

Moreover, ~~the attributes should be selected at optimal lead times~~at every lead time an optimal attribute must be selected. Hence the final prediction model is tuned for a specific lead time and will not be a step by step prediction forward in time. Apart from considering the physical mechanisms the variables represent, two methods will help to decide which variables can improve the prediction. First, correlation between the predictor and predictant is a commonly used measure for attribute selection (Hall, 1999). Therefore the Pearson cross-correlation is calculated for the attributes at lag τ to show the predictability of a time series:

$$R_\tau(p, q) = \max_\tau \left(\frac{\sum_{k=1}^{n} p(t_k) q(t_k - \tau)}{\sqrt{\left(\sum_{k=1}^{n} p^2(t_k)\right)\left(\sum_{k=1}^{n} q^2(t_k - \tau)\right)}} \right). \quad (7)$$

Here p is the predictor, q is the predictant and lag $\tau \leq 64$ weeks such that no information too far in the past is considered.

However, the effect of a variable on ENSO at a short lead time increases the cross-correlation at a longer lead time, due to the effect of autocorrelation (Runge, 2014). To solve this autocorrelation problem, a Wiener-Granger causality F-test (Sun et al., 2014) is performed between all predictors $x_1, \cdots, x_N$ and the predictant at lags τ. Note Granger causality is not the same as a 'true' causality. If the test results in a low p-value, the null hypothesis that x_i does not Granger cause the predictant is rejected at a low significance level (i.e. x_i is more likely to Granger cause the predictant). Notice both the cross-correlation and Wiener-Granger method give us merely an idea of which variables can be used for the prediction at different lead times. Both methods are linear, while the attributes will be used in a nonlinear method.

Finally, the $T \times N$ dataset with selected attributes is used to predict the residual between the ARIMA forecast and the observations in an ANN. Besides using the NINO3.4 sequence itself, the additional attributes can be applied to add important information and improve the prediction.

~~Generally~~In this paper, only a feed-forward ANN is applied, having a structure without loops. The input variables are linearly combined and projected to the first layer neurons according to (Bishop, 2006):

$$z_j = h\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right). \tag{8}$$

Here z_j is the value of the j-th neuron of the layer, $w_{ji}^{(1)}$ is the weight between input x_i from neuron i to neuron j, where the (1) denotes the first layer. $w_{j0}^{(1)}$ is referred to as the bias. h is the ~~so-called (nonlinear)~~ sigmoid activation function, essential for incorporating the nonlinearity in the prediction model.

These z_j can again be used as input for a second layer, which can be used for a third layer etc.. Eventually this leads to some output which can be compared with the time series that must be predicted. Using a backward propagating technique, the squared error $\sum_t (y_t - \hat{y}_t)^2$ between the residual we are predicting y_t and the output of the ANN $\hat{y}_t$, will be minimized over the weights for the training set. The optimized function can then be tested on the test set. Initially, some random distribution of weights is used. The ANN part of the prediction will be performed with the toolbox ClimateLearn (Feng et al., 2016).

To summarize the tuning of the hybrid model: the ARIMA order and the hyperparameters controlling the ANN structure are tuned on the data, i.e. such that the prediction result is optimal. However, we will consider whether some set of different parameter values converges to similar predictions, which can show whether the hyperparameter tuning was a one lucky shot or not. The choice of the attributes is based on the ZC-model giving a more physical basis for the information needed for a good prediction. To select them at a specific lag also their cross-correlation and Wiener-Granger causality with the ENSO index and performance are considered, which could lead to the replacement of an attribute with another attribute which is physically related.

3 Analysis of network properties and selection of ML attributes

In this section, topological properties of ~~CNs~~ Climate Networks are analysed within the ZC model and observations, which ~~leads~~ lead to specific choices of attributes in the hybrid prediction model.

3.1 Network variables from the ZC model~~results~~

Weekly spatial-temporal data on a 31×30 grid in the Pacific region are obtained for forty-five years from the ZC model, to construct the ~~CNs~~Climate Networks. The first five years are not considered, to discard the effect of the initial conditions. A sliding window approach is used to calculate the network variables. This implies that a different network is calculated at each time, which is sliding four weeks ahead every time step. For the ZC model, either the thermocline network (from h), SST network (from T), ~~wind~~ wind-stress network (from τ^x) or a combination of these are considered for ~~CN construction. Multiple network variables are presented here, containing interesting information about the ZC model, although only one will eventually be used~~ network construction. Only the network variable which showed the same behaviour in the observations ~~due to its predictive power.~~

~~Determining how strong noise can excite the ENSO mechanisms in the sub-critical case, or determining whether the feedbacks sustain an oscillation in the supercritical state, could provide information to increase the prediction skill. Feng (2015) found that the skewness of the degree distribution S_d of the CN reconstructed from SST decreases monotonically with increasing coupling strength μ. Although S_d relates to the climate stability and coupling strength, it does not inform whether the system is in either the supercritical or sub-critical state. Global cross clustering between the SST and wind network in blue and its variance in green in the ZC model. The coupling strength μ defined as a sinusoid around $\mu_c = 3$ with an amplitude of 0.25 in red. The sliding window is applied with a window of five years. Here, we introduce a NetOfNet variable which may represent properties of the stability of the background state: the global cross clustering (C_{vw}) between the SST and wind network. A sliding window of five years with $\epsilon = 0.6$ was used to compute the networks. In this case, the global cross clustering coefficient is a measure of the amount of triangles in the networks, containing one wind node and two SST nodes. In Fig. A1, this cross clustering is calculated from data from the ZC model , when coupling strength μ changes periodically in time around the critical value $\mu_c \sim 3.0$. Under sub-critical conditions, the noise has a larger influence on local correlations. This causes triangles to break and the variance of the cross clustering coefficient to increase. The cross clustering C_{vw} is hence a diagnostic network variable which informs whether the state of the system is in the supercritical or sub-critical regime.~~

~~Zonal skewness of the degree field of the thermocline network with $\epsilon = 0.6$ and a sliding window of one year in red, NINO3.4 index in black in the ZC model. (a) The sub-critical ($\mu = 2.7$) and (b) the supercritical ($\mu = 3.25$) case. From the classical view of the oscillatory behavior of ENSO, waves in the thermocline should contain memory of the system, because of their negative delayed feedback. The changing structure of the thermocline network is therefore of interest when predicting ENSO. Calculating this network with threshold $\epsilon = 0.6$ and a sliding window with a length of one year, a zonal pattern in the change of the network close to the equator can be observed during an ENSO cycle. To compare network structures in the super- and sub-critical state, now constant $\mu = 2.7$ (sub-critical) and $\mu = 3.25$ (supercritical) are taken. Generally, the degree field is quite spatially symmetric, but when ENSO turns either from upward to downward, or from downward to upward, the degree of the nodes in the east decreases. This is at the peak El Niño or La Niña.~~ in the ZC model is presented here. Other network variables with interesting properties can be found in Appendix A2.

~~To capture this zonal asymmetry around the equator with a variable, the zonal skewness of the degree field will be used between 7°S to 7°N. The higher the skewness, the more the degree will be located west of the basin. If the skewness is close to zero, the degree is symmetrically distributed over the basin. If it is low, most of the degree is situated in the east. The skewness will show a negative peak when the sign of the first ENSO derivative changes (Fig. A2). In the supercritical case $\mu = 3.25$ this effect is indeed observed. Nevertheless, in the sub-critical case, the pattern is only visible once ENSO shows a clear oscillation (around year 32).~~

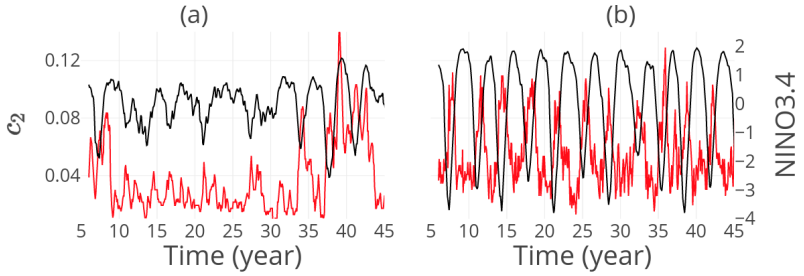


Figure 2. The network variable c_2 of the thermocline network with a sliding window of one year in red and NINO3.4 in black in the ZC model. (a) The sub-critical ($\mu = 2.7$) case with threshold $\epsilon = 0.99999$ and (b) the supercritical ($\mu = 3.25$) case with $\epsilon = 0.999$.

