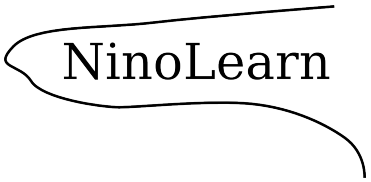
Master's thesis

Deep Ensemble and Encoder-Decoder neural networks for ENSO forecasting

Name: P. J. Petersik

Student number: 6207146

Supervisor: Prof. dr. ir. H.A. Dijkstra
Dated: August 15, 2019



Contents

Summary

This study introduces two novel statistical models for the prediction of the El Niño-Southern Oscillation (ENSO). The first model is termed Deep Ensemble (DE). This neural network method was recently developed by Lakshminarayanan et al. (2017). A DE is applied for the first time in the prediction of the ENSO. Instead of only predicting one value for Oceanic Niño Index (ONI), as it is done by other statistical ENSO models, the DE predicts the mean and the standard deviation of a Gaussian distribution. In this way, the forecast comes with an estimation of the predictive uncertainty, which is a novel feature in the prediction of the ENSO by statistical models. The predictor variables for the DE are chosen such that they represent different aspects of the ENSO dynamics. The memory component of the subsurface is included by the use of the warm water volume (WWV) of the equatorial Pacific. As a driver of more stochastic effects, the area averaged zonal wind stress anomaly in the west Pacific (WP) is added to the predictor variables. To include information about the interdecadal changes in the background state of the Pacific, the amplitude of the leading empirical orthogonal function (EOF) of the 5-year running-mean sea surface temperature anomaly (SSTA) field is used. Other predictor variables which are used are the dipole mode index (DMI) of the Indian Ocean Dipole (IOD), two variables from the theory of evolving complex networks as well as the ONI itself. The trained DEs show similar prediction skills as other statistical ENSO models do that are currently used in operational ENSO forecasting. As expected, the DEs assign low predictive uncertainties to forecasts with a small lead time. In contrast, for very long lead times, the DEs predict the climatological distribution of the ONI. Unfortunately, decadal variations in the estimated predictive uncertainty are not clearly visible within the forecasts. This can be attributed to the low amount of available training data.

The second model is a so-called Encoder-Decoder (ED) model which is inspired by the architecture of Autoencoders (AEs). The ED is used to predict the entire SSTA field in the Pacific ocean between 30°S-30°N and 120°E to 80°W. Therefore, the prediction provides information on the spatial pattern of the anomalies. The model shows generally weaker skills than the DE regarding the ONI. Hence, it has a weaker prediction skill in comparison to other statistical models. However, it is still remarkable that the ED can make skillful predictions given the little amount of data which was available for the training in respect to the complexity of the model. This shows that the bottleneck architecture of the ED effectively prevented overfitting.

Both models are analyzed onto their predictive skill during different decades and for different seasons. It is proven that the models have a good predictive skill for long lead times when the background state of the Pacific is relatively warm (El Niño-like decades), whereas they perform considerably worse for a colder background state (La Niña-like periods). Based on the findings of this study and of a literature review, a hypothesis is proposed to explain why it is possible to make meaningful prediction beyond the spring predictability barrier during El Niño-like periods whereas in La Niña-like periods this is usually impossible.

Chapter 1

General introduction

The inter-annual climate variability of the tropical Pacific is strongly dominated by the El Niño-Southern Oscillation (ENSO). Unusually high sea surface temperatures (SSTs) develop in the equatorial Pacific during an El Niño event. In contrast, during a La Niña event, cold sea surface temperature anomalies (SSTAs) are present in equatorial Pacific. Typically, the phase of the ENSO is quantified by the area-averaged SSTA in specific regions of the equatorial Pacific ocean. The most common region for the quantification of the ENSO is situated in the central Pacific between 5°N-5°S and 170°W-120°W. It is termed the NINO3.4 region (see Fig. 1.1a). The 3-months running-mean in the NINO3.4 region defines the Oceanic Niño Index (ONI) which is usually used to quantify and predict the ENSO. The time series of the ONI is shown in Fig. 1.1b. One can see recurrent warm (El Niño) and cold (La Niña) anomalies of various strengths with strong El Niño events appearing about every 15 years (1982/1983, 1997/1998, 2015/2016).
Two characteristics make ENSO forecasting very interesting for the climate and weather research community: On the one hand, the predictive horizon of the ENSO is by far greater than the one of the weather prediction. This is for one part because of the strong autocorrelation of the ENSO dynamics. For another part, there is a memory component within the ENSO system that makes it (sometimes) possible to make skillful prediction for a year ahead (e.g. Balmaseda et al., 1995). On the other hand, the ENSO has distinct influences on the climate around the globe (e.g. Diaz et al., 2001). This combination of large predictive horizon and global influence makes ENSO forecasting a very promising field of research.
In this study, two novel statistical models for the ENSO prediction are introduced. The remainder of this section will give a short introduction into the theory and physics of the ENSO. It highlights important aspects of the ENSO dynamics to motivate the choice of certain predictors for the later applied models. Then, Chapter 2 introduces the data which is used throughout this study. Moreover, a data analysis is conducted to highlight important aspects of the ENSO dynamics. Afterwards, the first model which is based in the family of *artificial neural networks* (ANNs), a so-called *Deep Ensemble* (DE), is applied for the ENSO prediction in Chapter 3. In contrast to other statistical models that have been applied for ENSO-forecasting, the DE is able to estimate the predictive uncertainty of a particular forecast. Although this is a very common feature for forecasts made by dynamical models, i.e. by ensemble forecasts, this is a novel feature for forecasts made by statistical models. The second model is introduced in Chapter 4. This other kind of ANN is termed Encoder-Decoder (ED). The model architecture is inspired by the architecture of Autoencoders (AEs). Instead of just predicting a single index value, the ED model predicts the entire SSTA field in the Pacific ocean between 30°S-30°N and 120°E-80°W. Finally, the research is concluded with an outlook for further research in Chapter 5.
During the work on this Master's thesis, a research framework for statistical ENSO prediction, which is named `NinoLearn`, was developed. The Appendix A introduces some of the core concepts and ideas behind this open-source Python package.

1.1 ENSO dynamics

This section introduces the most relevant theories of the ENSO dynamics. These theories can accurately describe certain aspects of the ENSO. However, it appears that there are periods in which certain theories are better suited to explain the ENSO dynamics than others. In the end of this section, I motivate a hypothesis how different theories could be unified into one theory.

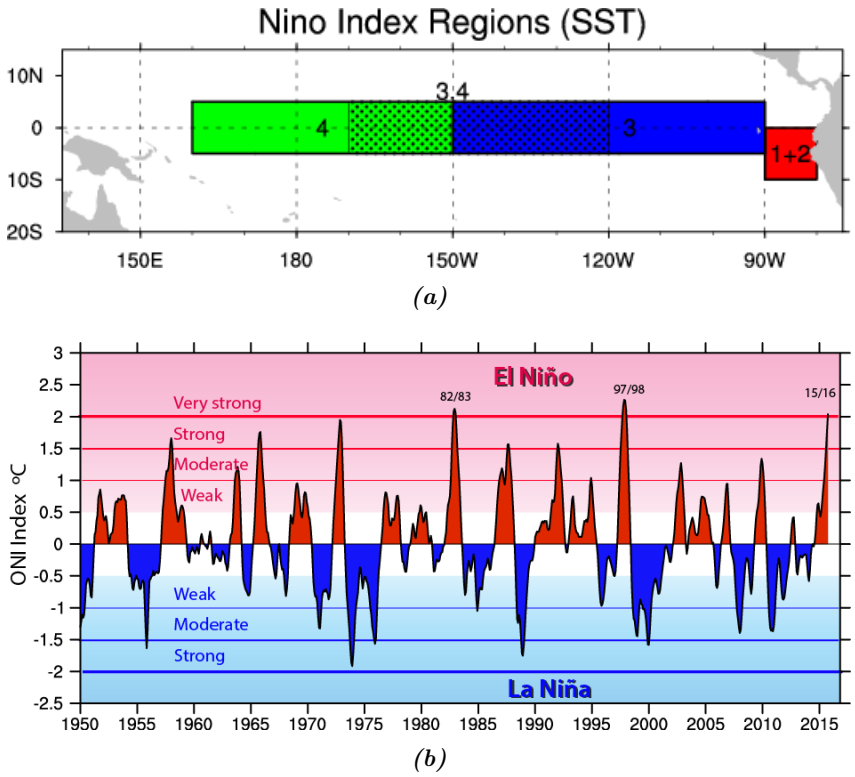


Figure 1.1: ***(a)*** *The NINO1+2 (red), NINO3 (blue), NINO4 (green) and NINO3.4 (dotted) regions for which the common indices are computed.* ***(b)*** *Time series of the ONI. (Trenberth and National Center for Atmospheric Research Staff (Eds), 2019)*

1.2 Feedbacks

For the ENSO, three feedbacks can be identified which play key roles in the development of an/a El Niño/La Niña. To understand these feedbacks, at first the mean state - also-called background state - of the equatorial Pacific has to be introduced. The presence of the mean state is a combination of an external atmospheric forcing, which would be present without a coupling of the atmosphere and the ocean, and internal feedbacks (Neelin and Dijkstra, 1995).
For the mean state, initially easterly trade winds (externally forced) are present over the equatorial Pacific. They lead to local upwelling by diverging Ekman transports around the Equator. Additionally, the easterly winds pile up water in the west Pacific (WP). This causes the thermocline to tilt such that it is shallower in the eastern Pacific (EP) and deeper in the WP. The thermocline is a thin layer of strong vertical temperature gradients in the ocean which separates the relatively warm upper mixed layer from the colder deep waters. The shallow thermocline in the EP combined with the Ekman upwelling gives rise to the relatively colder SST in the EP in respect to the WP. One refers often to the colder EP as the *Cold Tongue*. On the other side of the Pacific ocean, relatively warm water is brought to the surface by the Ekman upwelling because of the deeper thermocline. This gives rise to the so-called *Warm Pool* in the WP.
However, the external component of the trade winds account only for a small part of the mean zonal wind stress over the equatorial Pacific (Neelin and Dijkstra, 1995). An internal component arises due to the zonal SST gradient which is initially induced by the externally forced trade winds. Because of the SST differences between the cold EP and the warm WP, an additional surface easterly wind component is generated by the thermally-induced pressure difference between the EP and the WP. This additional easterly wind further cools the EP by the upwelling of cold water and hence strengthens the temperature difference. This is the so-called *Bjerknes feedback* (Bjerknes, 1969). The amplification of the internal wind forcing is limited due to various reasons. First, the cooling of the EP by the Ekman upwelling of cold water is limited by the constant deep water temperature. Second, the negative feedback by the so-called *thermodynamic damping* weakens the amplification of the SST gradient by two effects:

1. The warm SSTs in the WP (EP) lead to increased (decreased) evaporation and more (less) clouds over the warm WP (cold EP). Increased evaporation cools the warm WP and decreased evaporation

warms the cold EP.

