Predicting ENSO using Gaussian Density Neural Networks trained on distorted physics data

Ivo J. Goede (6975054)

supervisor: prof. dr. ir. Henk A. Dijkstra
second examiner: dr. Claudia E. Wieners

February 11, 2022

Abstract. This research explores the performance of the Gaussian Density Neural Network (GDNN) framework's performance predicting the El Ǹino Southern Oscillation (ENSO) when trained using distorted physics data produced by the Cane & Zebiak 1987 climate model (CZ87). Distorting the wavespeed and thermocline feedback in CZ87 shows a deterioration of the prediction skill of the GDNN at a 9 month lead time. Also shown is a notable capacity of the GDNN to account for differences in amplitude and period in the oscillation the target variable between test and distorted training datasets. Subsequent study may uncover new dynamical relationships at the core of the ENSO phenomenon.

1 Introduction

About twice per decade, measurements of sea surface temperatures (SST) show unusually high values in the Pacific ocean west of Peru [3]. Additionally, a reversal of trade winds and increase in precipitation over western South America are observed. This phenomenon is commonly known as El Niño, which is the warm phase of the El Niño Southern Oscillation (ENSO). This event, which happens every 3 to 7 years, has far-reaching economic consequences for the region (e.g. diminishing fish stocks and tourism) the damage of which is estimated on the order of billions of US dollars.

ENSO is characterized by inter-annual anomalies of the SST in the pacific around the equator. The metric describing this variability is called the Nino 3.4 index (NINO3.4), defined as the 5-month running mean of the SST from 120W to 170W and 5S to 5N. Related to NINO3.4 is the Oceanic Nino Index (ONI) which is defined as the three month running mean of the NINO3.4 index. The behaviour of ENSO is an irregularly periodic oscillation with positive (warm), neutral and negative (cold) phases with transitions marked by the ONI falling

below 0.5 and -0.5 respectively. The warm phase is colloquially better known as El Niño with the cold phase being La Niña.

The effects of El Niño on society make its prediction highly sought after. As such, ENSO has been an important topic of climate research for some time. Consequently, a large amount of climate data has been gathered in the region. A notable project, which significantly increased measurement density, is the Tropical Atmosphere Ocean (TAO) project. The TAO project deployed an array of measurement buoys in the 1980's. This has led to the development of statistical (using statistical relations in the climate data) and dynamical (using physical laws to describe the behaviour) models to predict when the next El Niño will happen. A focus on the development in dynamical models since the 90's have caused dynamical models to outperform their statistical counterparts by a notable margin. Such models are now capable of making predictions with Pearson correlations of $r > 0.6$ up to lead times of about 6 months. Prediction skill of periods starting before boreal (southern hemispheric) spring are generally unreliable (known as the spring predictability barrier (SPB)) which is caused by the high coupling strengths of oceanic processes in that season [3]. Further limitations to dynamical and statistical models are respectively the incomplete understanding of ENSO physics and the computing power required to model a process with so many relevant scales.

Recently a class of new models have been developed that outperform both the statistical and the dynamical models. These models making use of Machine Learning (an algorithm learning from training data to make a prediction for the future using test data) have been able to predict well beyond the SPB. Ham et al. 2019 have achieved predictions with significant skill ($r > 0.5$) for lead times up to 18 months with Transfer Learning (TL) using a vast amount of CMIP5 data as training set and normal weather data for the predictions [5]. Similar results also outperforming traditional prediction methods have been produced with a more simple method by Nooteboom et al. using the data reduction technique Autoregressive Integrated Moving Average (ARIMA) combined with Deep Learning (DL) on the residuals of the data [4].

However, the Machine Learning (ML) methods mentioned are limited in that they produce a single estimate without uncertainty measure of the variable of interest. Therefore Petersik et al. used Gaussian Density Neural Networks (GDNN) and quantile regression neural networks (QRNN) to obtain an estimate of the uncertainty. Notably Petersik et al. were able to get a similar result as Ham [5] but with much simpler design and using less features.

There is currently no complete explanation for the improvements in prediction of ENSO by using ML. A possible explanation is the existence of certain system dynamics that can be used to predict the future state of the equatorial pacific which could be inferred (or learned) from data by an artificial neural network (ANN). These dynamics would not be present in dynamical models since these are designed by humans, thereby explaining the difference in prediction skill between ML and traditional prediction methods.