~~For the ZC model,~~ The network variable of interest is c_2 (the proportion of nodes belonging to clusters of size two) of the thermocline network~~is found to indicate~~ , because it indicates the approach to a percolation transition of the network during an El Niño event (Fig. 2). ~~Again a~~ A window of one year is used. c_2 increases approximately one to two years before an El Niño event. This is mainly clear in the supercritical case. In the sub-critical case, a clear warning of an event occurs when the oscillation of ENSO is more clear and the El Niños are stronger. Because c_2 is ~~the only network variable which will not only be applied~~ a warning signal of an El Niño event in the ZC model, ~~but also in the observations later. The quantity Δ of the same network behaves similar to c_2. Although Δ does not depend on a chosen threshold like c_2, it peaks closer to an El Niño event.~~ we will look in the next section how it behaves when it is calculated from observations.

~~Finally, the algebraic connectivity (λ_2) can show the spread of information within a network. Specifically, when considering an unweighted NetOfNet from thermocline depth (h) and zonal wind (τ^x) with threshold $\epsilon = 0.6$. The spread of information is relatively high before an event, but also after an event, such that λ_2 peaks both before and after an El Niño event (both for $\mu = 2.7$ and $\mu = 3.25$).~~

3.2 Selecting attributes from observations

The ZC model results have given an indication of the network variables that could be used as attributes in the hybrid model to predict El Niño. Although the network variables show interesting behavior in the ZC model for prediction, this is not always the case in observations. This section describes which variables, including a network variable, are implemented in the hybrid model and the selection of these attributes at different lead times. Notice that only anomalies of the time series in observations are considered.

First, from the recharge/discharge oscillator point of view, the WWV shows great potential for the prediction of ENSO (Bosc and Delcroix, 2008; Bunge and Clarke, 2014). Therefore it is used in the attribute set. The second attribute is a WWV

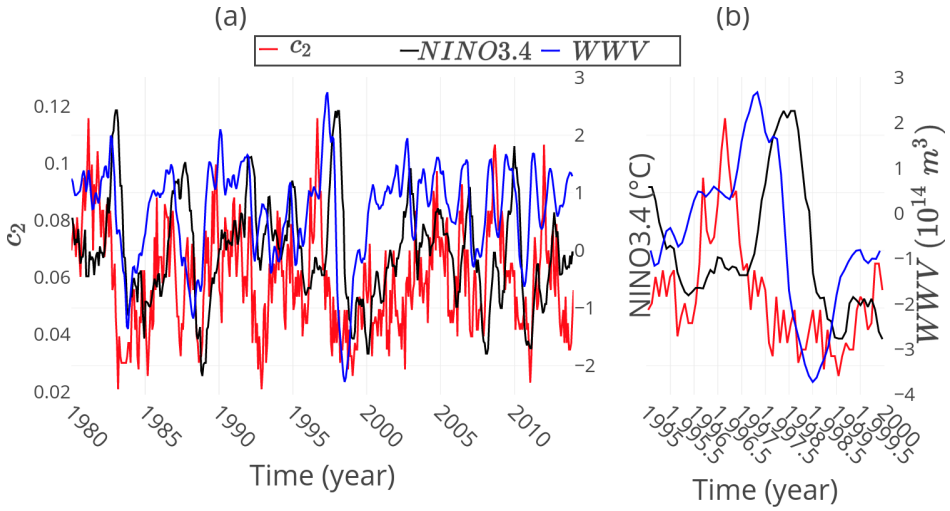


Figure 3. The WWV, c_2 and the NINO3.4 index from observations for (a) the whole considered time series and (b) only during the 1997 El Niño. A warning of the El Niño event is visible for the WWV and c_2. c_2 gives a warning almost a year before the 1997 El Niño, while the WWV warns almost seven months ahead.

related network variable. The correlations of the SSH time series on a grid of 27 latitude points and 30 longitude points in the Pacific area are used to reconstruct a ~~CN~~ network with a threshold $\epsilon = 0.9$ and a sliding window of one year. The sea surface height (SSH) is used instead of thermocline depth, because more data is available and it is by approximation proportional to the thermocline depth (Rebert et al., 1985). During an El Niño event, the link density of this network increases in the warm pool and the cold tongue specifically, causing a percolation-like transition. As discussed in the previous section, an early warning could be obtained with c_2. This variable allows us to extend the lead time of the WWV (Fig. 3). Third, atmospheric noise from the

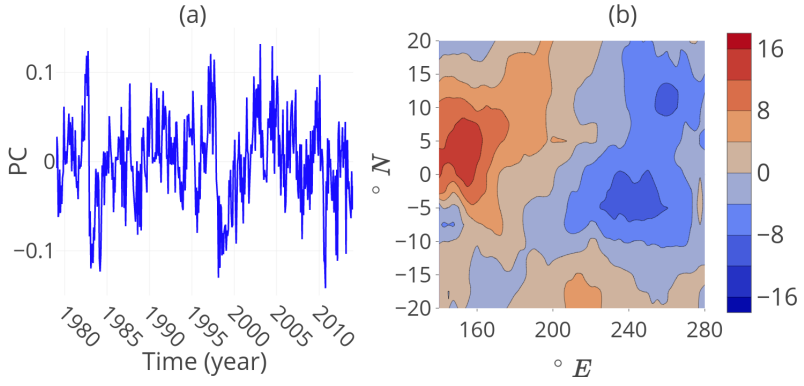


Figure 4. (a) The second ~~PC~~ principal component of the residual of the wind stress (PC_2) and (b) its EOF, associated with the WWBs.

WWBs are a limitation for the prediction of ENSO (Moore and Kleeman, 1999; Latif et al., 1988). To obtain a variable related to the WWBs, the linear effect of the SST is subtracted from the zonal component of the wind stress. The second principal component (PC_2), explaining 8 % of the variance, is associated with these WWB's. In Fig. 4, the ~~PC~~ principal component and its EOF are presented. The peaks in the ~~PC~~ principal component are visible before the great El Niño events of 1982 and 1997. Thereby, the EOF has the typical WWB structure, being positive west from the dateline and negative east. Finally, the attribute set does not yet contain any information about the seasonal cycle (SC) yet. The phase locking of an El Niño event to boreal winter is very typical to ENSO. Therefore a sinusoid with the period of a year is used as attribute, to see if it can improve the prediction skill.

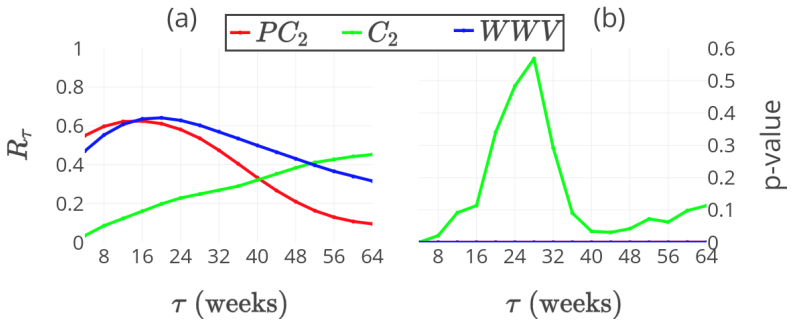


Figure 5. (a) The cross correlation of the PC_2, WWV and c_2 with respect to NINO3.4 for different lags τ. (b) The p-value of the Wiener-Granger hypothesis test for the same lags. A low p-value implies the variable is likely to Granger cause the NINO3.4 index at the specific lag. The p-values of the PC_2 and WWV are almost zero for all lags.