2. More (less) clouds reduce (increase) the incoming short-wave radiation.

After the background state is now introduced, the relevant positive feedbacks for the ENSO can be explained. Already for the mean state the positive Bjerknes feedback was introduced. This feedback does not only play a role for the mean state but as well for the development of anomalies. For a better understanding of the Bjerknes feedback, it can be split into three components. The first component is the *thermocline feedback*. Assume for simplicity that the background upwelling stays constant. Further, let there be positive SSTAs in the EP. This will cause westerly wind anomalies to develop over the equatorial Pacific that will deepen the thermocline in the EP. Because of this, relatively warm water is brought to the surface in the EP by the background upwelling which leads to an amplification of the SSTAs.
Another feedback is the *zonal advection feedback*. Starting with positive SSTAs in the EP that cause westerly wind anomalies to develop over the equatorial Pacific. Now, the zonal currents in the ocean have a positive anomaly (flow direction eastward) due to the wind anomalies. This advects warm water from the West towards the EP which leads to an amplification of the SSTAs. Note, that the easterly anomalies that are induced to the east of the SSTAs have no effect on the SSTAs because they are situated on land. However, even when the positive SSTAs would be present in the central Pacific (CP), the zonal advection feedback may still amplify the SSTAs. The explanation to this is related to the findings of Clarke (1994), which indicate that the wind responses to positive SSTAs are for most part westerly wind anomalies (see the following Section 1.3 for a more detailed explanation of this hypothesis). Therefore, just weak easterly wind anomalies develop to the east of the SSTAs. This means, that the induced advection by the eastward currents to the west of the SSTAs overcompensate the advection by the westward currents to the east of the SSTAs. Hence, positive SSTAs in the CP can be additionally amplified by the zonal advection.
The third feedback is the *upwelling feedback*. Positive SSTAs in the EP with their westerly wind anomalies over the equatorial Pacific lead to a weaker upwelling. Hence, less cold water is brought to the surface which is a positive amplification of the warm SSTA.
The mentioned feedbacks are varying in their importance throughout the equatorial Pacific. In the EP, the thermocline and the upwelling feedback are most effective because of the shallow thermocline in this region. In contrast, the zonal advection feedback is most important in the CP. This is because the strongest SST gradients in the background state are present in this region (e.g. McPhaden, 2002; Chen et al., 2015). More details on the feedbacks in the equatorial Pacific can be found in, e.g., Dijkstra (2008).

1.3 Equatorial waves in the atmosphere and ocean

An important component of the ENSO dynamics is the influence of equatorial Kelvin and Rossby waves in the ocean and the atmosphere. In the ocean, Kelvin and Rossby waves are generated by wind stress anomalies. For instance, a positive zonal wind stress anomaly excites eastward travelling Kelvin waves and westward travelling long Rossby waves. These eastward travelling Kelvin waves raise the sea surface height and depress the thermocline. Therefore, they are called *downwelling* Kelvin waves. In contrast, in the descrobed case Rossby waves depress the sea surface and raise the thermocline. That is why they are called *upwelling* Rossby waves. Note, for an easterly wind stress anomaly the Kelvin wave would be upwelling and the Rossby wave downwelling. The Kelvin waves need about 3 months to cross the Pacific ocean, whereas the fastest Rossby waves (first mode) need about 9 months. For a more thorough mathematical introduction to oceanic equatorial waves, see for instance Dijkstra (2008).
Oceanic equatorial waves play a key role in the ENSO dynamics because they can set off and damp the thermocline feedback, especially in the EP. In the delayed oscillator, the downwelling Kelvin waves lead to positive SSTA anomalies in the SSTA upon their arrival. However, the westward travelling upwelling Rossby waves get reflected at the western boundary of the Pacific into upwelling Kelvin waves that again raise the thermocline upon their arrival in the EP (Suarez and Schopf, 1988).
While the contribution of oceanic Kelvin and Rossby waves for the ENSO dynamics is already included into conceptual models, the role of atmospheric Kelvin and Rossby waves still is a matter of research and debate. Atmospheric Kelvin and Rossby waves can be excited by convection in the atmosphere. Again, Kelvin waves travel eastward and Rossby waves westward. The phase velocity of Kelvin waves is with 10-50 $\mathrm{m\,s^{-1}}$ considerably faster than in the ocean. The fastest Rossby wave has a phase velocity of 1/3

of the Kelvin wave. For a comprehensive treatment of atmospheric Kelvin and Rossby waves, see i.e. Wheeler and Nguyen (2015).
Atmospheric Kelvin and Rossby waves are of special importance for the ENSO because they transport wind stress anomalies over great distance and hence influence the ocean far from the site of their excitation. One important aspect was highlighted by Clarke (1994). Traditionally it is assumed that anomalous winds over the equatorial Pacific are mainly forced by SSTA gradients. However, this stands somewhat in contradiction to the observations during a usual El Niño event for which strong winds are just present in the western half of the Pacific, whereas the SSTA gradients are present throughout the entire ocean (see Chapter 2 for a more detailed analysis).
The convective response during a warming event is usually strongest around the warm pool edge. Clarke hypothesizes that the convection mostly excites atmospheric Rossby waves which are accompanied by westerly winds along the Equator. In contrast, the convection does only excite weak Kelvin waves (with easterly wind anomalies at the Equator)because of the strong mismatch between the meridional extent of the convection and the Kelvin wave structure. In specific, the meridional scale of atmospheric Kelvin waves ($\approx 4500\,\mathrm{km}$) is by far greater than the typical meridional scale of the equatorial convective anomalies ($\approx 1000 - 2000\,\mathrm{km}$). Note, that little other research was done regarding the theory that convection is the main driver of the westerly anomalies during an El Niño event. However, findings by Zebiak (1990) support this view. Herein, Zebiak notes that a parameterization for convection in an earlier study (Zebiak, 1986) led to too strong easterly winds ("the easterly wind problem") because the meridional extent of the convection was too large. Due to the meridionally wide convection, Kelvin waves with eastward anomalies were effectively excited. This shows, that the meridional extent of the convection is a decisive factor for the strength of the excited atmospheric Kelvin waves.
The asymmetric excitation of atmospheric Kelvin and Rossby waves was used in Clarke (2014) to point out a convective ocean-atmosphere instability which is the topic of the next section.

1.4 A convective ocean-atmosphere instability

Based on the theory of convection being the main driver of westerly wind anomalies, Clarke (2014) describes an ocean-atmosphere instability that he claims to be the "engine" for the ENSO. For this, it is important to realize that the convective response to SSTAs strongly depends on the background SST state. A warmer background state favours a stronger convective response to the same SSTA because of nonlinear relationships such as the Clausius-Cleyperon relation for the water vapour pressure. This is the reason why convection during an El Niño event is most often strongest at the warm pool edge in the western Pacific (WP). Now, because this convection is only accompanied by westerly winds (and just weak easterlies), the warm pool edge is pushed towards the East and hence the convection moves to the West, too. This means, a positive feedback atmosphere-ocean feedback is present. An illustration of the convective ocean-atmosphere instability is given in Fig. 1.2.
Interestingly, this feedback might be a reason for the phase-locking of the warming events. During boreal winter, the feedback becomes weaker because the SST background state along the Equator becomes colder. Therefore, the convective response to the SSTAs becomes weaker and so do the westerly wind anomalies. This causes the SSTAs to decrease because damping effects such as Newtonian cooling are dominating.
Besides the role of atmospheric Rossby waves discussed above, Hameed et al. (2018) states that atmospheric Kelvin waves are an important component of the ENSO because they can have a self-limiting effect on the EP during an El Niño event. This is because the excited Kelvin waves are accompanied by easterly wind anomalies that enhance the upwelling in the EP and hence counter-balancing the thermocline feedback (see Section 1.7 for more details).

1.5 Recharge oscillator theory

One highly popular ENSO theory is the recharge oscillator proposed by Jin (1997). Fig. 1.3 shows an illustration of the different phases of the recharge/discharge oscillation. The thermocline and upwelling feedbacks play the key role in this theory. Starting with a warm anomaly in the EP (Phase I in Fig. 1.3), SSTAs are amplified by the thermocline and upwelling feedback. The westerly winds over the Pacific have their maximum around the Equator and are weaker towards the Pole. Hence, a positive (negative) torque is exerted onto the water column north (south) of the Equator which induces a poleward Sverdrup transport. Therefore, water from the warm mixed layer is transported out of the equatorial Pacific which

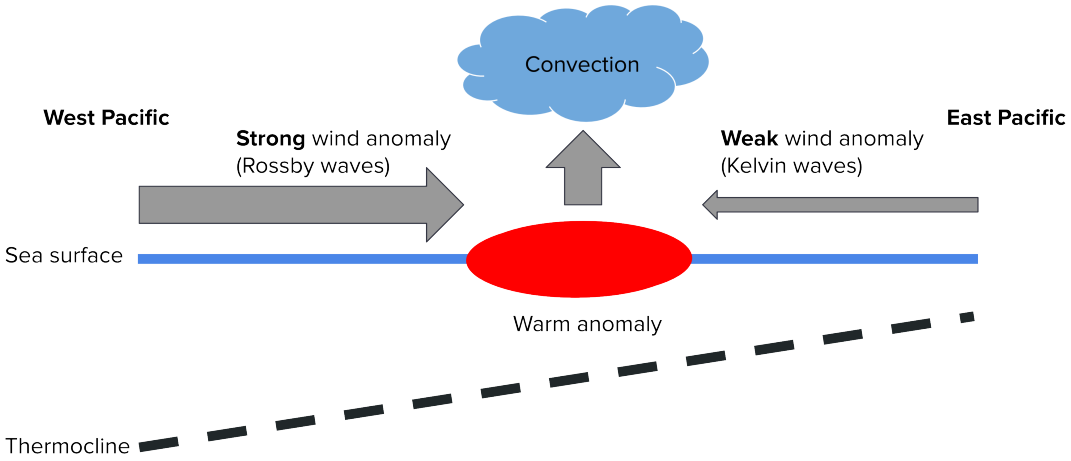


Figure 1.2: *Illustration of the ocean-atmosphere instability as proposed in Clarke (2014).*

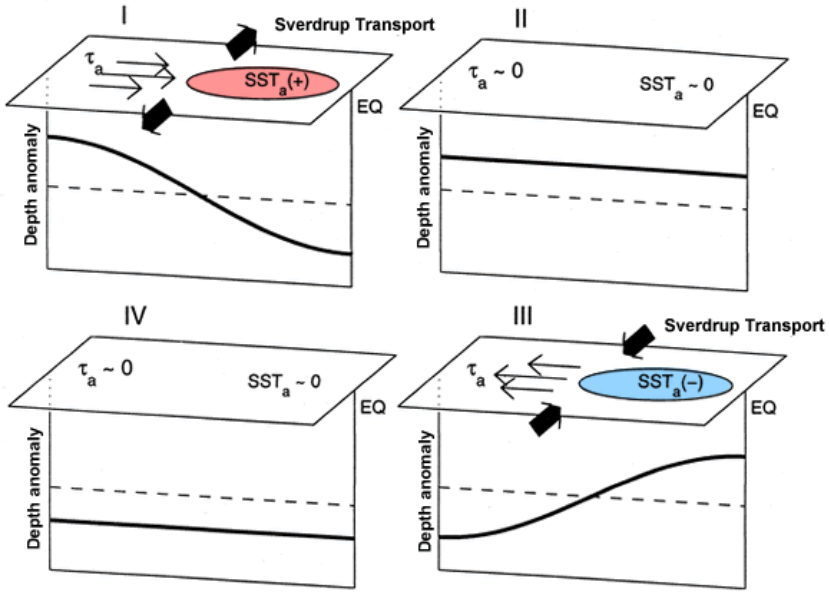


Figure 1.3: *Illustration of the recharge/discharge oscillation from Meinen and McPhaden (2000).*

leads to a rising thermocline. This phase is the so-called *discharge phase*. Eventually, the thermocline will become so shallow that cold water is brought to the surface in the EP and the system overturns into the La Niña phase in which the Sverdrup transport will recharge the mixed layer in the equatorial Pacific with warm water (Phase II and III). At a certain point the thermocline will be deep enough that relatively warm water is brought to the surface and the ENSO overturns into an El Niño event (Phase VI and Phase I).

In the recharge oscillator model in Jin (1997), a Hopf bifurcation is present that gives rise to the oscillation. If the coupling between the ocean and the atmosphere is strong enough, oscillations can grow where growth is limited by nonlinear processes. However, even if the coupling strength is too weak to sustain the growth of oscillations, stochastic forcing by the wind stress and heating can lead to the development of oscillations.

One variable that is closely linked to the recharge oscillator mechanism is the warm water volume (WWV) which is the volume of water above the 20°C isotherm (D20) in the equatorial Pacific (5°S to 5°N and 120° to 80°W). Research by Meinen and McPhaden (2000) indicated that the WWV has a strong correlation with the NINO3 index for 2-3 seasons ahead for the period between 1980-1999. However, Bunge and Clarke (2014) show that lead time for the strongest correlation of the WWV substantially decreased from 1999 onwards which diminished the prediction skill of the WWV for the ENSO.