Most machine learning methods provide little insight in the relationships or information that the model uses to make a prediction, which could improve

scientific understanding of ENSO while perhaps also aiding in furthering predictions. In this study we seek to gain such insight by training a ANN on distorted physics data (changing feedback mechanisms, coupling strength, wave speed of Kelvin and Rossby waves) produced by the Cane Zebiak 1987 model [2] to see if this results in a difference in prediction skill when tested on undistorted data. A change in skill implies a dependency of the predicting model on the distorted physical quantity thereby hopefully gaining some insight into currently unknown dynamics of ENSO.

2 Model & Methods

This section outlines the experimental setup for this paper, commencing with the generation of training- and test data using the Cane & Zebiak 1987 (CZ87) climate model [2], then training a model using the GDNN architecture in a distorted physics (DP) setup to produce the final results.

2.1 Cane & Zebiak 1987 model

The CZ87 dynamical model was developed in the 1980s by Mark Cane and Stephen Zebiak and is one of the first climate models to accurately resolve ENSO dynamics. This has provided a breakthrough in ENSO prediction by utilizing advances in computing power and increased understanding of ENSO dynamics, replacing Walker's method of comparing sea level air pressure anomalies between the Eastern Pacific and Northern Australia.

For the atmospheric dynamics the model uses a methodology as published by Gill in 1980 [14], using steady state linearized shallow water equations and assuming an equatorial beta plane, additionally using Rayleigh friction for energy dissipation[2]. The oceanic part of the climate model is based on a linear reduced gravity model with an added shallow frictional layer of 50m depth to resolve the wind driven currents found at the surface.

The CZ87 model depends on a number of parameters, of particular relevance to this research are: atmosphere ocean coupling (μ), wavespeed (δ), upwelling feedback (δ_s) and SST feedback (δ_{sst}) [2]. The atmosphere ocean coupling determines the exchange between the ocean and atmosphere while also determining the (in)stabillity of oscillations in the simulation. Whereas δ determines the velocities of the eastward Kelvin wave and westward Rossby wave by which the ENSO cycle is propagated ($\delta = 0$ being the fast wave limit). The oceanic response of the upwelling strength to the wind stress is parametrized by δ_s whereas δ_{sst} represents the advection feedback responsible for an increased zonal heat transport caused by an increase in SST.

A seasonal cycle is implemented by adding a sine term with a one year period to the ocean atmosphere coupling parameter $\mu = \mu_0(1+0.25\sin(2\pi t/T))$ thereby varying the coupling by 25% throughout the year. The variation of μ causes a change in the stability of the model thereby representing the presence of the SPB

as a period of high sensitivity to perturbation making it a difficult to predict period.

2.2 Gaussian Density Neural Network

The Gaussian Density Neural Network (GDNN) is a type of artificial neural network (ANN) designed by Petersik [10] based on earlier work by Lakshminarayanan at Google[1] on Deep Ensemble (DE) neural networks. Gaussian Density refers to the networks purpose of predicting a gaussian distribution by producing both a mean and standard deviation as output. Somewhat confusingly the GDNN consists of an ensemble of ANN's whose outputs are ultimately averaged in a final result, perhaps GDNN Ensemble or Gaussian Ensemble Neural Network would be more fitting. The DE methodology was developed as a more mathematically rigorous and less computationally expensive alternative to using a Bayesian Neural Network (BNN) for predicting a distribution (or value and uncertainty).

The variable to be predicted (or target variable) is the ONI at a (lead) time in the future. The features used in the GDNN are partly as described by [10]: ONI, network graph connectivity metric c_2, adjusted Hamming distance $\mathcal{H}^*$ (measure of change in the network graph) and a seasonal cycle (SC) in the form of a cosine. The warm water volume (WWV, volume of water above the 20 °C thermocline) is not available in the output of CZ87 therefore the thermocline height (h) itself was used in this paper. All feature datasets are normalized before training.

Network graph metrics as mentioned in the previous paragraph are calculated by first creating a graph in which the nodes are gridpoints of the SST field timeseries. Edges are then added to the graph between all the highly correlated nodes ($r > 0.97$) to get network metrics series, a timeseries of network connectivity. The metric c_2 is defined as the fraction of nodes that are connected to only one other node by an edge, c_3 is the fraction of nodes connected to exactly two other nodes by an edge. $\mathcal{H}$ is a measure of the change of the graph between two adjacent times in the network metrics timeseries. Research [4] found that before the onset of the positive phase of ENSO there is often a sudden increase in connectivity (increasing c_2) called a percolating transition.