To determine at which lead time the different attributes should be applied, the cross-correlation and the p-value of the Granger test between the attributes and NINO3.4 are considered (Fig. 5). The cross-correlations of PC_2 and the WWV show peaks at respectively 12 and 20 weeks, indicating their optimal lead times, since the p-values of the Granger tests are low at every lag and autocorrelation does not play an important role. For c_2 however, the cross-correlation increases up to the maximum considered lag, but the p-value of the Granger test has a local minimum close to a lag of 44 weeks. According to these methods, c_2 is especially predictive at the longer lead times close to 44 weeks.

To summarize, we are interested in the variables that represent specific physical characteristics related to the prediction of ENSO, to select the attributes. Both c_2 and the WWV are related to the recharge/discharge mechanism. PC_2 is related to the atmospheric noise from WWBs. The seasonal cycle (SC) is related to the phase locking of El Niño events to boreal winter. The hybrid model allows us to implement different variables in the attribute set at different lead times. Therefore, the cross-correlations and Wiener-Granger causality were used to determine which attribute is more optimal at various lead times. This showed that it is better to use c_2 instead of WWV at lead times of more than 40 weeks. The other network variables which were interesting for the ZC model output (~~as shown in the previous subsection~~see the Appendix) are performing worse when applied to observations and hence are not used as attributes in the hybrid model.

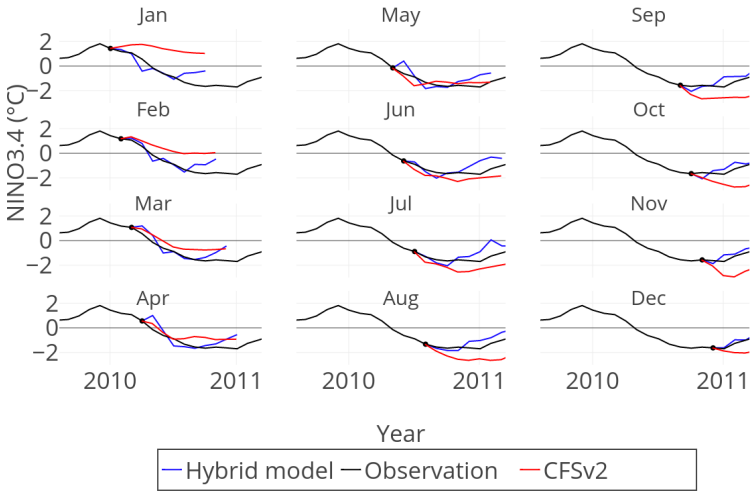


Figure 6. Nine-month ahead prediction starting from every month in the year 2010. Blue is the hybrid model prediction with ARIMA(12,1,1), $2 \times 1 \times 1$ ANN structure and attributes are the three-month running mean of WWV, PC_2 and SC. The black line is the observed index. Red is the mean of the CFSv2 ensemble prediction.

4 Prediction results

This section presents the predictions of the hybrid model, as compared with observations and with alternative predictions from the CFSv2 model ensemble of NCEP. The skill with ANN structures up to three hidden layers is investigated. First, a comparison between both predictions is made for the year 2010 (Fig. 6). Moreover, several lead time predictions are shown and compared to the available CFSv2 lead time predictions. Next it is shown that these prediction models converge to similar results for different hyperparameters and when using different training and test sets in a cross-validation method. Finally, a recent forecast is made and it is shown how the hybrid model predicts the development of ENSO the coming year.

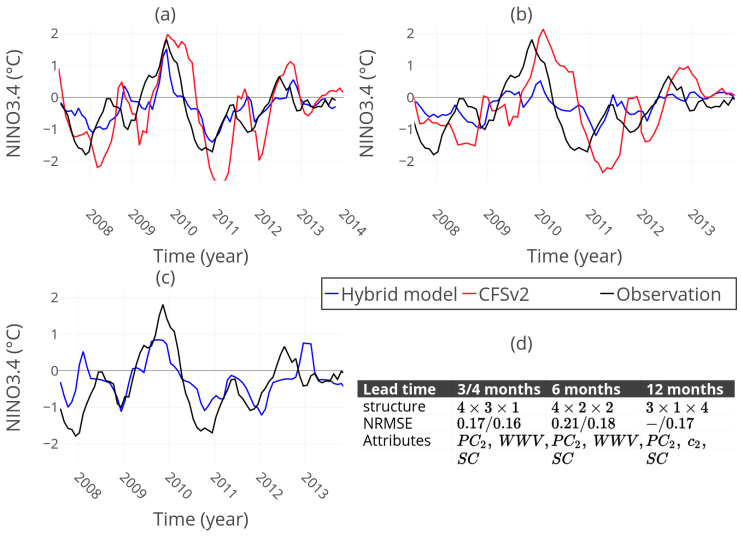
From now on, the Normalized Root Mean Squared Error (NRMSE) is used to indicate the skill of prediction within the test set:

$$NRMSE(y^A, y^B) = \frac{1}{\max(y^A, y^B) - \min(y^A, y^B)} \sqrt{\frac{\sum_{t_1^{test} \le t_k \le t_n^{test}} \left(y_k^A - y_k^B\right)^2}{n}}.$$

Here y_k^A, y_k^B are respectively the NINO3.4 index and its prediction at time t_k in the test set. n is the number of points in the test set. A low NRMSE indicates the prediction skill is better. For all presented hindcasts, the ARIMA prediction had a significant residual, which implies that the addition of the ANN part improved prediction.

The year 2010 is a recent example of an under-performing CFSv2 ensemble. Especially in January, all members of the ensemble overestimate the NINO3.4 index, resulting in an overestimation of the ensemble mean (see Fig. 6). The hybrid model is used to predict the same period, with ARIMA(12,1,1) and a $2 \times 1 \times 1$ ANN structure with the three-month running mean

of the WWV, PC_2, the ~~SC~~ seasonal cycle and NINO3.4 itself as attributes. In this case the hybrid model performs better than the CFSv2 ensemble. A $2 \times 1 \times 1$ structure means a feed-forward structure with three layers of respectively two, one and one neuron. This ANN structure is found to be the best performing structure in terms of NRMSE at a three-month lead time prediction. It will probably not be the most optimal ANN structure at other lead times.



(d)

Lead time	3/4 months	6 months	12 months
structure	$4 \times 3 \times 1$	$4 \times 2 \times 2$	$3 \times 1 \times 4$
NRMSE	$0.17/0.16$	$0.21/0.18$	$-/0.17$
Attributes	PC_2, WWV, SC	PC_2, WWV, SC	PC_2, c_2, SC

Figure 7. NINO3.4 predictions of the CFSv2 ensemble mean (red) and the hybrid model with ARIMA(12,1,0) (blue), compared to the observed index (black). For the hybrid model predictions, from an ensemble of eighty-four different ANN structures, structures resulting in a low NRMSE are presented. (a) The three-month lead time prediction of CFSv2 and four-month lead time prediction of the hybrid model, (b) the six-month lead time predictions and (c) twelve-month lead prediction. The CFSv2 ensemble does not predict twelve months ahead. (d) Table containing information about all predictions: ANN structures of the hybrid model, NRMSEs of the CFSv2 ensemble mean and the hybrid model, and attributes used in the hybrid model predictions.

Considering the three, six and twelve-month lead time predictions, both the three and six-month lead time prediction of the CFSv2 ensemble show some lag and amplification of the real NINO3.4 index (Fig. 7). The hybrid model predictions with ARIMA(12,1,0) resulting in a low NRMSE and relatively simple ANN structure within an ensemble consisting of eighty-four different ANN structures are also shown in Fig. 7. The eighty-four different structures ~~,~~ are all structures up to three hidden layers with up to four neurons.

Comparing the three-month lead prediction of the CFSv2 ensemble with the four-month lead prediction of the hybrid model, ~~the~~ both the amplification and the lag of the ~~prediction is less and the amplification is not as large in the hybrid model~~ hybrid model prediction are smaller. While the lead time of the hybrid model is one month longer, the prediction skill ~~of the hybrid model~~ is better in terms of NRMSE. The prediction skill of the hybrid model decreases at a six-month lead compared to the four-month lead time prediction. Thereby the lag and amplification of the CFSv2 prediction increase. Although the hybrid

model does not suffer as much from the lag, it underestimates the El Niño event of 2010. In terms of NRMSE the hybrid model still obtains a better prediction skill.