Bunge and Clarke find that the first EOF (EOF1) of the D20 describes a zonal “tilt” with changing

sign across the ocean. The amplitude of the EOF1 is in phase with the ENSO indices. In contrast, the second EOF (EOF2) is a basin wide component with equal sign over the entire equatorial Pacific. Bunge and Clarke (2014) showed that during 1980-1998, the EOF2 was strongly correlated with the WWV ($r = 0.98$). However, the correlation between the EOF2 and the WWV was considerably weaker during 1999-2010 ($r = 0.61$). In their study, it is shown that the tilt mode had a non-zero contribution onto the WWV during 1999-2010 by having a more asymmetrical structure over the Pacific Ocean. The likely reason for this was a westward shift of the inter-annual wind forcing after 1998 due to decadal variations. The non-zero influence of the EOF1 (which has an amplitude that is in phase with the ENSO indices) onto the WWV, however, decreased the lead time of the WWV. This might explain why the WWV was a weaker predictor for the ENSO during 1999-2010.
The changing predictive skill of the WWV might be, as well, explained by a hypothesis of Clarke and Zhang (2019). Clarke and Zhang claim that the recharge oscillator theory neglects an important part of the ENSO dynamics, namely the zonal flow acceleration. By adjusting for this flaw, they claim that they can explain why the WWV has changing lead times on the ENSO indices. In their explanation, they show that the WWV has two main contributions. The first contribution comes from the zonal wind stress and the second one from the zonal flow. Whereas the wind stress term is strongest correlated with the NINO3.4 index for about a 0-months lead time[1], the zonal flow acceleration has a lead onto ENSO. Hence, only if the contribution of the zonal flow acceleration term to the WWV is relatively large, WWV is a predictor for ENSO. Based on this argument they show that the contribution of the zonal acceleration term onto WWV was relatively weak in respect to the wind stress term during La Niña-like periods (1959-1973 and 1999-2010) in which the long-term mean warm pool edge is shifted towards the West. Because of this diminished contribution of the zonal acceleration term onto the WWV, the WWV was a weak predictor during these periods.
The explanation of Clarke and Zhang (2019) sounds reasonable at first sight. However, without rigorous mathematical treatment, I want to point out an inconsistency in the study of Clarke and Zhang (2019). In their entire study, they did not discuss how the contribution of the Sverdrup transport to the WWV is related to their findings. This is because they do their analysis only for a simplified system that is located exactly at the Equator. This makes them conclude that the effect of the zonal wind stress is just the tilting of the thermocline. This process is, as one would expect, in phase with the ENSO indices and does not lead it (zonal tilt evolves simultaneously with the temperature anomalies). Based on this finding, they argue that there must be another contribution to the WWV which can explain the lead of WWV onto the ENSO indices. They argue that the zonal flow acceleration is this contribution because it has a lead correlation onto the ENSO indices. However, to my understanding it is questionable and incomplete to argue on the basis of neglecting the influence of the Sverdrup transport that the zonal flow is the core reason why WWV can lead the ENSO indices.
Despite the fairly questionable explanation of Clarke and Zhang (2019), it is still important to point out that the WWV has a changing lead time onto the ENSO indices with a larger lead time during El Niño-like periods and smaller lead time during La Niña-like periods. The changing lead time was most likely one of the reasons why statistical models underperformed between 2002-2011 in the study of Barnston et al. (2012) in comparison to their reported skill.

1.6 ENSO diversity

In scientific literature it is hypothesized that El Niño comes in different *flavours* (Kug et al., 2009; Kao and Yu, 2009). The so-called eastern Pacific (EP) type, also named *canonical* or *Cold Tongue* El Niño, refers to an El Niño event that is located in the East of the equatorial Pacific. The other type is the central Pacific (CP), also called *El Niño Modoki* or *Warm Pool El Niño*, which has its maximum temperature anomaly around the dateline. These two types can be distinguished by multiple methods as summarized in Capotondi et al. (2014).
Whereas for the EP El Niño vertical advection of anomalously warm water plays a key role in the development of the anomaly, meridional and zonal advection is most important for the development of a CP type (Chen et al., 2015). This indicates that the thermocline and the upwelling feedback control the development of EP El Niños and the zonal advection feedback the development of CP types (An and Jin, 2000).
In the recharge oscillator theory, the thermocline feedback is the key mechanism that explains the recurrent occurrence of SSTAs. Therefore, it is an adequate theory to understand the development of

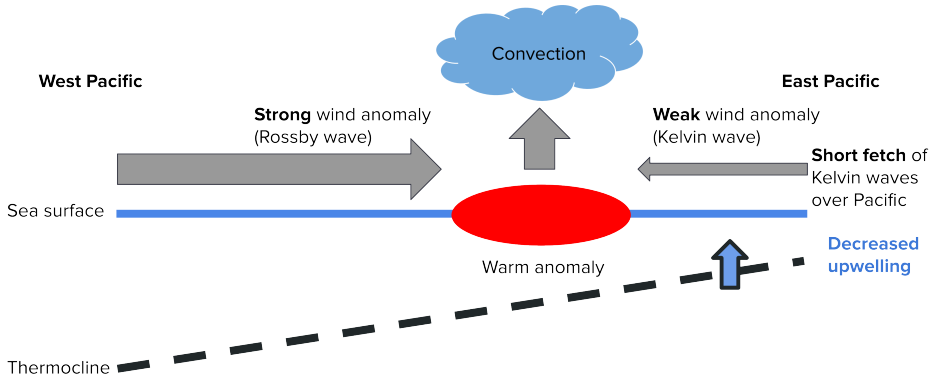
[1]Here, 0-months lead time refers to correlation between variables coming from the same time period.

EP type El Niños (and subsequent La Niñas) but may not be suitable to understand the occurrence of CP type El Niños. Kug et al. (2009) and Kao and Yu (2009) argue that the mechanisms behind a CP El Niño are substantially different to the mechanisms behind the development of an EP El Niño. In contrast to the EP El Niño, the sea surface height (SSH) anomaly field does not have a slope over the entire equatorial Pacific but rather has a positive bulge in the CP. Hence, the induced geostrophic currents nearly balance each other during a CP El Niño leading to no effective discharge of warm water from the equatorial Pacific and hence CP El Niños are rarely followed by La Niña events (Kug et al., 2009). Due to this, it is assumed that a CP El Niño is more a stochastic event than a phase of the recharge/discharge oscillation. Without further discussion, I want to point out that in contrast to this view, Ren and Jin (2013) explain the CP type in terms of a recharge oscillation.
Interestingly, Horii et al. (2012) associates the mentioned breakdown of the prediction skill of forecast models during the first decade of the 21st century (Barnston et al., 2012) with the frequent occurrence of CP El Niños during this period. This fits the assumption of CP El Niños being a rather stochastic event that is excited by weather noise such as westerly wind bursts (WWBs).
The synopsis of the literature on ENSO diversity and the changing prediction skill of WWV (described in the previous Section 1.5) indicates that the background state of the equatorial Pacific plays a crucial role for the ENSO dynamics and its predictability. It can be hypothesized that a more El Ninõ-like background state favours a more frequent development of EP El Niños whereas a more La Niña-like background state diminishes their occurrence. Already in An and Jin (2000), it was shown that the interdecadal climate shift in the late 1970s changed the ENSO dynamics. SST anomalies had a maximum in the CP during 1965-1975 whereas the maximum was situated in the EP for 1981-1991. This might be a reason why forecast models performed significantly better during the 1980s and 1990s than before (e.g. Tangang et al., 1997) and after (Barnston et al., 2012) this period.
Note, that this hypothesis is strongly influenced by the method how EP and CP El Niños are distinguished. For instance, on the hand, during the 1980s and 1990s, a more El Ninõ-like background state was present with the warm pool edge shifted relatively towards the East. In this period the WWV was a good predictor for upcoming El Niño events. Following Wiedermann et al. (2016) there have been 2 EP and 4 CP El Niños during this period. On the other hand, the 1960s and 1970s, as well as the 2000s, had a more La Niña-like background state with the warm pool edge shifted relatively towards the West. This period was accompanied by a poor prediction skill of the WWV. Whereas for the 2000s CP El Niños occurred more frequent (0 EP, 4 CP El Niños), the 1960s and 1970s showed an increased occurrence of EP El Niños (3 EP, 5 CP El Niños). This shows that the definition by Wiedermann et al. (2016) just supports the hypothesis from above for the period after 1980.
The observation of the decadal variability of the equatorial Pacific can be associated with the Pacific decadal oscillation (PDO). The PDO is defined by the amplitude of the EOF1 of the monthly SSTA in the north Pacific (20°N-70°N, Mantua et al., 1997). When the PDO is high, the SSTA in the equatorial Pacific is in average positive and therefore more El Niño-like, whereas when it is low SSTAs in the equatorial Pacific are negative and hence more La Niña-like (Fig. 1 in Newman et al., 2016). Newman et al. show that a high PDO favours EP El Niños, while during a low PDO, CP El Niños seem to occur more frequently (Fig. 13 in Newman et al., 2016).
To sum up, it is crucial to include information about the climatic background state into a predictive scheme. Therefore, it is important to train and evaluate the models onto periods of more La Niña and more El Niño-like periods, respectively.

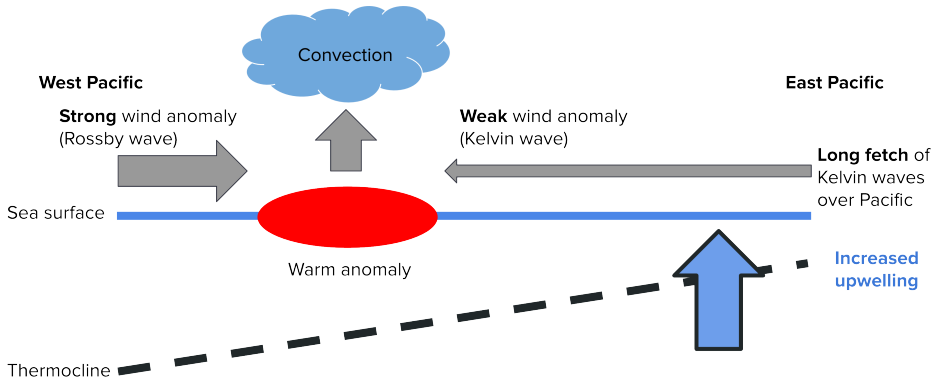
1.7 Self-limiting effect

Recently, Hameed et al. (2018) proposed a *self-limiting effect* for the development of positive SSTAs in the EP that appears due to enhanced convection over the equatorial ocean. In contrast to the common view of the El Niños types, the self-limiting effect makes the CP El Niño appear as the normal type and the EP El Niño as the special case[2].
As explained in Section 1.3, convective anomalies excite relatively strong westward travelling Rossby waves and weaker atmospheric eastward travelling Kelvin waves. The convection generates convergence in the atmospheric boundary layer. Because of this, the Kelvin waves transport easterly anomalies eastward and the Rossby waves westerly anomalies westward. Whereas the westerly winds from the atmospheric Rossby waves have been included into conceptual ENSO models, the effect of the atmospheric Kelvin waves has been neglected up until the study of Hameed et al. (2018). However, the easterly

[2]Hameed et al. (2018) call them *super El Niños*.



(a) El Niño-like period.



(b) La Niña-like period.

Figure 1.4: *Self-limitation during an El Niño event when the background state is* ***(a)*** *El Niño-like (warmer) and* ***(b)*** *La Niña-like (colder). In a warmer background state anomalous convection appears further to the East because the warm pool extends further to the East.*

anomalies damp SSTAs in the EP by increasing the Ekman upwelling. Hence, the developing convection introduces a negative feedback to the SSTA amplification in the EP. Therefore, it counteracts the effect of the downwelling oceanic Kelvin wave (excited by the westerly winds in the WP) in the EP. This self-limiting effect is the stronger the further West the convective anomalies are present. This is because the atmospheric Kelvin waves then have a longer fetch (distance over which they act) over the Pacific. Hence, self-limitation might be a reason why CP El Niños develop more often in La Niña-like periods because during these relatively colder decades convective anomalies are in average shifted towards the West which leads to a long fetch of the atmospheric Kelvin waves. In contrast, self-limitation is weaker during El Niño-like periods because the fetch of the atmospheric Kelvin waves is decreased during these periods. The difference in self-limitation during El Niño-like and La Niña-like periods is sketched in Fig. 1.4.
Hameed et al. (2018) argue that an additional positive contribution to the amplification of warm SSTAs is needed to explain why sometimes warm SSTAs can extend towards the EP. One probable contribution is the topic of the next Section 1.8, namely the Indian Ocean Dipole.