Training the GDNN consists of a number of ensemble members that are trained in parallel. Each of the members is trained for 100 iterations over 500 epochs with a batch size of 100. The training starts with a random selection of hyperparameters within bounds defined by the user and is then optimized using the ADAM algorithm [15] with user specified learning rate, drop out and gaussian noise. The resulting ensemble members will later be called upon to each predict a mean and standard deviation of the target variable, these predictions are then averaged over the ensemble for the final prediction.

2.3 Distorted physics

In order to explore the behaviour and sensitivities of the GDNN method this research uses several similar setups of the CZ87 model using a method called Distorted Physics (DP). DP refers to the varying of parameters in the dynamical equations of (in this case) a model, running a simulation and analysing the change in output of the model. Originally this was done by Neelin [11] to better understand the representation of the SST mode and wave modes and their influence on inter annual coupled oscillations (such as ENSO). In this research the DP experiments are performed similarly as in the original paper however an additional step is added by using the output of the GCM as training and testing data for the GDNN model.

The experiments broadly consist of two steps, first the CZ87 model is ran for standard parameter values to produce *reference case* data, then CZ87 is ran again but for a range of values around the standard parameter value (shown in table 1) to get the *distorted* data. This ultimately results in four different kinds of datasets: reference case, distorted wavespeed, distorted thermocline feedback and distorted SST damping. There are are no simulations where more than one parameter is distorted at the same time.

In the second step the distorted datasets are used as training data for the GDNN whose performance is then determined by using the reference case as the test set. As a consistency check the GDNN is also trained on reference case data and then tested on reference case data, this should produce the highest performance because this GDNN is tested on data it has already seen. GDNNs trained on distorted datasets without a SC have test sets without a SC and vice versa. The seasonal cycle feature is present in all the GDNNs however without a SC in the data this feature should be irrelevant [1].

name	μ	δ	δ_s	δ_{sst}	SC
distorted wavespeed	2.7	0.5-1.5 (0.1)	0.3	1.0	On
distorted wavespeed	2.7	0.5-1.5 (0.1)	0.3	1.0	Off
distorted upwelling	2.7	1.0	0.1-0.6 (0.05)	1.0	On
distorted upwelling	2.7	1.0	0.1-0.6 (0.05)	1.0	Off
distorted SST feedback	2.7	1.0	0.3	0.5-2.0 (0.1)	On
distorted SST feedback	2.7	1.0	0.3	0.5-2.0 (0.1)	Off
reference (SC)	2.7	1.0	0.3	1.0	On
reference (no SC)	2.7	1.0	0.3	1.0	Off

Table 1: Parameter settings of CZ87 used to generate the data used in the DP experiments, step size shown within brackets. Parameter ranges are chosen to cover roughly a 50% increase and decrease compared to reference value, stepsize is chosen to get around 10 points within this range. The parameters are from left to right: name, coupling strength, wave speed, upwelling feedback, SST damping, seasonal cycle.

[1] this could have been another consistency check

The Deep Ensemble Model originally uses a negative log likelyhood loss function [10] (y being the predicted variable given another variable x).

$$\mathcal{L} = -\log(y|x) \tag{1}$$

However this function applies poorly to a model that produces an output estimate of both mean and standard deviation such as a Gaussian Density Neural Network (GDNN) because there are now two outputs of the model that also should be weighted differently in the loss function. To this end the following approximation of the loss function is proposed [1].

$$\mathcal{L} = \frac{\log(\sigma_\theta^2(x))}{2} + \frac{(y - \mu_\theta(x))^2}{2\sigma_\theta^2(x)} + \text{constant} \tag{2}$$

In the above equation, μ_θ and σ_θ are the mean and standard deviation given as output of the model for a given hyperparameter settings θ. However this loss function is not bounded to positive values in the current form and can actually go to minus infinity. The paper by Lakshminarayanan [1] proposes to apply a softplus function to the variance predicted by the model to ensure positiveness, however this does not put the minimum of the loss at zero. To bound the loss at 0 we can add the inverse of the softplus function at x = 1 to the argument of softplus as suggested by Jia Fulow [13]. The softplus function is similar to the better known ReLu activation function but smooth and described by $f(x) = \log(1 + \exp(x))$, the inverse of softmax is called the softinv.