Although the shorter lead time predictions show slightly better results than the conventional models, most important is a good prediction skill for larger lead times that appears to overcome the spring-predictability barrier. To perform a twelve-month lead prediction which could overcome this barrier, the attributes from the shorter lead time predictions are found to be insufficient. However, c_2 of the SSH network has shown to be predictive at this lead time, according to its Granger causality and cross-correlation. Therefore the WWV is replaced by c_2 for this prediction, which is related to the same physical mechanism. In terms of NRMSE, the twelve-month lead prediction even improves the six-month lead prediction of the hybrid model. On average the prediction does not contain a lag in this period.

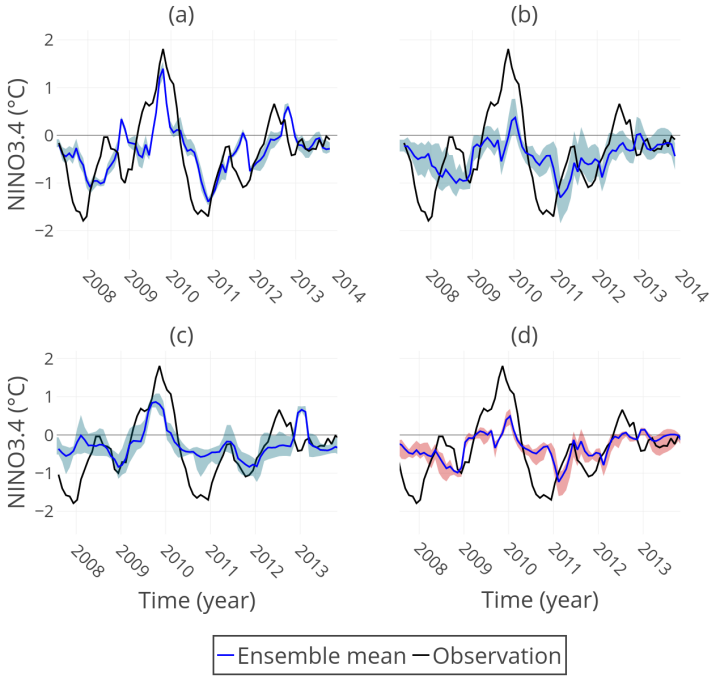


Figure 8. ~~Predictions~~ Spread and mean (blue line) of ~~the NINO3.4 index from an ensemble~~ ensembles of hybrid ~~models~~ model predictions with different ~~ANN structures and ARIMA~~hyperparameter values. The nine optimal (~~12,1,0~~in terms of NRMSE) ~~(blue) compared to~~ predictions from the ~~observed index~~ eighty-four different ANN structures at the ~~(black) with (~~a) ~~four-month~~ four month lead time (b) ~~six-month~~ six month lead time and (c) ~~twelve-month~~ twelve month lead time. ~~The ensemble consists of nine predictions~~ (d) Ensemble with $9 \leq p \leq 14$ in the ~~lowest NRMSE out of eighty-four predictions. The shaded blue area denotes~~ ARIMA order with their optimal ANN structure at six month lead time prediction (at the ~~spread of the nine predictions~~ four and ~~the blue line the mean~~twelve month lead there is almost no spread). ~~The NRMSE of~~ Black is the ~~ensemble mean predictions are respectively 0.15, 0.18, 0.17~~observed NINO3.4 index.

~~To show that the results of~~ The hyperparameter values (i.e. the ARIMA order and the ANN structure) of the predictions in Fig. 7 ~~can be generalized, the mean~~ could still be a lucky shot. Therefore the spread of the predictions with ~~the nine~~

~~lowest NRMSE of the ensemble with eighty-four different ANN structures is considered (~~different hyperparameter values is shown in Fig. 8~~). This ensemble does not differ much from the best predictions of Fig. 7. The spread of the ensemble remains limited, although it is a bit larger~~ . For the ANN structures, nine optimal (in terms of NRMSE) predictions from the ensemble of eighty-four are considered. This resulted in a higher spread in the six and ~~twelve-month~~ twelve month lead prediction compared to the ~~four-month~~ four month lead prediction. For the ARIMA order all $9 \leq p \leq 14$ are chosen, which resulted in almost no spread for the six and twelve month lead prediction and a higher spread in the six month lead prediction. Overall the models converge to similar predictions for those different hyperparameter values.

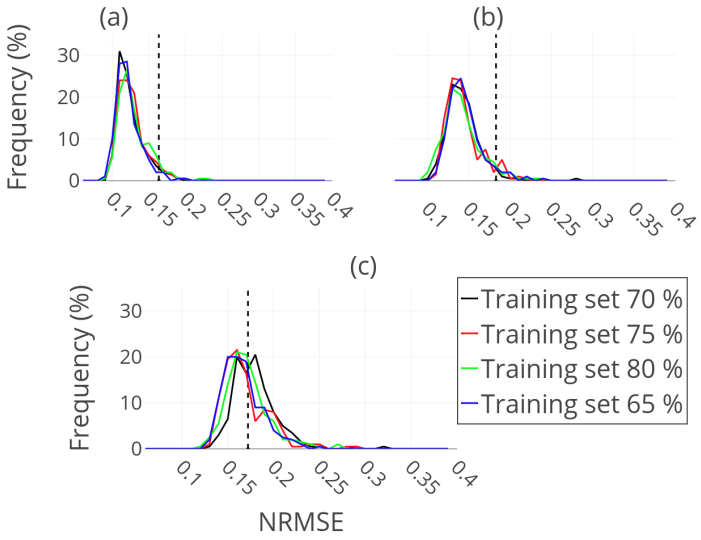


Figure 9. Cross validation results of the (a) four, (b) six and (c) twelve-month lead predictions of hybrid models from Fig. 7. Each line presents the frequency every NRMSE is obtained for 200 different initial test sets with a specific training set/test set percentage split . The vertical dashed line denotes the NRMSE of the predictions of Fig. 7.

To test the robustness of these results, a series of cross-validations has been performed ~~.~~ on the prediction models of Fig. 7. Several percentage splits have been chosen for the training and test set (65-35, 70-30, 75-25 and 80-20), but 200 different initial times of the test set t_i^{test} are randomly chosen between March 1985 and December 2014. This implies that $t_i^{test} > t_f^{train}$ is not necessarily satisfied anymore. This allows us to make full use of the short time series we have (Bergmeir and Benítez, 2012). If the results for different training and test sets do not deviate much, it is evidence that the model ~~also generalizes~~ can generalize to an arbitrary training and test set. The different percentage splits are chosen since the size of a training set could possibly have influence on the prediction model. The cross validation results of the hybrid models of Fig. 7 are presented in Fig. 9. At all three prediction lead times, the peaks coincide at the same NRMSE for different training-test set ratios. Therefore the different sizes of training and test sets do not seem to influence the result. However, the width of the peaks increases when the prediction lead time increases. This implies the prediction skill becomes more sensitive to the choice of the training and test

set ~~in~~ with higher lead time. Interestingly, at the four and six-month lead time predictions, the average NRMSE is lower than the NRMSE of the prediction of Fig. 7. This implies the predictions with a different training and test set are on average even better than the prediction shown in Fig. 7.

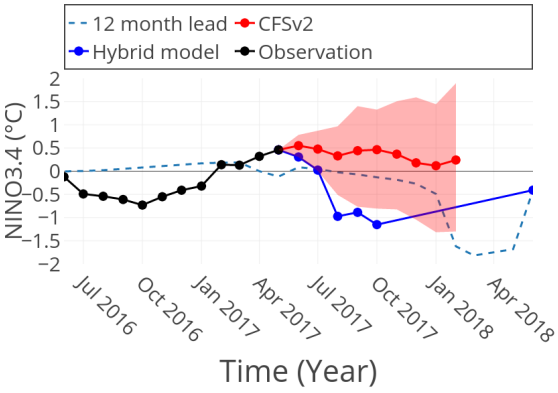


Figure 10. NINO3.4 prediction from May 2017. In black the observed index until May 2017. Red is the CFSv2 ensemble prediction mean and the shaded area is the spread of the ensemble. The hybrid model prediction in blue is given by predictions from hybrid models found to be most optimal at the different lead times with ARIMA(12,1,0). The dashed blue line is the running twelve-month lead time prediction.