1.8 The role of the Indian Ocean Dipole

Not only the Pacific ocean contains an inter-annual mode of variability but also the Indian ocean (Saji et al., 1999). This mode is commonly referred to as the *Indian Ocean Dipole* (IOD). The positive phase appears with high SSTAs in the western Indian ocean and low SSTAs off the coast of Sumatra. The IOD is quantified by the so-called Dipole Mode Index (DMI). It is computed by the difference between the area-averaged SSTA of western (10°S-10°N, 50°E-70°E) and southeastern (10°S-0°N, 90°E-110°E) Indian ocean. The time series of the 3-months running-mean of the DMI is shown in Fig. 1.5 in comparison to the ONI.

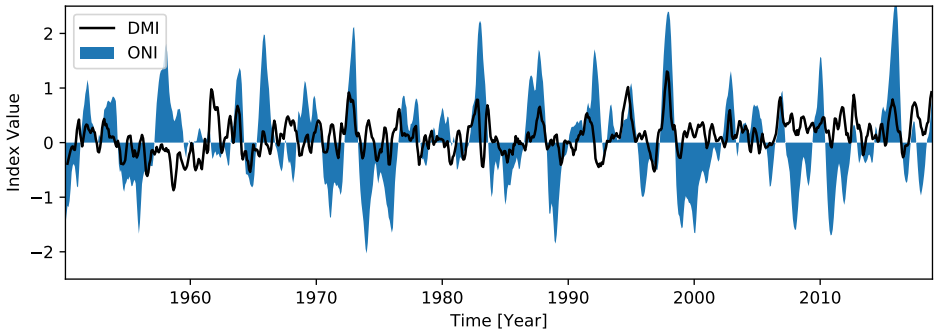


Figure 1.5: *Time series of the ONI (blue shading) and the 3-months running-mean of the DMI (black line).*

Very strong El Niños seem to co-occur with positive IOD phases. Hameed et al. (2018) show that a positive IOD causes an important additional westerly wind forcing in the WP which is necessary to compensate for the self-limiting effect and to generate an EP El Niño. The negative convective anomalies which are present over the eastern Indian ocean during the positive phase of the IOD generate this additional westerly wind forcing. Negative convective anomalies cause low-level divergence which excites atmospheric Kelvin waves with westerly wind anomalies to the west of the convection which travel towards the East (towards the WP). Hence, there is an additional component that drives the westerly winds in the WP. The IOD-intensified signal is then transmitted to the EP by the excitation of eastward travelling downwelling oceanic Kelvin waves. In the EP these downwelling Kelvin waves set off an intensified amplification of the SSTAs by the thermocline feedback.
Moreover, Izumo et al. (2010) find that the DMI has a significant prediction skill on the occurrence of El Niño/La Niña events about 14 months in advance. During a negative IOD anomaly, convection is enhanced in the eastern Indian ocean. Therefore, easterly wind anomalies are tranported towards the WP due to the excited atmospheric Kelvin waves. These easterly winds recharge the WP with warm water by an equatorward Sverdrup transport. When the IOD anomaly disappears, usually in November-December, the zonal wind anomalies disappear as well. Now, because of the recharged WP, the development of an El Niño can be set off (Wieners et al., 2016).

1.9 Summary of the theories

This chapter showed various important aspects of ENSO dynamics. The combination of two theories from the literature might be able to explain a large part of the observed ENSO dynamics. On the one hand, the recharge oscillator from Jin (1997) describes the deterministic part of the system, in which the subsurface already carries information about the future state of the ENSO. The thermocline feedback is the most important feedback mechanism for the recharge oscillator. Hence, it is well suited to explain EP El Niños. On the other hand, in the ENSO dynamics described in Clarke (2014), wind anomalies by stochastic weather noise in the WP can set off an ocean-atmosphere instability in which zonal advection is the key player for the development of the SSTAs. This theory might be more appropriate to explain the CP El Niños.
Depending on the importance of the more deterministic recharge oscillation and the more stochastic convective ocean-atmosphere instability, the predictability of the ENSO is altered as well as the flavour regarding El Niño events. The literature analysis revealed that the decadal variability of the background state of the Pacific ocean seems to strongly influence if the deterministic or the stochastic part of the system is “ruling”. During El Niño-like periods the self-limiting effect is weakened such that the thermocline feedback can effectively come into play and connect the surface to the subsurface which carries the memory of the system. This nudges the ENSO system towards a more deterministic nature and the more frequent development of EP El Niños. In contrast, during La Niña-like periods the self-limiting effect is enhanced such that the thermocline feedback is not effectively compensated by the enahnced upwelling. Therefore, the ENSO system is just weakly influenced to the memory of the subsurface. This, in turn, nudges the ENSO towards a more stochastic nature and more frequent development of CP El Niños.
This hypothesis is supported by the findings of Kirtman and Schopf (1998) which show that the ENSO is more (less) predictable when the 10-year mean of the SSTA in the equatorial Pacific is positive (negative).

Chapter 2

Data analysis

This chapter introduces the data which is used throughout this study (Section 2.1). Furthermore, a brief data analysis with focus on the ENSO diversity in Section 2.2 and the decadal variability in Section 2.3 is done to draw a clearer picture of these aspects of the ENSO dynamics.

2.1 Data

All data used in this study if note stated differently is retrieved for a period between January 1960 and December 2017 and in case of spatial-temporal data for the region between 30°S to 30°N and 120°E to 80°W. In addition, anomalies are computed by removing the seasonal climatology from the data corresponding to the reference period from January 1981 to December 2010.
The Oceanic Niño Index (ONI) is used in this study to capture the ENSO phenomena. The ONI is the 3-months running-mean of the NINO3.4 index which is the monthly average sea surface temperature anomaly (SSTA) in the NINO3.4 region (-5° to 5°N and 170° to 120°W) computed from the ERSSTv5 data set (Huang et al., 2017). The value of the ONI is assigned to the last month of the considered season, e.g the value of season December-January-February is assigned to February. This is done to make sure that later the prediction scheme is not accidentally using any data from the future to predict the ONI.
The warm water volume (WWV), which is the water volume above the 20°C isotherm in the area 5°S to 5°N and 120° to 280°E, is retrieved from the Bureau National Operations Centre[1] for the period between January 1980 and December 2017. The WWV is computed from data of moorings from the Tropical Atmosphere-Ocean Array (TAO), Argo floats and Expendable Bathythermographs (XBTs). To extend the time series backward, the WWV proxy index from Bunge and Clarke (2014) is used for the time between January 1960 and December 1979. In the following, the combination of the WWV and the WWV proxy index will be referred to as the WWV.
Another index value that is used is the Dipole Mode Index (DMI) of the Indian Ocean Dipole (IOD). It is retrieved from the website of the Earth System Research Laboratory's (ERSL) Physical Sciences Division (PSD)[2]. The DMI is the difference between the SSTA of the western (10°S-10°N, 50°E-70°E) and southeastern (10°S-0°N, 90°E-110°E) equatorial Indian ocean.
To compute the wind stress, the zonal and the meridional wind component is retrieved from the NCEP/NCAR reanalysis data set (Kalnay et al., 1996) on a 2.5°×2.5° grid. Then, the wind stress can be computed by

$$\tau_x = \rho C_D |\vec{U}| u \tag{2.1}$$

with ρ the density of the air, C_D the drag coefficient, and $\vec{U}$ the wind vector. However, in this study

$$\hat{\tau}_x = |\vec{U}| u \tag{2.2}$$

is used since it is proportional to the wind stress and can be easily derived from the wind field. Finally, the monthly Sea Surface Height (SSH) field from the ORAS4 data set (Balmaseda et al., 2013), the monthly Sea Surface Temperature (SST) from the ERSSTv5 data set (Huang et al., 2017) and the

[1] https://www.pmel.noaa.gov/elnino/upper-ocean-heat-content-and-enso
[2] https://www.esrl.noaa.gov/psd/gcos_wgsp/Timeseries/

monthly interpolated Outgoing Longwave Radiation (OLR) from the National Oceanic and Atmospheric Administration (NOAA, Liebmann and Smith, 1996) are used. Note, that the OLR is only available from June 1974 onwards with missing values during March - December 1978. In this study, the OLR is used to identify convection. Negative OLR anomalies are associated with anomalous strong convection because high clouds have cold cloud tops.

2.2 ENSO diversity

Literature introduced in 1.6 indicated that El Niño (and maybe even La Niña) comes in two flavours, namely a central Pacific (CP) and an eastern Pacific (EP) type. Wiedermann et al. (2016) proposed a new index based on the transitivity metric of an evolving complex network (ECN) of the SSTA field to classify an El Niño event into one of the two flavours which is used as the definition for CP and EP El Niños in this study. Fig. 2.1 shows composite plots for boreal winter (December-February) for the two flavours that occurred during 1974 and 2014. In this period, ten events are classified as CP El Niños (1977, 1979, 1986, 1987, 1991, 1994, 2002, 2004, 2006, 2009) and three (1976, 1982, 1997) as EP El Niños. Hence, based on this classification, the CP El Niño appears as the normal case and the EP El Niño as the exceptional one.
Fig. 2.1a shows, as expected, that CP El Niños have strongest SSTAs between 180°W and 120°W (coloured contours). Rather low anomalies are observed in the EP. As for the SSTAs, also the SSH anomalies (SSHAs, contour lines) show a bulge of positive anomalies in the CP which is shifted slightly to the east with respect to the SSTA bulge. Interestingly, OLR anomalies (OLRAs) do not coincide with the SSTAs but have a rather asymmetric structure over the equatorial Pacific as shown in Fig. 2.1b (coloured contours). Whereas strong positive OLRAs are found in the western Pacific (WP) at the western part of the SSTA bulge with a maximum around 180°W, slightly negative OLRAs are found in the East of the SSTA bulge. These convective anomalies coincide with anomalies in the zonal wind stress (contour lines) with strong positive anomalies in the WP and slightly negative anomalies in the EP.
For the EP classified El Niños, the composites show significant differences. Fig. 2.1c shows that both, SSTAs and SSHAs, have the strongest anomalies in the EP which weaken and even reverse towards the West. Moreover, the OLRAs and the zonal wind stress anomalies are rather symmetric over the equatorial Pacific with strongest anomalies around 150°W.
As hypothesized in the introduction, the recharge oscillator which is based on the dominance of the thermocline feedback in the EP appears as an incomplete conceptual model for the ENSO. It fits the appearance of EP El Niños with the deepened thermocline (positive SSHAs) in the EP and the overall positive zonal wind stress anomalies over the equatorial Pacific. However, for CP El Niñois, the thermocline feedback is weaker due to two reasons. First, the thermocline anomalies are less strong which is indicated by a weaker maximum SSHA. Second, the thermocline anomalies appear in the CP were the mean thermocline is significantly deeper than in the EP. Next, to the weaker thermocline feedback, the discharge of warm water by the Sverdrup transport is comparably small for CP El Niños because zonal wind stress anomalies reverse over the equatorial Pacific. This makes a transition of the ENSO into a La Niña phase in the following year less likely (Kug et al., 2009). All in all, this shows that other processes, for instance, the convective ocean-atmosphere instability (Clarke, 2014) with the accompanied zonal advection (Chen et al., 2015), might be more relevant for the development of CP El Niños.