$$\tilde{\sigma}_\theta^2 = \text{softplus}(\sigma_\theta^2(x) + \text{softinv}(1)) \tag{3}$$

$$\tilde{\sigma}_\theta^2 = \log(1 + e^{(x + \log(e^1 - 1)}) \tag{4}$$

$$\mathcal{L} = \frac{\log(\tilde{\sigma}_\theta^2)}{2} + \frac{(y - \mu_\theta(x))^2}{2\tilde{\sigma}_\theta^2(x)} + \text{constant} \tag{5}$$

Equation 5 now has the desired property that $\sigma_\theta^2 = 0$ (perfectly predicted standard deviation) leads to $\tilde{\sigma_\theta}^2 = 1$ which then causes the first term of the loss to be zero ($\log(1) = 0$). Since the second term of the new loss function was already zero for $y = \mu_\theta$ (perfectly predicted mean) a perfect prediction will now result in a loss of exactly zero (given constant = 0).

3 Results

This section contains first results of the data produced by the CZ87 model and the network analysis in preprocessing, then some individual predictions from this data using ML are shown and finally results on the effect of DP on the ONI and the final resulting predictions using the DP data. Running the CZ87 model for distorted values of $\delta_s st$ was unsuccessful and has therefore yielded no results.

3.1 CZ87 results (reference case)

Shown in figure 1 are the network graph metrics for the reference case dataset. The c_2 values are small (compared to c_2 calculated from *in situ* data, not shown) for most of the domain. This means the correlations between nodes in the network can be either small or much higher. In our experiment we find the former is the case. This low node connectivity has led to an increase of the correlation threshold for nodes in the network to be considered connected from 0.95 to 0.99 compared to the original design of the GDNN. Increasing the correlation threshold results in a decrease of the amount of nodes that have more than two edges (c_3 and higher) thereby increasing the value of c_2.

The resulting c_2 in figure 1 is similar in shape to fig 2a from Nooteboom et al. [4] however notably less smooth, this is likely caused by the use of an even higher ($\epsilon = 0.9999$) correlation threshold. The predictive property of c_2 where a high c_2 precedes an El Niño event shown in Nooteboom is also found in 1 around the El Niño events from 1958 and 1978.

The $\mathcal{H}$ values are actually higher in the CZ87 data than in reality, which can be expected as the points in the timeseries are less correlated (more random) and which could be explained by the artificial climate being more noisy and random in general (see figure 1).

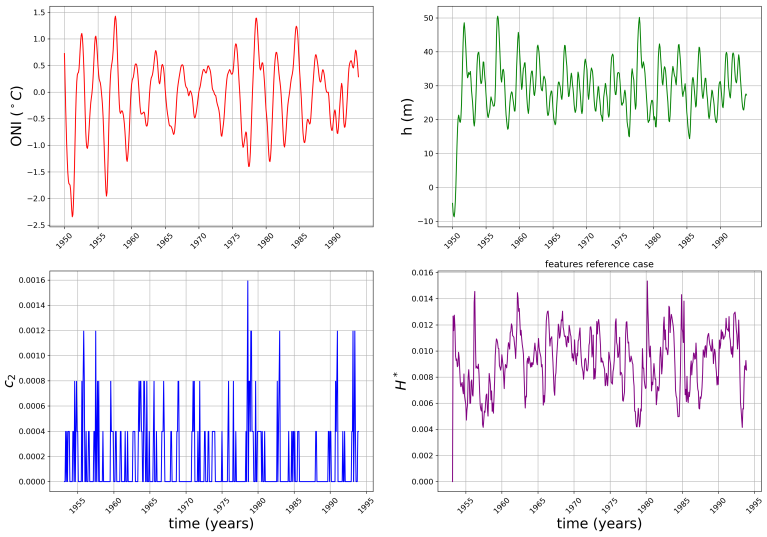

Figure 1: Network graph analysis results for the reference case dataset without SC . Shown metrics are: ONI (top left), thermocline hight h (top right), two edge node fraction c_2 (bottom left) , adjusted Hamming distance $\mathcal{H}^*$ (bottom right).