Finally, a prediction is made for the coming year in Fig. 10. Different hybrid models are used at different lead times with ARIMA(12,1,0). ANN structures are chosen that are found to be optimal at the different lead times. For the predictions up to five months, the attributes WWV, PC_2 and ~~SC~~ the seasonal cycle are used from 1980 until present. For the twelve-month lead prediction, the WWV is replaced by c_2 again. This time c_2 is computed from the SSALTO/DUACS dataset. Therefore, only a dataset from 1993 until present has been used to train the model and perform the twelve-month lead prediction.

Interestingly, as can be seen in Fig. 10, the hybrid model typically predicts much lower ENSO development than the CFSv2 ensemble. The uncertainty of the CFSv2 ensemble is large, since the spread of predictions is between a strong El Niño (NINO3.4 index between 1.5 and 2) and a moderate La Niña (NINO3.4 index between -1 and -1.5) for the coming 9 months. The hybrid models predict development to a strong La Niña (NINO3.4 index lower than -1.5) the coming year. From the time of writing, only time will tell which prediction is better. By the time of submission in early March 2018, La Niña conditions are present according to the Climate Prediction Centre of NCEP.

5 Summary and Discussion

A successful attempt was made in this paper to use Machine Learning techniques in a hybrid model to improve the skill of El Niño predictions. Crucial for the success of this hybrid model is the choice of the attributes applied to the Artificial Neural Network~~(ANN)~~. Here, we have explored the use of network variables as additional attributes to several physical ones. Results of the ZC model provided several interesting network variables~~, such as the cross clustering between wind and SST network,~~

~~the zonal skewness in the degree field of the thermocline network, and two variables anticipating a percolation-like transition (Δ, c_2) and finally the diffusivity of information in the wind and SST network (λ_2).~~

. Of these network variables, c_2 the amount of clusters of size two in a SSH network constructed from observations, is found to provide a warning of a percolation-like transition in the SSH network. This percolation-like transition coincides with an El Niño event. This variable relates to the WWV and hence the recharge/discharge mechanism, but extends the prediction lead time of the WWV when applied in the prediction scheme. Furthermore, apart from both these 'recharge/discharge' related quantities, the PC_2 and ~~SC~~ the seasonal cycle improve the prediction skill, representing respectively the WWBs and the phase locking of ENSO. The flexibility of implementing different variables at different lead times, allows the hybrid model to improve on the CFSv2 ensemble at short lead times (up to six months). Furthermore, it had a better prediction result than all members of the CFSv2 ensemble in January 2010.

By including the network variable c_2, we obtained a twelve-month lead time prediction with comparable skill to the predictions at shorter lead times. This prediction shows a step towards beating the spring predictability barrier. Using ML has the advantage of recognizing the early warning signal of c_2 as either a false or true positive. Therefore, it can be a more reliable method then considering a warning when the signal exceeds a certain threshold (Ludescher et al., 2014). Moreover, the early signal from the network variable is not only used to predict an El Niño event, but the development of ENSO, as the hybrid model provides a regression of the NINO3.4 index. ML serves as a tool which is able to recognize important, but subtle patterns. Something the conventional statistical and dynamical models fail to do in the chaotic system. In the end, the predictions from May 2017 are discussed. By the time of writing, this is the prediction for the coming year. The CFSv2 ensemble mean predicts neutral conditions the coming nine months, with the spread between different members ranging from a strong El Niño to a moderate La Niña. The hybrid model predicts moderate to strong La Niña conditions for the coming year.

Although the results of the methods are promising, some adaptations to the methods ~~attributes~~ which select attributes could still improve predictions. Several network variables resulted in a clear signal in the ZC model, but not necessarily for the observations. Perhaps the cross-correlation and a Granger causality test are not enough to determine the suitability of a variable in the observations. Testing all possible attribute sets in the prediction scheme and comparing results costs time. As a solution, the nonlinear methods 'lagged mutual information' and 'transfer entropy' can be ~~considered~~ techniques to select variables at different lead times. After all, the attributes are applied in the nonlinear part of the prediction scheme. Consequently, more variables might be found to increase the prediction skill.

Even though the currently applied network measures showed interesting properties, different ~~CN~~ Climate Network construction methods can still be interesting to apply. The Pearson correlation is a simple, effective method to define links between nodes. However, different properties of ~~CNs~~ Climate Networks could be found when using mutual information instead. Moreover, the effect of spatial distance between nodes can be investigated and corrected for (Berezin et al., 2012). Besides, we have limited ourselves to networks within the Pacific area itself. As ENSO is an important mode in the whole climate system, the area used for ~~CN~~ network construction might as well be extended. More specifically, it can be interesting to include the Indian Ocean in the ~~CN~~ network construction. Evidence is found that a cold SST in the West of the Indian Ocean is related to a WWB

a few months later (Wieners et al., 2016). This could result in a variable related to WWBs, but increasing the lead compared to PC_2, which is comparable to c_2 increasing the lead compared of the WWV.

By applying the ARIMA as simple, yet effective statistical method to apply in the first step of the scheme, the hybrid model shows promising results. However, the exact reason how this model works, remains a topic of investigation. The ARIMA prediction could be related to the linear wave dynamics. It can be interesting to replace the ARIMA part of the scheme by a dynamical model accounting for these linear wave dynamics. For the same reason Vector Autoregression ~~(VAR)~~ can be used instead of ARIMA. Being a multivariate generalization of an autoregressive model, this can implement the linear effect of other variables on ENSO.

Next to investigation of the exact reason the hybrid model works, some adaptations could still improve the prediction scheme. For example, it is assumed the linear and nonlinear part of the model are additive (see Eq. (3)). This is not necessarily the case for the real system (Khashei and Bijari, 2011). Besides, the current model does not take into account possible nonlinear effects from the history, since the ANN describes a nonlinear function which does not depend on the history. The ANN probably succeeds here because of its performance for nonlinear time series in general. However, it could be interesting to investigate whether Climate Network properties comprise enough of the nonlinear dynamics by themselves, by combining them with a purely linear model. Moreover, the applied methods searched for a prediction model which is most optimal in terms of least squares minimization. However, it could be interesting to put larger weight at predicting the extreme events in the optimization scheme (as the six-month lead predictions missed the 2010 El Niño event in Fig. 8), or find a function which is more simple (e.g. applying a support vector machine instead of ANN (Pai and Lin, 2005)).

~~Although the~~ A general difficulty in El Niño prediction is the short available observational time series, also in other statistical prediction models (Drosdowsky, 2006). Although different hyperparameters (the ANN structure and ARIMA order) converge to a similar prediction and the prediction models perform well at different training and test sets, the short time series make it difficult to perform another cross-validation method which completely rules out that the model is overfitting.

Although the hybrid model and the attribute selection can clearly be improved, the results here have shown the potential for ML methods, in particular with network attributes, for El Niño prediction. The underlying reason for this success is likely that through the network attributes, more global correlations are taken into account which are needed to be able to overcome the ~~spring predictability~~ spring-predictability barrier.

Appendix A

This appendix summarizes the methods to calculate Climate Network properties. They improved the prediction in the ZC model, but not for observational data. Thus, they are not discussed in the main text. Appendix A1 defines the different quatities and Appendix A2 their application to the ZC model.

A1 ~~Pearson correlation~~ Alternative Network methods

~~The Pearson correlation between variables p_i, p_j associated with two points on a spatial grid i, $j \in [0, \cdots, N]$ is defined as:~~ From the unweighted network we compute the local degree d_i of node i in the network as,

$$d_i = \sum_j A_{ij}, \tag{A1}$$

i.e. degree is equal to the amount of nodes that are connected to node i.