2.3 Decadal variability

To investigate the decadal signal of the ENSO, the 5-year running-mean time series of the SSTAs, the OLRAs and the zonal wind stress anomalies are computed. Then, the linear trend of the SSTAs is removed to exclude the signal due to global warming. A principal component analysis (PCA) is performed for the detrended SSTA. Fig. 2.2 shows the leading EOF and the corresponding amplitude of the 5-year running-mean SSTA that accounted for 46.1% of the variance. One can see that the equatorial Pacific exhibits a decadal oscillation with warmer periods, more El Niño-like, during 1980-2000 and colder periods, more La Niña-like before and after this period. The spatial structure of the leading EOF shows a triangle-like structure that is symmetric over the Equator starting in the CP and extending towards the East.
Fig. 2.3a shows the regression of the 5-year running-mean OLRAs on the amplitude of the EOF1 5-year running-mean SSTA. The plot reveals, that during El Niño-like periods (positive amplitude) more

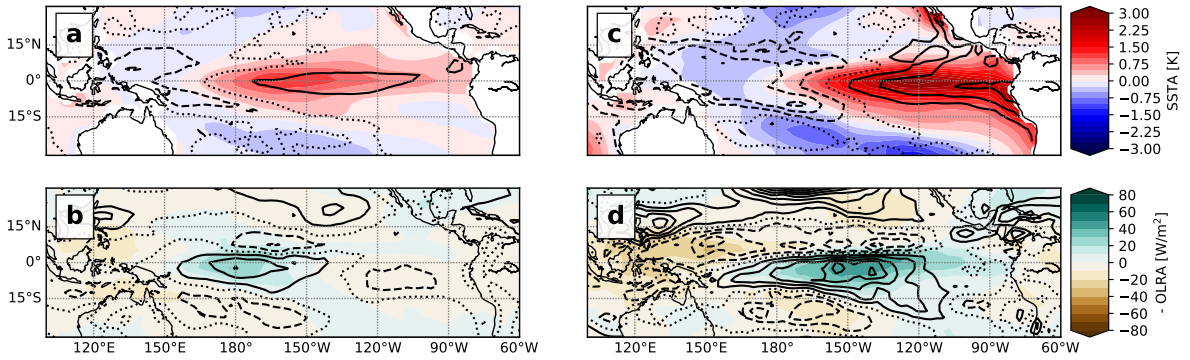


Figure 2.1: *The left (right) side shows composites for CP (EP) El Niños averaged over December-February when the CP (EP) El Niños was present according to Wiedermann et al. (2016).* ***a*** *and* ***c*** *show the SSTA (coloured) and the SSHA (contours). The 0 m SSHA level is dotted, positive levels are solid and negative levels are dashed. The SSHA levels are drawn with a spacing of 5 cm.* ***b*** *and* ***d*** *show the OLRAs (coloured) and the zonal wind stress anomalies (contours). The 0 $m^2\ s^{-2}$ zonal wind stress anomaly level is dotted, positive levels are solid and negative levels are dashed. The zonal wind stress anomaly levels are drawn with a spacing of 5 $m^2\ s^{-2}$.*

convection (decreased OLRAs) is present over the CP, whereas less convection (increased OLRAs) is present over the WP. This fits with the hypothesis that convection is shifted towards the East during El Niño-like periods and towards the West during La Niña-like periods (see Section 1.9). Moreover, Fig. 2.3b shows that the decadal convective signal in the CP during El Niño-like periods is accompanied by strengthened decadal the westerly wind anomalies. This is in accordance with the hypothesis of Clarke (1994) that positive convective anomalies in the CP mostly trigger a westerly wind response.
In the following chapters, two machine learning (ML) models are applied for the ENSO prediction. The insights from the literature review (Chapter 1) and the data analysis are applied to select reasonable predictor variables, interpret the results and draw conclusions from these findings.

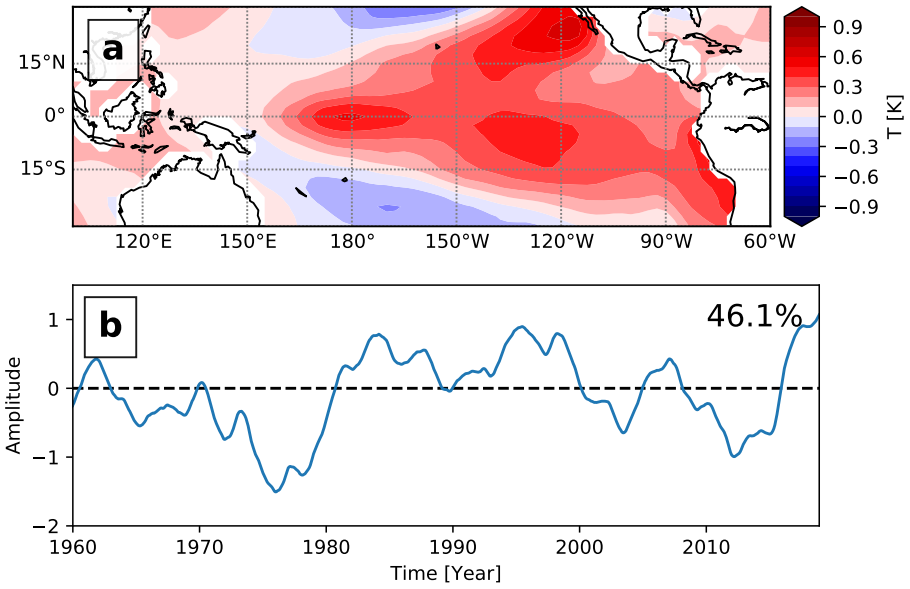


Figure 2.2: *EOF1 of the 5-year running-mean SSTAs* ***(a)*** *and the corresponding amplitude* ***(b)****.*

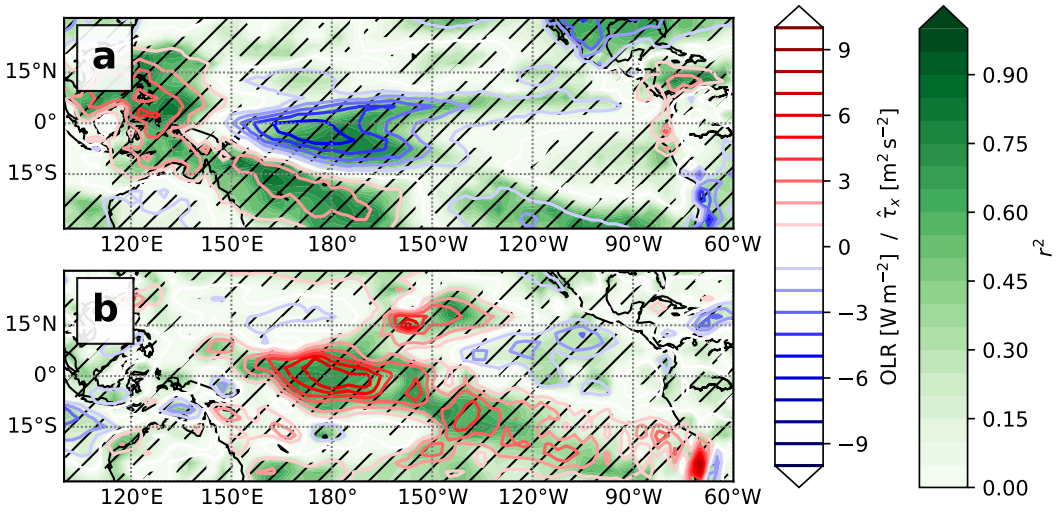


Figure 2.3: *Regression of the 5-year running-mean OLRAs* ***(a)*** *and the 5-year running-mean zonal wind stress anomalies* ***(b)*** *on the amplitude of the EOF1 of the 5-year running-mean SSTA. Coloured contours are the regressed anomalies for a positive one-standard deviation of the amplitude. A positive amplitude indicates El Niño-like periods. The filled contours indicate the explained variance (r^2) and the hatches indicate the areas in which the correlation coefficient is at least 99.9% significant.*

Chapter 3

Deep Ensembles for ENSO-forecasting

3.1 Introduction

3.1.1 Machine learning for ENSO-forecasting

Due to the strong autocorrelation of the ENSO, its prediction has a by far greater predictive horizon than weather prediction. In general, one can divide forecast models for the ENSO into dynamical and statistical models. On the one hand, dynamical models predict the future state of the ENSO by integrating the physical equations that determine the system in time. On the other hand, statistical models are fitted to past observations. The fitted model then can be used to make a forecast over the future state of the ENSO. For the ENSO, most statistical models aim to predict the ONI. Whereas in early years of ENSO forecasting, statistical and dynamical models had an equal prediction skill (Barnston et al., 1994), in recent years, forecasts by dynamical models such as the Climate Forecast System (CFSv2) model (Saha et al., 2013) started to outperform statistical models (Barnston et al., 2012). In general, forecasts that need to pass the spring season often decrease dramatically in prediction skill (Duan and Wei, 2013). Hence, whereas a prediction for December that was made in June can be fairly accurate, a prediction for June made in December is usually not very trustworthy. This is because disturbances have a relatively short lifetime during spring which can be seen in Fig. 3.1.
In the ENSO forecast literature, various definitions of lead time have been applied that can make it confusing to compare results. Here, the definition of Barnston et al. (2012) is used if not stated differently. This means lead time is the time that passed between the date of the last observation and beginning of the target period[1].

3.1.2 Evolving climate networks

Complex network (CN) theory became a popular method to analyze various problems, for instance in social science, biology, finance or environmental science (see e.g. Newman, 2003, for an extensive review). In climate science, CNs were build in earlier studies for a full time series (e.g. Tsonis and Swanson, 2008; Donges et al., 2009). More recently, Radebach et al. (2013) introduced evolving complex networks (ECNs) to analyze the development of a network over time. Instead of computing the network for an entire time series, successive networks are computed for a shifting time window. Interestingly, the two El Niño flavours discussed above are distinguished in Wiedermann et al. (2016) using the transitivity metric from an ECN analysis. Moreover, time series of certain network metrics were found to contain early warning signals for upcoming El Niño/La Niña events more than a year ahead (e.g. Ludescher et al., 2014; Rodríguez-Méndez et al., 2016).
Pecolation-based early warnings in climate networks for an upcoming El Niño/La Niña event are introduced by Rodríguez-Méndez et al. (2016). Specifically, the fractions of small isolated clusters peak before the ENSO transitions into an/a El Niño/La Niña phase. If one views ENSO as a dynamical system with a Hopf bifurcation (Jin, 1997) at which oscillations start to grow, one can give an intuitive explanation for the percolation transition. Sufficiently close to the bifurcation point, the response to

[1] E.g. when the last observed period is December-February and the target season is March-May, the lead time is 0 months.

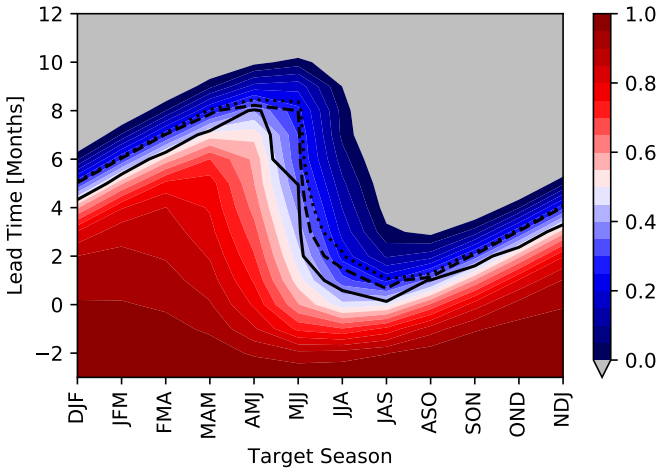


Figure 3.1: *Autocorrelation of the ONI as a function of target season and lead time for the period between 1962 and 2017. The contour lines indicate the 90% (dotted), 95%(dashed) and 99%(solid) significance levels for a Pearson correlation coefficient (at least as strong as the observed one) that is not produced by an actually uncorrelated system.*

perturbations of the system slows down (real part of the Eigenvalues of the Jacobian approach 0). This is the so-called *critical slowdown*. The slowdown in turn leads to stronger autocorrelations at longer lead times within the system. For spatially coupled systems, these local temporal correlations translate into spatial correlations. It follows that isolated nodes can suddenly become connected to another node, building a cluster of size two (in case the ECN is build using a correlation coefficient). The fraction of nodes that are part of a cluster of size s can be measured with the following variable:

$$c_s = \frac{sn_s}{N} \tag{3.1}$$

Here, n_s is the number of clusters of size s and N the total amount of nodes N. For a proceeding transition, more small clusters of size two emerge. Hence, c_2 increases. For a further transition, small clusters can form even bigger clusters, counter-balancing the increase for fractions of smaller clusters. Therefore, a first sign of a percolation transition is indicated by a peak of c_2 which is followed by peaks of c_s with increasing s when the bifurcation point is approached. At one point, spatial correlations will be strong enough such that a so-called *giant component* emerges which incorporates a large fraction of all nodes. This point is called the percolation point. Hence, in a stereo typical case the giant component is present around the bifurcation and hence the bifurcation point is surrounded by two percolation points. Rodríguez-Méndez et al. (2016) show that this percolation transition occurs before the actual dynamical transition takes place. If the system moves again away from the bifurcation point, peaks in c_s occur in reversed order.