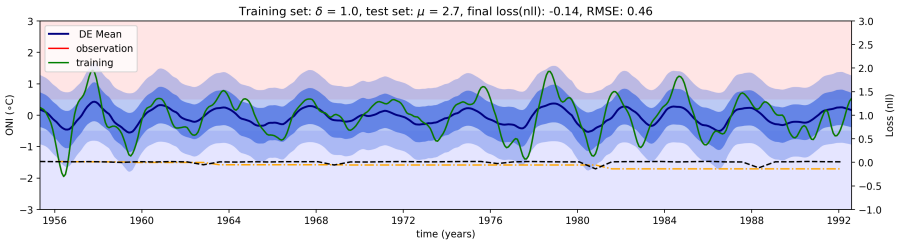

Figure 2: Predictions of the GDNN with test and training data from the reference case dataset. Since both training and observation data is the same only the latter is visible (shown in green). Predicted mean of the GDNN is shown as the dark blue graph, the predicted standard deviation is shown as dark blue shading, two standard deviations are shown in light blue shading. The best loss (orange) and average loss (black) is also shown for each iteration increasing from left to right (x-axis hidden).

Shown in figure 2 is a prediction of the ONI by the GDNN trained on the reference case data set (parameter settings shown in table 1). Most of the actual ONI values are within one standard deviation of the prediction and the mean behaviour of the ONI is generally captured by the model. This result is one of the best found in this project with an Root Mean Square Error (RMSE) value of 0.357, which is expected as training and testing data are the same in this case.

3.2 Distorted wavespeed (δ)

Figure 3 shows that changing the δ value in CZ87 causes changes in more general parts of the GCM such as in the ONI. Most noticably the amplitude of the oscillation becomes much smaller for $\delta < 1$, so much even that by definition only ENSO neutral conditions ($-0.5 < ONI < 0.5$) are present. Increasing δ above the reference value of 1.0 initially leads to an increase in the oscillations amplitude and then decreases again for higher values of δ. A staggering of graphs caused by change in periodicity is expected in this case because the ENSO period depends on the velocity of Rossby and Kelvin waves crossing the Pacific basin, however this is not evident from the figure.

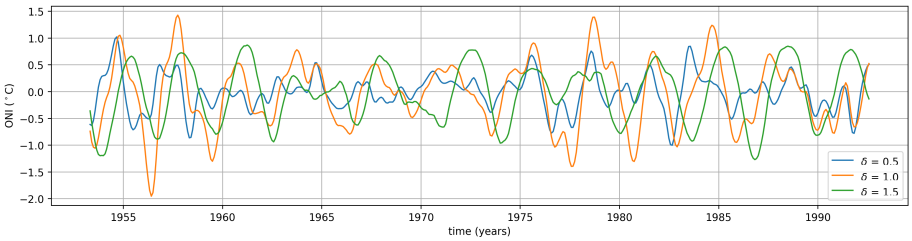

Figure 3: Several timeseries of ONI calculated from DP simulations using δ parameter values of 0.5, 10 and 1.5. with $\mu = 2.7$ and including a Seasonal Cycle

Figure 4 shows the GDNN performing relatively poorly on the $\delta = 1.2$ training dataset. The amplitude of the reference case data is mismatched in the first part of the dataset (until around 1968) and later there is a mismatch in the timing of the El Ǹino events (between around 1980 to 1990).

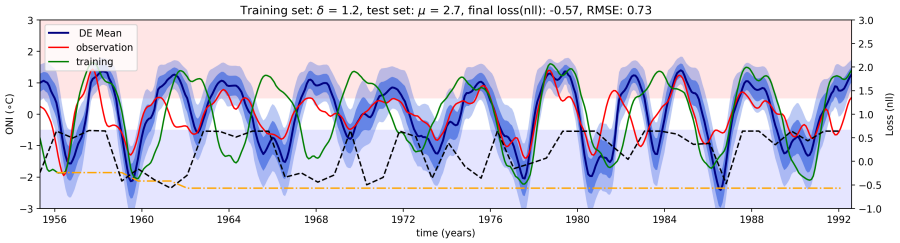

Figure 4: Predictions of the GDNN using distorted physics $\delta = 1.2$ as training data with reference case test data. Predicted mean of the GDNN is shown as the dark blue graph, the predicted standard deviation is shown as dark blue shading, two standard deviations are shown in light blue shading. The best loss (orange) and average loss (black) is also shown for each iteration increasing from left to right (x-axis hidden).

3.3 Distorted upwelling feedback (δ_s)

Figure 5 shows that the ONI's amplitude increases (decreases) for larger (smaller) values of δ_s. This behaviour is expected because the thermocline feedback is positive, enhancing the existing SST anomaly further and increasing its magnitude will thereby increase de amplitude of the ONI.