The spatial symmetry of the degree distribution is of interest, since it informs where most links of the network are located. More specifically, our interest will be in the symmetry in the zonal direction in a network. Therefore, the skewness of the meridional mean of the degree in the network is calculated. This defines the zonal skewness of the degree distribution in a network.

The following two Climate Network properties are derived from a so-called NetOfNet approach. This is a network constructed with the same methods as previously, but using multiple variables at each grid point (as specified in Appendix A2). This gives a network consisting of the networks from the different variables interacting with each other. Only NetOfNet of two different variables are considered. First, the cross clustering contains information about the interaction between two unweighted networks. The local cross clustering of a node is the probability that two connected nodes in the other network are also connected to each other. The global cross clustering C_{vw} is the average over all nodes in subnetwork G_v of the cross clustering between G_v and G_w:

$$R(i,j)C_{vw} = \frac{\sum_{k=1}^{n} p_i(t_k)p_j(t_k)}{\sqrt{\left(\sum_{k=1}^{n} p_i^2(t_k)\right)\left(\sum_{k=1}^{n} p_j^2(t_k)\right)}} \frac{1}{N_v} \sum_r \frac{1}{k_r\,(k_r-1)} \sum_{p \neq q} A_{rp}A_{pq}A_{qr}. \tag{A2}$$

~~Here p_i is a vector of size n of~~ Here r is a node in subnetwork G_v of size N_v, p and q are the nodes in the other subnetwork G_w and k_r denotes the cross degree of node r (i.e. amount of cross links node r has with the other subnetwork).

The second NetOfNet property is the ~~time series at time steps t_k. The temporal mean is subtracted and data is detrended before the correlation is calculated.~~

A2 ~~Algebraic connectivity~~

~~Let ψ_i be the time series at node i in a network. How much ψ_i changes by a hypothetical diffusionprocess occurring~~ algebraic connectivity. This is the second smallest eigenvalue (λ_2) of the Laplacian matrix as in Newman (2010) and describes the 'diffusion' of information in the network~~depends on the values of nodes j itis connected to according to the unweighted adjacency matrix A (Newman, 2010):~~ In general, $\lambda_2 > 0$ if the network has a single component.

A final network property Δ makes use of differently calculated network which is also undirected, but weighted. To construct it, the cross-correlation $C_{ij}(\Delta t)$ at lag Δt, i.e. the Pearson correlation between the variables $p_i(t)$ and $p_j(t + \Delta t)$ is considered. Then the weights between the nodes are calculated by:

$$\frac{d\psi_i}{dt} = C \sum_j A W_{ij} \psi_j - \psi_i = \frac{\max_{\Delta t}(C_{ij}) - \text{mean}(C_{ij})}{\text{std}(C_{ij})}. \tag{A3}$$

~~Here C is the diffusion constant and A the adjacency matrix. By separating the sum in~~ Here $\max_{\Delta t}$ denotes the maximum, std the standard deviation and mean the mean value over all time steps that are considered.

To calculate the property Δ of the network, links are added to a network one by one, adding the link with the largest weight first (Eq. (~~??), this can be rewritten as:~~

$$\frac{d\psi_i}{dt} = C \sum_j (A_{ij} - \delta_{ij} d_i)\, \psi_j,$$

~~δ_{ij} is the Kronecker delta function and d_i the degree of node i. In matrix notation this reduces to~~ A3)). At every step T that a link is added, the size of the largest cluster $S_1(T)$ is calculated. At the point of the percolation transition, $S_1(T)$ increases rapidly. The size of this jump is Δ:

$$\frac{d\psi_i}{dt} = C\,(\mathbf{A} - \mathbf{D})\,\psi.$$

~~Here~~~~$\mathbf{D}$ with $\mathbf{D_{ij}} = \delta_{ij}\mathbf{d_i}$ is a square matrix that contains the degrees at the diagonal and zero elsewhere. Now we define the graph Laplacian of the network as:~~

$$\mathbf{L} = \mathbf{A} - \mathbf{D}.$$

~~Equation ?? reduces to the diffusion equation, but with the graph Laplacian matrix $\mathbf{L}$ instead of ∇^2. By calculating the eigenvalues of~~

$$\Delta = \max\left[S_1(2) - S_1(1), \ldots, S_1(T+1) - S_1(T), \ldots\right]. \tag{A4}$$

The quantity Δ can be used to capture the percolation-like transition (Meng et al., 2017).

A2 Climate network properties of the ZC model

Determining how strong noise can excite the ENSO mechanisms in the sub-critical case, or determining whether the feedbacks sustain an oscillation in the supercritical state, could provide information to increase the prediction skill. Feng (2015) found that the skewness of the degree distribution S_d of the network reconstructed from SST decreases monotonically with increasing coupling strength μ. Although S_d relates to the climate stability and coupling strength, it does not inform whether the system is in either the supercritical or sub-critical state. Here, we introduce a NetOfNet variable which may represent properties of the stability of the background state: the global cross clustering (C_{vw}) between the SST and wind-stress network. A sliding window of five years with $\epsilon = 0.6$ was used to compute the networks. In this case, the global cross clustering coefficient is a measure of the amount of triangles in the networks, containing one wind node and two SST nodes. In Fig. A1, this cross clustering is calculated from data from the ZC model, when coupling strength μ changes periodically in time around the critical value $\mu_c \sim 3.0$. Under sub-critical conditions, the noise has a larger influence on local correlations. This causes triangles to break and the variance of the cross clustering coefficient to increase. The cross clustering C_{vw} is hence a diagnostic network variable which informs whether the state of the system is in the supercritical or sub-critical regime.

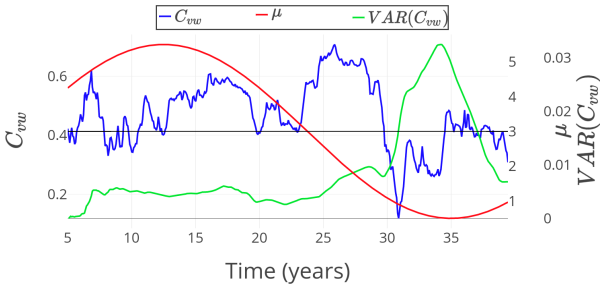


Figure A1. Global cross clustering between the SST and wind-stress network in blue and its variance in green in the ZC model. The coupling strength μ defined as a sinusoid around $\mu_c = 3$ with an amplitude of 0.25 in red. The sliding window is applied with a window of five years.

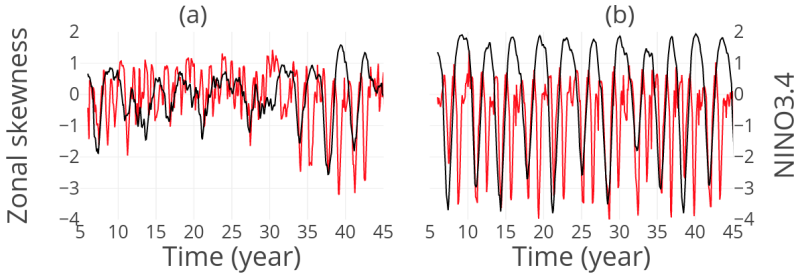


Figure A2. Zonal skewness of the degree field of the thermocline network with $\epsilon = 0.6$ and a sliding window of one year in red, NINO3.4 index in black in the ZC model. (a) The sub-critical ($\mu = 2.7$) and (b) the supercritical ($\mu = 3.25$) case.