3.1.3 Artificial neural networks for ENSO forecasting

Artificial neural networks (ANNs) are powerful statistical models. They are inspired by biological neural networks which are one of the core building blocks of human brains. ANNs can be used for regression and classification tasks in a supervised learning setup. A so-called *feed-forward neural network* (FFNN) is a specific type of ANN in which neurons are structured in layers. In these FFNNs, information is passed from an input layer through an arbitrary number of hidden layers to an output layer. If the FFNN has at least one hidden layer, it is usually called a Multilayer Perceptron (MLP). MLPs with at least one hidden layer have the remarkable ability to approximate any continuous nonlinear function (regression task) given that there model architecture is complex enough to cope with the characteristics of the nonlinear function and that there is enough data to train the MLP. This is the so-called *universal approximation theorem* (e.g. Cybenko, 1989; Hornik, 1991; Csáji, 2001).
Extensive research was undertaken by a working group from the University of British Columbia (UBC) using MLPs for the prediction of the ENSO. In their first study by Tangang et al. (1997), the amplitudes of the first seven leading wind stress EOFs as well as the ONI are used as input to an MLP to predict the ONI at various lead times. For a test period between 1982 and 1992, the ANN archived remarkable

prediction skills, i.e. for the 12-months lead time, the Pearson correlation coefficient for the correlation between the predicted and the observed ONI was $r \approx 0.65$. However, as already discussed above, this period was an El Niño-like period and therefore relatively easy to predict. Tangang et al. (1997) already point out that the prediction skill was considerably lower when the test data set was chosen to be in the 50s, 60s or 70s.

An extension of the model from Tangang et al. (1997) was published a year later in Tangang et al. (1998a). In this study, the best MLP used the amplitudes of the first seven leading EOFs of the sea level pressure (SLP) field as well as the considered ENSO index itself as predictor variables. For the ONI, the MLP, again, archived an impressive prediction skill for the test period between 1982-1993 ($r \approx 0.65$, 12-months lead). For the decades 1962-1971 and 1972-81, the results were significantly worse ($r \approx 0.4$, 12-months lead time). Note, that these periods of lower predictability coincide with the more La Niña-like background state described above.

A very similar MLP model (using just the four leading EOFs of the SLP) from the UBC group dramatically underperformed for the period 2002-2011 in the study of Barnston et al. (2012) ($r \approx 0.0$, 7-months lead time). Since the 2002-2011 period had a La Niña-like background state, low predictability was expected. However, the skill was even considerably lower than the one reported for other La Niña-like decades (1962-1971 and 1972-81). Since the authors do not indicate that a third data set was used for the hyperparameter optimization, one can suspect that hyperparameters were tuned onto the test data. Such a practice invalidates the test data set as a truly independent data set because it "leaks" knowledge from the test data set into the model. Hence, reported skills are expected to be better than for a truly independent test data set. See Section 3.2.3 for more details on the importance of the split of the full data set into a train, validation and test data set.

One of the latest papers by the UBC group was published by Wu et al. (2006). Here, they applied the method of extended EOFs (EEOFs) on the SLP anomaly (SLPA) and SSTA field to generate a feature set. In contrast to previous studies, they chose the amplitudes of the five leading EOFs of the SSTA field as target variables. This enabled them to make forecasts for the full SSTA field by reconstructing the SSTA field from these five amplitudes. Furthermore, an ensemble method was applied to generate an ensemble of MLPs. For this, each time series of the predictor and target variables were divided into 10 segments. Each segment was one time the test data set[2] for which an ensemble of 30 MLPs was generated based on the remaining nine segments. 85% (D85) of the remaining time series were used to train the model whereas 15% (D15) were used to check for overfitting[3]. Unfortunately, the D85 and D15 data sets were formed by randomly choosing data points from the considered nine segments. However, random sampling must not be applied when dealing with the ENSO to split a time series. This is because of the strong autocorrelations in the system (for more details see Section 3.2.3). Despite this methodical mistake, their ensemble models still had a very good overall prediction skill on the test data sets (1948-2005). The overall correlation skill i.e. for the 15-months lead time[4] was $r \approx 0.5$.

A recent study made by Nooteboom et al. (2018) reported ENSO predictions that might overcome the spring predictability barrier. They use a combination of an autoregressive integrated moving average (ARIMA) model and an MLP. The ARIMA model is trained to predict the ONI based on the 12 proceeding months of the ONI. Afterwards, the ANN model is trained to predict the residual. For the 4 and the 6-months lead time, the inputs to the ANN are the amplitude, called principal component (PC), of the EOF2 of the wind stress field (PC_2), the WWV and the seasonal cycle (SC) represented by a sinusoid with a period of one year (all from the last observed month). Instead of PC_2, the c_2 variable from the SSHA field was used for the 12-months lead time. The model was evaluated for a test period between 2007 and 2013. Interestingly, the model showed a signal for the 2009/2010 El Niño for the 12-months lead time that was poorly predicted by other models[5]. However, this signal was not present in the 6-months lead time forecast, which makes the model seem to be inconsistent. However, this is probably because the c_2 variable was not included as input for the shorter lead times.

In Fig. 8 of Nooteboom et al. (2018), an ensemble forecast is shown to prove that the good performance of the model was not a "lucky shot". The ensemble is generated by training multiple MLPs with different architectures using the same training data set. Note, the reported spread of the ensemble prediction does not cover the observation of the ONI. This indicates that such an ensemble method cannot be used to estimate the predictive uncertainty within the forecast. In the following section, some methods will be

[2] Wu et al. (2006) call the test data set "validation data set" which is inconsistent with the nomenclature in data science nowadays.

[3] This D15 data set is the data set one would call validation data set nowadays.

[4] 12-months lead time when the Barnston et al. (2012) definition of lead time would have been applied.

[5] `https://iri.columbia.edu/our-expertise/climate/forecasts/enso/archive/200903/SST_table.html`

introduced that are better suited to estimate the predictive uncertainty of ANNs.

3.1.4 Predictive uncertainty of neural networks

In recent years, advanced methods became available to estimate the uncertainties of ANN predictions. In *Bayesian neural networks* (BNNs) weights and biases are not fixed values but distributions. The weight distributions can be learned using Bayesian inference. Based on prior distributions (distribution before data was seen) for the weights the posterior distributions (after data was seen by the model) can be computed. Performing Bayesian inference for an ANN is not a straight-forward procedure. Sophisticated methods have been developed to infer the posteriors of the weights. For instance, a Markov-Chain Monte-Carlo method (MCMC) is a clever algorithm that walks through the parameter space in such a way that it eventually can sample the posterior distributions (e.g. by Gibbs sampling as described in Geman and Geman, 1987). Further, Blundell et al. (2015) introduced a method called *variational inference* where an approximated posterior is computed based on the assumption that the posterior is a Gaussian. In one of the examples in McDermott and Wikle (2019), the application of a BNN is shown for the ENSO prediction. Unfortunately, results are just presented for the years 2015-2016. Hence, a comprehensive analysis of the application of BNNs for ENSO forecasting is still missing in the literature.
Recently, Lakshminarayanan et al. (2017) introduced a conceptually simpler method, so-called *Deep Ensembles* (DEs), to estimate the predictive uncertainty of an ANN model. DEs consist out of multiple MLPs where each MLP has two output neurons that predict the mean and the variance (or standard deviation) of the target variable. They are trained using the negative log-likelihood of a Gaussian distribution (more details in Section 3.2.3). In this Master's thesis, DEs are for the first time applied for the ENSO prediction.

3.2 Data and methods

3.2.1 Predictor variables

The data sources, as well as some general preprocessing, were already described in Section 2.1. This section focuses on motivating the use of the specific predictor variables and describes some additional preprocessing which is of particular importance for this chapter, namely the computation of ECN metrics. Predictor variables, also-called features, are the variables which are used in a statistical model to predict the so-called target variable, which is sometimes called label. The first predictor variable which is used is the ONI itself. This is done to introduce information that comes from the strong autocorrelation of the ONI (see Fig. 3.1). Moreover, the WWV is used as a predictor variable because it is known to be a good predictor of an upcoming El Niño/La Niña when the thermocline feedback is the major feedback in the development of the anomaly (see Section 1.5). Another index which is used as a predictor variable is the DMI. This is because the IOD was found to be a modulator as well as a predictor of the ENSO (see Section 1.8). Furthermore, a cosine function with a period of one year is included to represent the seasonal cycle (SC). This variable is included to supply the model with information about the specific time of the year. This is useful for the model to better represent the phase-locking of the ENSO as well as the seasonal cycle of the standard deviation of the ONI.
Next to the described predictor variables, the zonally averaged (2.5°S to 2.5°N) zonal wind stress, $\bar{\tau}_x$, in the western Pacific (WP, 120°E to 160°W) is used. Westerly wind anomalies and westerly wind bursts (WWB) are an important factor in the development of an El Niño because they can excite equatorial Kelvin and Rossby waves in the ocean (see Section 1.3) and trigger the convective ocean-atmosphere instability (see Sec 1.4).
As seen in various parts of Chapter 1 and Section 2.3, the decadal variability of the Pacific ocean seems to have a strong influence on the dynamics of the ENSO. Because of this, the amplitude of the EOF1 from the 5-year running-mean SSTA field (as in Section 2.3) is used as a predictor variable. It will be abbreviated as PC_1-$\mathrm{SSTA}_{5\,\mathrm{year}}$ from now on.
In the following subsection, the last two predictor variables are described which are derived from the ECN time series of the SSHA field.

Evolving climate networks

The ECN time series is computed for the monthly SSHA field between 30°S to 30°N and 120°E to 160°W. To do so, the SSHA data is first re-gridded to the 2.5°×2.5° grid of the NCEP/NCAR data set.