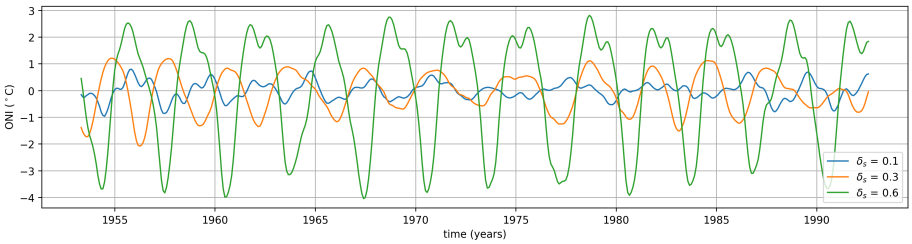

Figure 5: Several timeseries of ONI calculated from DP simulations using δ_s parameter values of 0.1, 0.3 and 0.6 using $\mu = 2.7$ including a Seasonal Cycle.

Figure 6 shows a regime where the GDNN still generalizes well between a training set with a larger amplitude compared to the test set. However in figure 7 this is no longer the case and the prediction overestimates the amplitude of the signal for most of the domain. The offsetting of the period is well accounted for in both cases more clearly so in the $\delta_s = 0.4$ case. The loss functions in figure 7 show only small improvements in the best loss found during training, illustrating the difficulty of fitting a model for this amplitude difference.

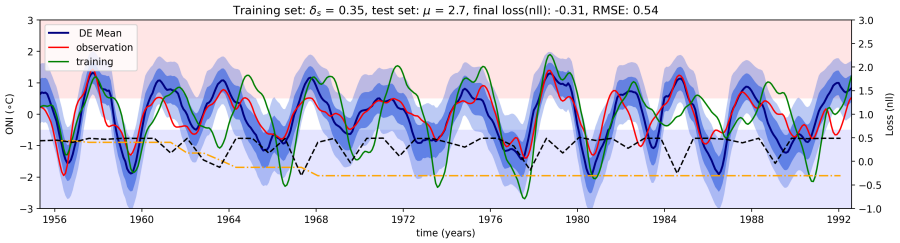

Figure 6: GDNN predictions using distorted physics $\delta_s = 0.35$ as training data with reference case test data. Predictions of the GDNN are shown as the dark blue graph, one standard deviation is shown as dark blue shading, two standard deviations are shown in light blue shading. The best loss (orange) and average loss (black) is also shown for each iteration increasing from left to right (x-axis hidden).

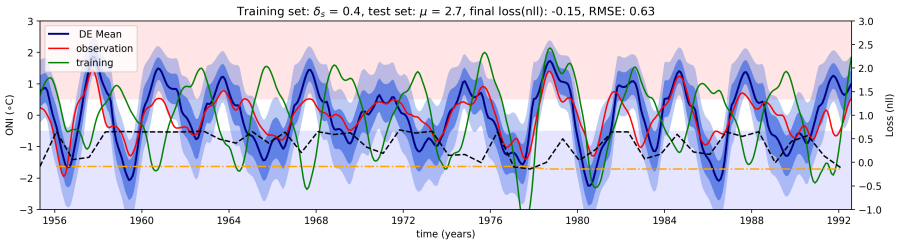

Figure 7: GDNN predictions using distorted physics $\delta_s = 0.40$ as training data with reference case test data. Predictions of the GDNN are shown as the dark blue graph, one standard deviation is shown as dark blue shading, two standard deviations are shown in light blue shading. The best loss (orange) and average loss (black) is also shown for each iteration increasing from left to right (x-axis hidden).

3.4 Machine Learning results

The experiments with DP trained GDNN's are assessed by quantifying the NN performance for each setup as shown in figure 8. Shown in all subfigures is an expected minimum in the loss and RMSE for the reference value of the parameter. Broadly similar behaviour is present in setups that are only different in the presence or absence of a SC.

Notably in the δ cases (fig 8a and 8b) there appears a monotonically increasing trend of the RMSE for values below the reference value (faster waves) while for larger delta values the increase in RMSE can be much larger and unpredictable.

Contrastingly in the δ_s cases (fig 8c, 8d) the RMSE skill deteriorates for smaller values of δ_s however not progressivly worsening while higher δ_s values lead to a monotonically decreasing RMSE skill in experiments with and without a SC.