Second, from the classical view of the oscillatory behavior of ENSO, waves in the thermocline should contain memory of the ~~Laplacian matrix $\lambda_1, \cdots, \lambda_n$ with $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, we can determine the diffusion within the network~~ system, because of their negative delayed feedback. The changing structure of the thermocline network is therefore of interest when predicting ENSO. Calculating this network with threshold $\epsilon = 0.6$ and a sliding window with a length of one year, a zonal pattern in the change of the network close to the equator can be observed during an ENSO cycle. To compare network structures in the super- and sub-critical state, now constant $\mu = 2.7$ (sub-critical) and $\mu = 3.25$ (supercritical) are taken. Generally, the degree field is quite spatially symmetric, but when the ENSO turns either from upward to downward, or from downward to upward, the degree of the nodes in the east decreases. This is at the peak El Niño or La Niña. ~~Since the matrix is symmetric, the eigenvalues are real. Moreover, the smallest eigenvalue λ_1 is always zero and no eigenvalues are negative. This means ψ_i will decay to a stable solution. The second smallest eigenvalue, called the algebraic connectivity, is of particular interest. In general, $\lambda_2 > 0$ if the network has a single component~~

To capture this zonal asymmetry around the equator with a variable, the zonal skewness of the degree field will be used between 7°S to 7°N. The higher the skewness, the more the degree will be located west of the basin. If the skewness is close to zero, the degree is symmetrically distributed over the basin. If it is low, most of the degree is situated in the east. The skewness will show a negative peak when the ENSO index is at its highest or lowest point in the cycle (Fig. A2). In the supercritical

case $\mu = 3.25$ this effect is indeed observed. Nevertheless, in the sub-critical case, the pattern is only visible once ENSO index shows a clear oscillation (around year 32).

Third, the quantity Δ behaves similar to c_2, when calculated from the same (thermocline) network. Although Δ does not depend on a chosen threshold like c_2, it peaks closer to an El Niño event.

Finally, the algebraic connectivity (λ_2) can show the spread of information within a network. Specifically, when considering an unweighted NetOfNet from thermocline depth (h) and zonal wind (τ^x) with threshold $\epsilon = 0.6$. The spread of information is relatively high before an event, but also after an event, such that λ_2 peaks both before and after an El Niño event (both for $\mu = 2.7$ and $\mu = 3.25$).

Acknowledgements. PN would like to thank the *Instituto de Física Interdisciplinar y Sistemas Complejos* (IFISC), for hosting his stay in Mallorca during part of 2017.

CL and EHG acknowledge support from Ministerio de Economia y Competitividad and Fondo Europeo de Desarrollo Regional through the LAOP project (CTM2015-66407-P,MINECO/FEDER)

References

Akaike, H.: A New Look at the Statistical Model Identification, IEEE Transactions on Automatic Control, AC-19, 716–723, https://doi.org/doi:10.1109/TAC.1974.1100705, 1974.

Al-Smadi, A. and Al-Zaben, A.: ARMA Model Order Determination Using Edge Detection: A New Perspective, Circuits, Systems Signal Processing, 24, 723–732, 2005.

Aladag, C. H., Egrioglu, E., and Kadilar, C.: Forecasting nonlinear time series with a hybrid methodology, Applied Mathematics Letters, 22, 1467–1470, https://doi.org/10.1016/j.aml.2009.02.006, 2009.

Berezin, Y., Gozolchiani, A., Guez, O., and Havlin, S.: Stability of Climate Networks with Time, Scientific Reports, 2, 1–8, https://doi.org/10.1038/srep00666, 2012.

Bergmeir, C. and Benítez, J. M.: On the use of cross-validation for time series predictor evaluation, Inf. Sci. (Ny)., 191, 192–213, https://doi.org/10.1016/j.ins.2011.12.028, 2012.

Bishop, C. M.: Pattern Recognition and Machine Learning, Springer-Verlag New York, 2006.

Bjerknes, J.: Atmospheric Teleconnections From The Equatorial Pacific, Monthly Weather Review, 97, 163–172, https://doi.org/10.1175/1520-0493(1969)097<0163:ATFTEP>2.3.CO;2, 1969.

Bosc, C. and Delcroix, T.: Observed equatorial Rossby waves and ENSO-related warm water volume changes in the equatorial Pacific Ocean, Journal of Geophysical Research, 113, 1–14, https://doi.org/10.1029/2007JC004613, 2008.

Bunge, L. and Clarke, A. J.: On the Warm Water Volume and Its Changing Relationship with ENSO, Journal of Physical Oceanography, 44, 1372–1385, https://doi.org/10.1175/JPO-D-13-062.1, 2014.

Chen, D., Cane, M. A., Kaplan, A., Zebiak, S. E., and Huang, D.: Predictability of El Niño over the past 148 years., Nature, 428, 733–736, https://doi.org/10.1038/nature02439, 2004.

Deza, J. I., Masoller, C., and Barreiro, M.: Distinguishing the effects of internal and forced atmospheric variability in climate networks, Nonlinear Processes in Geophysics, 21, 617–631, https://doi.org/10.5194/npg-21-617-2014, 2014.

Dijkstra, H. A.: The ENSO phenomenon: theory and mechanisms, Advances in Geosciences, 6, 3–15, https://doi.org/10.5194/adgeo-6-3-2006, 2006.

Drosdowsky, W.: Statistical prediction of ENSO (Nino 3) using sub-surface temperature data, Geophys. Res. Lett., 33, 10–13, https://doi.org/10.1029/2005GL024866, 2006.

Fedorov, A. V., Harper, S. L., Philander, S. G., Winter, B., and Wittenberg, A.: How predictable is El Niño?, Bulletin of the American Meteorological Society, 84, 911–919, https://doi.org/10.1175/BAMS-84-7-911, 2003.

Feng, Q. Y.: A complex network approach to understand climate variability, Ph.D. thesis, Utrecht University, 2015.

Feng, Q. Y. and Dijkstra, H. A.: Climate Network Stability Measures of El Niño Variability, 035801, https://doi.org/10.1063/1.4971784, 2016.

Feng, Q. Y., Vasile, R., Segond, M., Gozolchiani, A., Wang, Y., Abel, M., Havlin, S., Bunde, A., and Dijkstra, H. A.: ClimateLearn: A machine-learning approach for climate prediction using network measures, Geoscientific Model Development Discussions, pp. 1–18, https://doi.org/10.5194/gmd-2015-273, 2016.

Fountalis, I., Bracco, A., and Dovrolis, C.: ENSO in CMIP5 simulations: network connectivity from the recent past to the twenty-third century, Climate Dynamics, 45, 511–538, https://doi.org/10.1007/s00382-014-2412-1, http://dx.doi.org/10.1007/s00382-014-2412-1, 2015.

Gill, A.: Some simple solutions for heat-induced tropical circulation, Quart. J. Roy Meteor. Soc.,, 106, 447–462, 1980.

Goddard, L., Mason, S., Zebiak, S., Ropelewski, C., Basher, R., and Cane, M.: Current Approaches to seasonal-to-interannual climate predictions, International Journal of Climatology, 21, 1111–1152, https://doi.org/10.1080/002017401300076036, 2001.

Gozolchiani, A., Yamasaki, K., Gazit, O., and Havlin, S.: Pattern of climate network blinking links follows El Niño events, EPL (Europhysics Letters), 83, 28 005, https://doi.org/10.1209/0295-5075/83/28005, 2008.

Gozolchiani, A., Havlin, S., and Yamasaki, K.: Emergence of El Niño as an autonomous component in the climate network, Physical Review Letters, 107, 1–5, https://doi.org/10.1103/PhysRevLett.107.148501, 2011.

Guyon, I. and Elisseeff, A.: An Introduction to Variable and Feature Selection, Journal of Machine Learning Research, 3, 1157–1182, https://doi.org/10.1016/j.aca.2011.07.027, 2003.

Hall, M. A.: Correlation-based Feature Selection for Machine Learning, Ph.D. thesis, The university of Waikato, 1999.

Hibon, M. and Evgeniou, T.: To combine or not to combine: Selecting among forecasts and their combinations, International Journal of Forecasting, 21, 15–24, https://doi.org/10.1016/j.ijforecast.2004.05.002, 2005.

Huang, B., Banzon, V. F., Freeman, E., Lawrimore, J., Liu, W., Peterson, T. C., Smith, T. M., Thorne, P. W., Woodruff, S. D., and Zhang, H. M.: Extended reconstructed sea surface temperature version 4 (ERSST.v4). Part I: Upgrades and intercomparisons, Journal of Climate, 28, 911–930, https://doi.org/10.1175/JCLI-D-14-00006.1, 2015.

Hush, M. R.: Machine learning for quantum physics, Science, 355, 580, 2017.