Table 3.1: *Predictor variables which are used for the DE. For each variable a keyword for the physical reason for its usage as predictor variable is given. Furthermore, the data set from which the variable is retrieved and/or computed is indicated as well as the preprocessing that has been applied.*

Variable	Physical reason	Data set	Preprocessing
ONI	Autocorrelation (Fig. 3.1)	ERSSTv5	3-months running-mean SST over NINO3.4 area; removed long-time seasonal mean
SC	Seasonal phase-locking	Cosine function with yearly period	
WWV	Recharge oscillator (Section 1.5)	Computed by NOAA-PMEL and Bunge and Clarke (2014)(Section 2.1)	WWV proxy (1960-1979); WWV measured (1980-2017); removed long-time seasonal mean
IOD	Modulator and precursor for ENSO (Section 1.8)	HadISST	
$\tau_{x,\mathrm{WP}}$	Stochastic effects of the weather noise (Section 1.4)	NCEP/NCAR reanalysis	Computed by Eq. (2.1); removed long-time seasonal mean, average over WP
PC_1 $\mathrm{SSTA}_{5\,\mathrm{year}}$	Decadel variability	ERSSTv5	Amplitude of the leading EOF of the 5-year running-mean SSTA field
c_2	Percolation theory (Section 3.1.2)	ORAS4	Evolving climate networks; Rodríguez-Méndez et al. (2016)
$\mathcal{H}^*_{t,t-1}$	Percolation theory (Section 3.1.2)	ORAS4	Evolving climate networks; Radebach et al. (2013)

Then, multiple complex climate networks are computed for a shifting time window with a window size of 12 month. For each time step of the ECN time series, the window is shifted by a month. To build one complex climate network, the cross-correlation matrix, R, for the SSHA field is computed (using Pearson correlation). Then, two grid points are assumed to be linked if the correlation coefficient is greater than the threshold $\epsilon = 0.9$. In mathematical terms, the adjacency matrix of the undirected complex climate network can be computed from R as follows:

$$A_{ij} = \Theta(|R_{ij}| - \epsilon) - \delta_{ij} \tag{3.2}$$

Here, $\Theta(\cdot)$ is the Heaviside function and δ the Kronecker delta function.
In this study, the corrected Hamming distance $\mathcal{H}^*_{t,t-1}$, computed as in Radebach et al. (2013), and c_2 are computed for each complex climate network. The computed network metrics are assigned to the last month of the time window, e.g. if the time window was March 2011 to February 2012 the metrics are assigned to February 2012. This is done to ensure not to include any information from the future into the metrics.
In Section 3.1.2 it was already motivated why c_2 is a valuable predictor variable. The ECN metric $\mathcal{H}^*_{t,t-1}$ is included for similar reasons. It is basically a metric to estimate how much the complex climate network changed from one time step to the following (as indicated by the subscript $t, t-1$). Hence, it is assumed here that it contains, as c_2, signals regarding the percolation transition of the ECN when the bifurcation point that gives rise to the El Niño and La Niña events is approached.
The complete set of all predictor variables is briefly summarized in Table 3.1. For the final prediction scheme of the DE, all values from the past year previous to the latest observed value are used as predictor values for each of the nine predictor variables described above. Hence, the predictor variable set consists out of $9 \times 12 = 108$ values.

3.2.2 Multilayer Perceptrons

MLPs were already briefly introduced in Section 3.1.3. In this section, some more methodical details are provided. MLPs are layered fully-connected feed-forward neural networks. They consist of multiple layers of neurons where neurons are connected to all neurons in the proceeding and the subsequent layer. Through these connections, neurons can receive and pass information. In an MLP, a neuron gets multiple inputs from the neurons from the proceeding layer but just passes forward one output value to all of the neurons in the subsequent layer. For the output of the neuron, the inputs are first weighted, then summed up including a bias term and afterwards passed through a so-called *activation function*. This function is usually an nonlinear function such as the hyperbolic tangent or a Rectified Linear Unit (ReLU, $f(x) = \max(0, x)$). In mathematical terms the operation done by a neuron reads as follows:

$$y = f\left(\sum_{i=1}^{N} w_i x_i + b\right) \tag{3.3}$$

Here, x_i are the N input values to the neuron, w_i are the corresponding weights, b is the bias, f the activation function and y the output. MLPs can be trained using a gradient descent algorithm to minimize a so-called *loss function*. Depending on the kind of task (regression or classification), different loss functions can be used. For the regression task for the prediction of the ONI, usually, the root-mean-squared error (RMSE) is used. However, in this study, the negative log-likelihood (NLL) of a Gaussian distribution will be used because the MLPs of the DE predict two values, namely the mean and the standard deviation of a Gaussian distribution. The following section describes this in greater detail.

3.2.3 Deep ensemble

DEs were briefly introduced in Section 3.1.4. An example architecture for an DE member is shown in Fig. 3.2. DEs are used to forecast the ONI with a certain lead time. Before the training of the DE, each predictor variable is normalized by subtracting its mean and dividing by its standard deviation of the entire time series (1960-2017). Then, they are merged to from the feature vector $\mathbf{x}$.
The members of a DE are MLPs which have one hidden layer with 16 neurons. The activation function of the hidden neurons is chosen to be the ReLU. As described earlier, a member of the DE has two output neurons for the mean, $\hat{\mu}_{\theta_i}(\mathbf{x})$ and the standard deviation $\hat{\sigma}_{\theta_i}(\mathbf{x})$ of a Gaussian distribution, where θ_i refers to the parameter setting, i.e. the weights, of one member. The output neuron for $\hat{\mu}_{\theta_i}(\mathbf{x})$ has a linear activation function and the $\hat{\sigma}_{\theta_i}(\mathbf{x})$ the softplus function. The ensemble prediction of the mean, $\hat{\mu}_*(\mathbf{x})$, is the average of all $\hat{\mu}_{\theta_i}(\mathbf{x})$:

$$\hat{\mu}_*(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\hat{\mu}_{\theta_i}(\mathbf{x}) \tag{3.4}$$

And the ensemble prediction of the standard deviation, $\hat{\sigma}_*(\mathbf{x})$, is

$$\hat{\sigma}_*^2(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\sigma}_{\theta_i}^2(\mathbf{x}) - \hat{\mu}_{\theta_i}^2(\mathbf{x})\right) - \hat{\mu}_*^2(\mathbf{x}). \tag{3.5}$$

A DE member is trained using the NLL of a Gaussian distribution:

$$-\log P(y|\hat{\mu},\hat{\sigma}^2) = \frac{1}{2}\log\hat{\sigma}^2 + \frac{(x-\hat{\mu})^2}{2\hat{\sigma}^2} + constant \tag{3.6}$$

Here, y is the observation of the predicted variable which is in this study the ONI at some specific lead time. For numerical stability, a very small term of the order $\mathcal{O}(10^{-6})$ was added to $\hat{\sigma}$ during training.
The likelihood for given time series $\mathbf{Y} = \{y_1, y_2, ..., y_n\}$, $\mathbf{M} = \{\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_n\}$ and $\mathbf{\Sigma} = \{\sigma_1, \sigma_2, ..., \sigma_n\}$ is now

$$P(\mathbf{Y}|\mathbf{M},\mathbf{\Sigma}^2) = \prod_{i=1}^{N} P(y_i|\hat{\mu}_i,\hat{\sigma}_i^2). \tag{3.7}$$

Hence, the corresponding NLL is given by

$$-\log P(\mathbf{Y}|\mathbf{M},\mathbf{\Sigma}^2) = -\sum_{i=1}^{N}\log P(y_i|\hat{\mu}_i,\hat{\sigma}_i^2). \tag{3.8}$$

The quantity in Eq. (3.8) is minimized to fit a member to the given data set. For this, the weights and biases of the considered MLP are trained using the Adam optimizer (Kingma and Ba, 2014).
To regularize the network, various techniques are applied. First, Gaussian noise layers are installed after the input and just before the output layer. Then, the hidden layer was equipped with a dropout rate (Srivastava et al., 2014). Finally, the L_1 and the L_2 penalty terms were used in combination (referred to as *elastic net*) for the hidden and the output layer. Applying both penalty terms is a good regularization technique for neural networks when the number of observations is low (Zou and Hastie, 2005). Note, that the L_1-regularization is able to decrease weights to 0 and therefore removes useless features from the feature vector. Lastly, *Early Stopping* is applied when the loss on a validation data set (on which the model is not trained) of the trained MLP starts to become worse again.
Hence, before training, the data needs to be split into a training, validation and test data set. While the training data set is used to perform the training of the neural network using backpropagation, the

validation data set is used to tune hyperparameters such that over- and underfitting is avoided. The model that performs best on the validation data set is then evaluated on the test data set. It its very important to understand, that the test data set is reserved solely to evaluate the prediction of the trained DE but not to find the best hyperparameter setting. Studies in the past that applied ANNs to the ENSO forecast often just split the data into a train and validation data set and reported the skill of the model on the validation data set (Tangang et al., 1998b; Baawain et al., 2005; Nooteboom et al., 2018). Although the model is not trained on the validation data set, it is possible that it can learn from it. For instance, when Early Stopping is applied, the model training is stopped as soon one cannot see any improvement on the validation data set. The resulting weights are therefore somewhat nudged towards a good performance on the validation data set. It is because of this possible nudging, why a test data set is needed to better estimate the true performance of the model.
In addition, for the ENSO forecast problem, it is very important to split the dataset into connected time series instead of using randomly sampled points. This is because of the high autocorrelations in the ENSO system. If one would split the data set by randomly selecting points, a point from the test data set would likely be close to a data point from the training data set and because of the high autocorrelation very similar to that point such that the model indirectly has information over the test data set. Hence, the test data set is not independent. If such a mistake is made, prediction skills on the test data set are nearly as good as on the training data set. This occurred in the study of Wu et al. (2006). They first correctly performed a split into a test and a train/validation data set. However, in the following, the train/validation data is split into a training data set (D85) and a validation dataset (D15) by random sampling. Also Baawain et al. (2005) likely used random sampling to split their data set into a test and train data set. This is indicated by an unusual high prediction skill on the test data accompanied by a very similar prediction skill on the train data set.
In this study, to test the DE on the entire available time series, the data set is split into six test periods, namely 1962-1971, 1972-1981, 1982-1991, 1992-2001, 2002-2011 and 2011-2017. For each test period, the remaining data is used to train the model using a scheme that is similar to cross validation. For this, the remaining data set is divided into five segements. One ensemble member is trained on four segments and validated on the other segment using the NLL. This is done until each segment was once the validation data set. Fig. 3.3 illustrates this splitting process. Hence, the entire ensemble consists out of 5 members and in total 6 ensembles are needed per lead time for which the model is trained.
This process was repeated 200 times for each test decade with a random uniform choice of the dropout rate of the hidden layer $(0.1, 0.5)$, the Gaussian noise to the input and output $(0.1, 0.5)$ as well as the L_1 and L_2 penalty terms of the hidden and the output layers $(0.0, 0.2)$ with the boundaries of the uniform distribution in parenthesis. The DE that had the lowest average NLL on the validation data sets is then evaluated on the respective test period. The described hyperparameter optimization using random hyperparameters is usually referred to as *random search*.
In this study, DEs for the 0, 3, 6, 9, 12 and 15-months lead time are trained. All in all, 5(members per DE)$\times$6(considered lead times)$\times$6(test periods)$\times$200(random search iterations)$= 36,000$ single MLPs are trained out of which 180 MLPs eventually formed the 36 DEs that are necessary for the evaluation on six test periods regarding the six considered lead times.