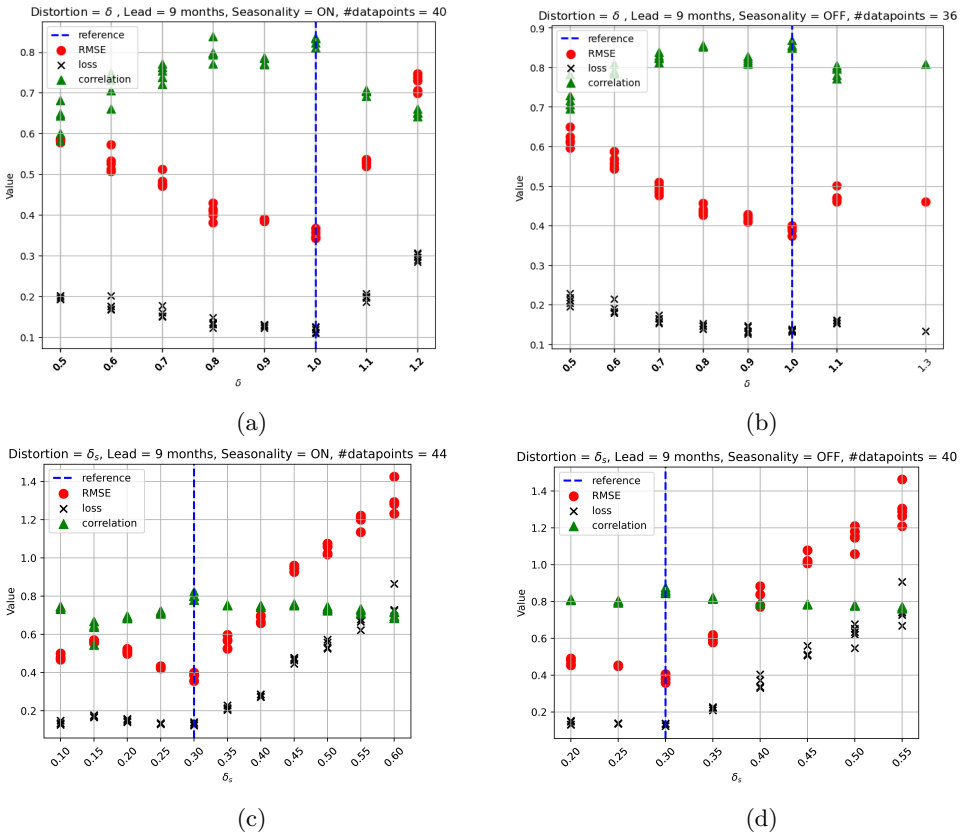

Figure 8: Metrics of DP run with a) a range of δ values including a SC b) a range of δ values without a SC c) a range of δ_s values including a SC d) a range of δ_s values without a SC. Generating data was unsuccesful in some cases leading to the absence of datapoints for some parameter values (when compared to table 1

The expected general behaviour is that the minimum of the RMSE and the loss is situated at the reference values of the parameters, while we should see a decrease in prediction skill for the distorted physics. This is exactly what we observe in figure 6, confirming the influence of the CZ87 physics on the prediction value.

In the case of δ (fig 8a, 8b)there is a linear decrease in skill when lowering the value of this parameter with and without a SC. This decrease can be explained by the ENSO signal travelling faster across the pacific ocean and back due to the higher wave speed thereby making a GDNN trained on this data to expect a shorter period of ENSO. Figure 3 shows that not only the period of ENSO in CZ87 is changed by varying δ but also the ONI is affected, this is of essential importance to the GDNN because the ONI's distribution is ultimately predicted

by the model. However the differences in ONI are relatively small for δ values below reference and become extreme only for much higher values. This change in behaviour could be explained by the model showing subcritical (supercritical) behaviour for low (high) values of δ. Although changing the wave speed should in theory not change the distance to the Hopf bifurcation and thereby the stability of the ONI. For higher values of δ the RMSE increases again but no trend could be reasonably discerned from the small amount of data points.

For δ_s (fig 8c, 8d) the RMSE increases linearly for higher than reference parameter values, this suggests harder to predict supercritical behaviour of the CZ87 model for these values. δ_s values below reference all have similar RMSE and do not show a clear trend which could be caused by the CZ87 model being in the subcritical regime. Contrary to the δ case this is actually expected because the δ_s parameter is known to directly affect the stabillity of the CZ87 climate model.