Jin, F.-F.: An Equatorial Ocean Recharge Paradigm for ENSO. Part II: A Stripped-Down Coupled Model, Journal of the Atmospheric Sciences, 54, 830–847, https://doi.org/10.1175/1520-0469(1997)054<0830:AEORPF>2.0.CO;2, 1997.

Jin, F.-F., Neelin, D. J., and Ghil, M.: El Niño on the Devil's staircase: Annual Subharmonic Steps to Chaos, Science, 264, 70–72, https://doi.org/10.1126/science.264.5155.70, 1994.

Khashei, M. and Bijari, M.: A novel hybridization of artificial neural networks and ARIMA models for time series forecasting, Applied Soft Computing Journal, 11, 2664–2675, https://doi.org/10.1016/j.asoc.2010.10.015, 2011.

Latif, M., Biercamp, J., and von Storch, H.: The response of a Coupled Ocean-Atmosphere General Circulation Model to Wind Bursts, Journal of the Atmospheric Sciences, 45, 1988.

Legler, D. M. and Brien, J. J. O.: 2. Tropical Pacific Wind Stress Analysis for TOGA, Intergovernmental Oceanographic Commission 11, 1988.

Ludescher, J., Gozolchiani, A., Bogachev, M. I., Bunde, A., Havlin, S., and Schellnhuber, H. J.: Very early warning of next El Niño., Proceedings of the National Academy of Sciences of the United States of America, 111, 2064–6, https://doi.org/10.1073/pnas.1323058111, 2014.

Madden, R. A. and Julian, P. R.: Observations of the 40–50-Day Tropical Oscillation—A Review, Monthly Weather Review, 122, 814–837, https://doi.org/10.1175/1520-0493(1994)122<0814:OOTDTO>2.0.CO;2, 1994.

Meng, J., Fan, J., Ashkenazy, Y., and Havlin, S.: Percolation framework to describe El Niño conditions, Chaos, 27, 1–15, https://doi.org/10.1063/1.4975766, 2017.

Moore, A. M. and Kleeman, R.: Stochastic forcing of ENSO by the intraseasonal oscillation, Journal of Climate, 12, 1199–1220, https://doi.org/10.1175/1520-0442(1999)012<1199:SFOEBT>2.0.CO;2, 1999.

Newman, M.: Networks: An introduction, vol. 6, Oxford university press, Oxford, https://doi.org/10.1017/S1062798700004543, 2010.

Pai, P.-F. and Lin, C.-S.: A hybrid ARIMA and support vector machines model in stock price forecasting, Omega, 33, 497–505, https://doi.org/10.1016/j.omega.2004.07.024, 2005.

Philander, S. G.: El Nino, La Nina, and the Southern Oscillation, vol. 46, International Geophysics Series, San Diego, 1990.

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., and Kaplan, A.: Global analyses of sea surface temperature , sea ice , and night marine air temperature since the late nineteenth century, Journal of Geophysical Research, 108, https://doi.org/10.1029/2002JD002670, 2003.

Rebert, J. P., Donguy, J. R., Eldin, G., and Wyrtki, K.: Relations between sea level, thermocline depth, heat content, and dynamic height in the tropical Pacific Ocean, Journal of Geophysical Research, 90, 11 719, https://doi.org/10.1029/JC090iC06p11719, 1985.

Rissanen, J.: Modelling by the shortest data description, Automatica, 14, 465–471, 1978.

Rodríguez-Méndez, V., Eguíluz M, V. M., Hernández-García, E., and Ramasco, J. J.: Percolation-based precursors of transitions in extended systems, Scientific Reports, 6, 29 552, https://doi.org/10.1038/srep29552, 2016.

Runge, J. G.: Detecting and Quantifying Causal Interactions from Time Series of Complex Systems, Ph.D. thesis, Humboldt-Universität zu Berlin, 2014.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D.: Mastering the game of Go with deep neural networks and tree search, Nature, 529, 484–489, https://doi.org/10.1038/nature16961, 2016.

Steinhaeuser, K., Ganguly, A. R., and Chawla, N. V.: Multivariate and multiscale dependence in the global climate system revealed through complex networks, Climate Dynamics, 39, 889–895, https://doi.org/10.1007/s00382-011-1135-9, 2012.

Stolbova, V., Martin, P., Bookhagen, B., Marwan, N., and Kurths, J.: Topology and seasonal evolution of the network of extreme precipitation over the Indian subcontinent and Sri Lanka, Nonlinear Processes in Geophysics, 21, 901–917, https://doi.org/10.5194/npg-21-901-2014, 2014.

Sun, Y., Li, J., Liu, J., Chow, C., Sun, B., and Wang, R.: Using causal discovery for feature selection in multivariate numerical time series, Machine Learning, https://doi.org/10.1007/s10994-014-5460-1, 2014.

Tsonis, A. A., Swanson, K. L., and Roebber, P. J.: What do networks have to do with climate?, Bulletin of the American Meteorological Society, 87, 585–595, https://doi.org/10.1175/BAMS-87-5-585, 2006.

Tupikina, L., Rehfeld, K., Molkenthin, N., Stolbova, V., Marwan, N., and Kurths, J.: Characterizing the evolution of climate networks, Nonlinear Processes in Geophysics, 21, 705–711, https://doi.org/10.5194/npg-21-705-2014, 2014.

Tziperman, E., Stone, L., Cane, M. A., and Jarosh, H.: El Nino chaos: Overlapping of resonances between the seasonal cycle and the pacific ocean-atmosphere oscillator, Science, 264, 72–74, https://doi.org/10.1126/science.264.5155.72, 1994.

Valenzuela, O., Rojas, I., Rojas, F., Pomares, H., Herrera, L. J., Guillen, A., Marquez, L., and Pasadas, M.: Hybridization of intelligent techniques and ARIMA models for time series prediction, Fuzzy Sets and Systems, 159, 821–845, https://doi.org/10.1016/j.fss.2007.11.003, 2008.

van der Vaart, P. C. F., Dijkstra, H. A., and Jin, F. F.: The Pacific Cold Tongue and the ENSO Mode: A Unified Theory within the Zebiak–Cane Model, Journal of the Atmospheric Sciences, 57, 967–988, https://doi.org/10.1175/1520-0469(2000)057<0967:TPCTAT>2.0.CO;2, 2000.

Von Der Heydt, A. S., Nnafie, A., and Dijkstra, H. A.: Cold tongue/Warm pool and ENSO dynamics in the Pliocene, Climate of the Past, 7, 903–915, https://doi.org/10.5194/cp-7-903-2011, 2011.

Wang, Y., Gozolchiani, A., Ashkenazy, Y., and Havlin, S.: Oceanic El-Niño wave dynamics and climate networks, New Journal of Physics, 18, 1–5, https://doi.org/https://doi.org/10.1088/1367-2630/18/3/033021, 2015.

Wieners, C. E., de Ruijter, W. P., Ridderinkhof, W., von der Heydt, A. S., and Dijkstra, H. A.: Coherent tropical Indo-Pacific interannual climate variability, Journal of Climate, 29, 4269–4291, https://doi.org/10.1175/JCLI-D-15-0262.1, 2016.

Wu, A., Hsieh, W. W., and Tang, B.: Neural network forecasts of the tropical Pacific sea surface temperatures, Neural Networks, 19, 145–154, https://doi.org/10.1016/j.neunet.2006.01.004, 2006.

Yeh, S.-W., Kug, J.-S., Dewitte, B., Kwon, M.-H., Kirtman, B. P., and Jin, F.-F.: El Niño in a changing climate, Nature, 461, 511–514, https://doi.org/10.1038/nature08316, 2009.

Zebiak, S. E. and Cane, M. A.: A model El Niño-Southern Oscillation, Monthly Weather Review, 115, 2262–2278, https://doi.org/10.1175/1520-0493(1987)115<2262:AMENO>2.0.CO;2, 1987.

Zhang, G.: Time series forecasting using a hybrid ARIMA and neural network model, Neurocomputing, 50, 159–175, https://doi.org/10.1016/S0925-2312(01)00702-0, 2003.