4 Summary & Discussion

Assessing the GDNN's performance when trained on DP data, this research finds that the model still performs consistently when the distortion leads to a change in period length (δ) or offset of the periodicity (δ_s). However a change in the ONI's amplitude in the training data (such as for higher than reference δ_s) is poorly corrected for, leading to a large overestimation of the predicted variable (e.g. $\delta_s = 0.40$, see fig 7). The model only tolerates a difference in amplitude between test and training dataset ONI if only a small distortion of the variable is used (e.g. $\delta_s = 0.35$, see fig 6). The ability to compensate for period but not amplitude could be explained by the relatively simple architecture of the GDNN. Whereas the former might only require a scalar addition to the input (or simply change of bias ($\vec{b}$) of the model) the latter would require some linear combination of (co)sines to be learned by the NN.

Further research might consider using a different climate model to generate the DP data because the CZ87 model (although computationally tractable) does not converge for all parameter settings that this project originally planned on using. Another climate model should also resolve ENSO to a meaningful degree however doing so dynamically (from equations) limits the possible research outcomes by the "garbage in, garbage out" (GIGO) principle. One might circumvent GIGO by using a climate model with a statistical ENSO representation (perhaps a NN) and then imposing the DP between every time step, thereby creating a climate model that might contain the missing ENSO physics and can also be distorted.

Another improvement might be made by tweaking the loss function to be more sensitive to the predicted mean thereby hopefully leading to more realistic results, currently there are some low amplitude training data sets that lead to quasi-horizontal lines as predictions.

In conclusion, we looked at the effects of DP on GDNN predictions and we found a maximum skill for the reference dataset confirming the influence of

the parameters on the predictive capabilities, leading the way to research more complicated distorted physics setups with ML methods.

References

[1] *Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles*. B. Lakshminarayanan, A. Pritzel, C. Blundell.

[2] *A Model El Nino-Southern Oscillation*. M. A. Cane, S. E. Zebiak, Monthly Weather Review, 97(3), 163-172, (1987)

[3] A. Timmermann, S. An, J. Kug, et al. *El Niño–Southern Oscillation complexity* . Nature, 559:535-545, 2018.

[4] P. D. Nooteboom, Q. Y. Feng, C. López, E. Hernández-García, H. Dijkstra *Using network theory and machine learning to predict El Niño*. Earth Syst. Dynam., 9, 969–983, 2018

[5] Y. Ham, J. Kim, J. Luo *Deep learning for multi-year ENSO forecasts*. Nature 573, 568–572, 2019.

[6] M. Reichstein, G. Camps-Valls, B. Stevens, et al. *Deep learning and process understanding for data-driven Earth system science*. Nature 566, 195–204, 2019.

[7] M.L. L'Heureux, K. Takahashi, A. B. Watkins, et al. *Observing and predicting the 2015/16 El Niño* Bulletin of American Metereological Society, July, 1363-1382, 2017

[8] A. G. Barnston, M. K. Tippett, M. L. L'Heureux *Skill of real-time seasonal ENSO model predictions during 2002-11* Bulletin of American Metereological Society, May, 631-651, 2012

[9] H. A. Dijkstra, P. Petersik, E. Hernández-Garcia, Cristóbal López *The Application of Machine Learning Techniques to Improve El Niño Prediction Skill*

[10] P. J. Petersik, H. A. Dijkstra *Probabilistic Forecasting of El Niño Using Neural Network Models* Geophysical Research Letters, 47-6, 2020

[11] J. D. Neelin *The slow sea surface termperature mode and the fast-wve limit: analytic theory for tropical interannual oscillations and experiments in a hybrid coupled model* Journal of the atmopsheric siences, 48(4) 584-606 (1990)

[12] Trenberth, Kevin National Center for Atmospheric Research Staff (Eds). Last modified 21 Jan 2020. "The Climate Data Guide: Nino SST Indices (Nino 1+2, 3, 3.4, 4; ONI and TNI)." Retrieved from https://climatedataguide.ucar.edu/climate-data/nino-sst-indices-nino-12-3-34-4-oni-and-tni.

[13] J. Fulow. Softplus and softminus (2019 July 11) https://jiafulow.github.io/blog/2019/07/11/softplus-and-softminus/

[14] A. E. Gill *Some simple solutions for heat-induced tropical circulation* Quarterly Journal of the Royal Meteorological Society, 106(499), 447-462.

[15] D. P. Kingma, J. Ba *ADAM: A Method for Stochastic Optimization* arXiv:1412.6980