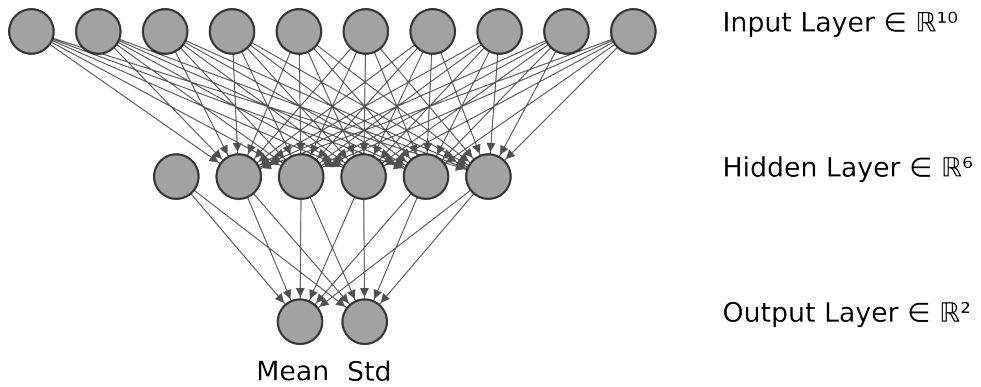


Figure 3.2: *An example for an architecture of a DE member with ten input neurons, six hidden layer neurons and two output neurons for the mean and the standard deviation. Note, that in the input layer and the hidden layer one neuron accounts for the bias term in the subsequent layer by having a constant output equal to 1. For the hidden layer this neuron is the one that is not connected to the input layer neurons.*

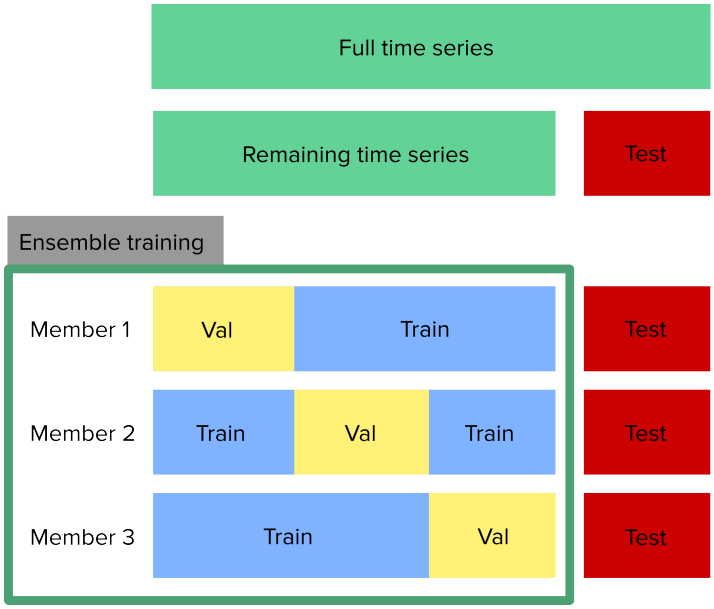


Figure 3.3: *An illustration of the data splitting and ensemble training which is applied in this study. From a full time series, first the test data (red) set is split apart. The remaining series is then split into (here) three segements which are used to train an ensemble model. The ensemble members are trained on the training data set (blue) and hyperparameters are optimized for the validation data sets (yellow). The model ensemble that performs best in terms of the average loss on the validation data set is evaluated on the test data set.*

3.2.4 Model evaluation

The DE is evaluated using different skill measures. The first measure is the Pearson correlation coefficient, r, between the observed ONI, y, and the predicted mean, $\hat{\mu}$:

$$r = \frac{\overline{y'\hat{\mu}'}}{\sigma_y \sigma_{\hat{\mu}}} \tag{3.9}$$

The dash indicates the deviation from the mean, the overline the average and σ the standard deviation of the corresponding variable written in the subscript.
Furthermore, the standardized root-mean-square error (SRMSE) and the corresponding seasonally SRMSE (SSRMSE) are used:

$$\mathrm{SRMSE}(y, \hat{\mu}) = \frac{\mathrm{RMSE}(y, \hat{\mu})}{\sigma_y} \tag{3.10}$$

$$\mathrm{SSRMSE}(y, \hat{\mu}) = \frac{1}{12} \sum_{i=1}^{12} \mathrm{SRMSE}(y_i, \hat{\mu}_i) \tag{3.11}$$

where the subscript i selects all data points that belong to the same season, e.g. DJF. This variable is used because the standard deviation of the ONI has a seasonal cycle which makes the seasonal skill of the simple RMSE not comparable to other seasons.
Finally, the NLL from Eq. (3.6) is used to evaluate the model.

3.3 Results

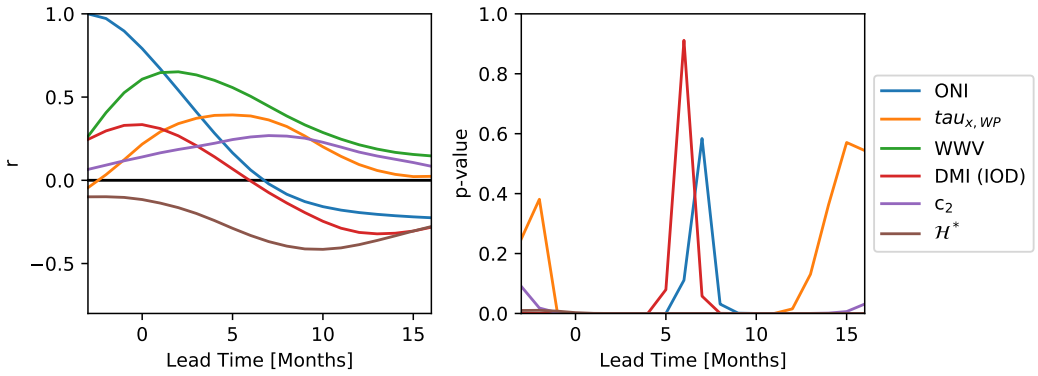
At first, the predictor variables are analyzed. Fig. 3.4a shows the correlation of the predictor variables with the ONI for various lead times for the period between 1962-2017. The WWV correlates strongest with the ONI at a lead time of about 2-3 months. For $\tau_{x,\mathrm{WP}}$, strongest correlation appears for a lead time of 3-5 months. The DMI has two peaks of strong correlation. One is observed for a lead time of about 0 months and the other for a lead time of about 13 months. The c_2 variable shows, in general, a fairly weak but still significant correlation with the ONI with the strongest correlation for a lead time of 8 months. Interestingly, $\mathcal{H}^*_{t,t-1}$ has a relatively strong correlation with the ONI for a lead time of 10 months ($r \approx -0.4$).

In Fig. 3.4b and 3.4d, the same correlation analysis is done for the time periods 1962-1981 and 2001-2017, respectively. These periods can be roughly identified as La Niña-like periods. One can see that the lead time of maximum correlation of the WWV decreases towards 0-1 months. Furthermore, the overall correlation coefficient of the DMI, c_2 and $\mathcal{H}^*_{t,t-1}$ for long times are weaker than for the entire time series in Fig. 3.4a. Whereas the mentioned predictor variables generally seem to be weaker predictors for the ONI during La Niña-like periods, $\tau_{x,\mathrm{WP}}$ is a better predictor during 2002-2017 and just weaker between 1962-1981. However, the overall picture shows that the majority of the predictors have a relatively weaker predictive skill in the La Niña-like periods.
In contrast, for more the El Niño-like period (1981-2001) in Fig. 3.4c, the lead time of maximum correlation of the WWV increases to 4-5 months. Furthermore, the correlation coefficient of the DMI, c_2 and $\mathcal{H}^*_{t,t-1}$ for long times are relatively strong. Additionally, one can see that the autocorrelation of the ONI is as well stronger for longer lead times during the El Niño-like period. This means disturbances are more persistent during this period. The prediction skill of $\tau_{x,\mathrm{WP}}$ is about the same as for the entire time series.
Fig. 3.5 shows the hindcast of the DEs for various lead times. For the 0-months lead time, the predicted standard deviations are generally small and the predicted mean quite accurately follows the observed ONI. The majority of the El Niño and La Niña events are as well signalled in the 3-months lead time hindcasts. However, the prediction already comes with a notably larger predictive uncertainty. In the 6 and 9-months lead time hindcasts, signals for some El Niño and La Niña events are only present for the El Niño-like period, e.g. El Niño events in 1982/1983, 1986/1987, 1987/1988 and 1997/1998 and subsequent La Niñas. Signals for El Niño and La Niña events are mostly gone in the 12-months hindcasts. For this lead time, the DEs mostly predict the climatology distribution of the ONI.
The skills of the DEs including all seasons are shown in Fig. 3.6 for the full time series and for different decades. As ANN models in older studies (see Section 3.1.3), the model has a good prediction skill for higher lead times in the El Niño-like decades (1982-2001) and performs considerably worse in the La Niña-like periods (1962-1981, 2001-2017). However, also for the La Niña-like periods the skills of the DEs are considerably better than the persistence skills.
The skill of the model for different target season is shown in 3.7a for the entire period between 1962 and 2017. One can see that the summer period is the most difficult period to predict. Interstingly, the spring season does not appear with a predictability barrier in this plot. This means that predictions for the winter period still have a significantly high prediction skill for lead times up to 9-months (predictions that are made in winter for the winter a year ahead). For instance, the DE signalled an upcoming El Niño event in 1986/1987 and 1997/1998 as well as predicted accurately the development between 2000 and 2003 a year ahead (see Fig. 3.5).
However, if one evaluates the model again onto its skill during the El Niño-like periods (see Fig. 3.7b) and La Niña-like periods (see Fig. 3.7c), it becomes clear that the model is just able to overcome the spring predictability barrier to some extent during El Niño-like but not during La Niña-like periods. During La Niña-like periods, the DEs do not even have a significant correlation skill for rather short lead times of 3 months if the target season in boreal summer. This is accompanied by an SRMSE for summer seasons that is greater than 1. This means that the predicted mean during La-Niña periods is on average more than one standard deviation of the observed value. Note, that this is worse than just predicting a mean 0 which would lead to an SRMSE value of 1.

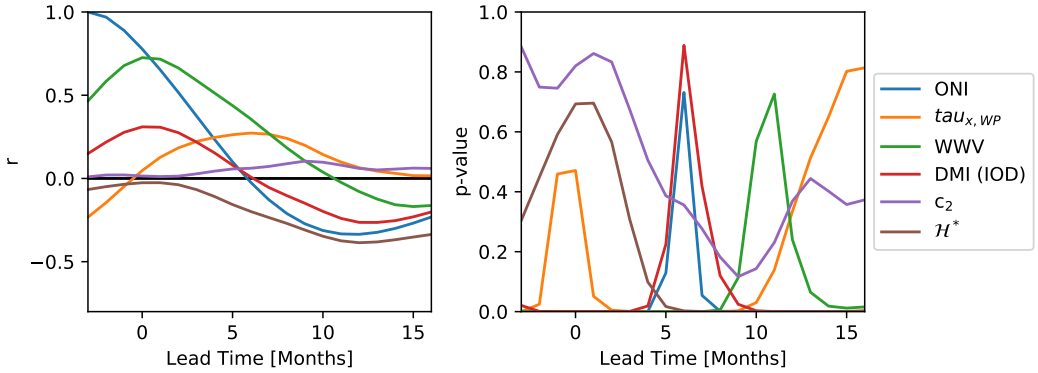
3.4 Discussion

3.4.1 Interdecadal changes in prediction skill

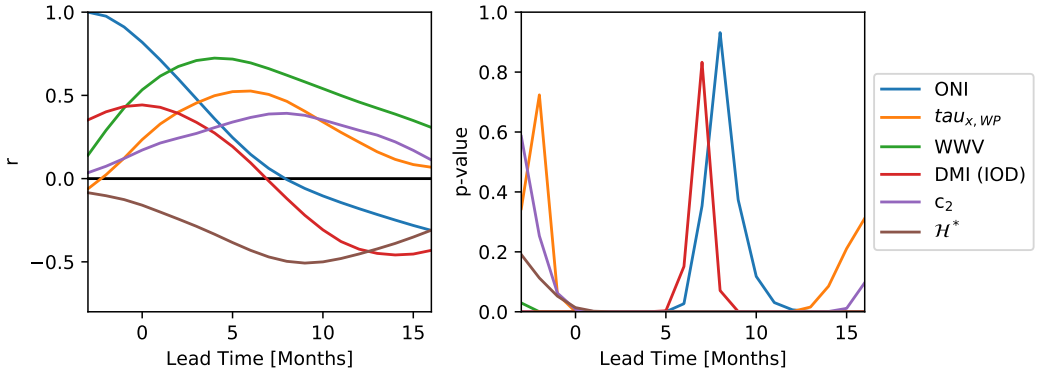
The DEs show notable prediction skills for the El Niño-like period between 1982-2001. In contrast, the prediction skills are worse for the La Niña-like periods (1962-1981, 2002-2017). Similar interdecadal differences in prediction skills were already seen in previous studies for ANN models but as well for other statistical and dynamical models.
Table 3.2 summarizes the interdecadal change of correlation skill for various ANN models. The ANN in Tangang et al. (1997) for instance, has a high correlation skill for the period 1983-1992 (6-months lead, $r = 0.70$) but a considerably weaker skill for the three decades between 1952-1982 (6-months lead, $r < 0.34$). Also Wu et al. (2006) observe relatively low skills in the 1950s (6-months lead, $r < 0.4$), moderate prediction skill in the 1960s, 1970s and 1990s (6-months lead, $r \approx 0.4 - 0.6$), and a strong increase in prediction skill in the 1980s (6-months lead, $r \approx 0.6 - 0.75$) for their ANN model. Similar



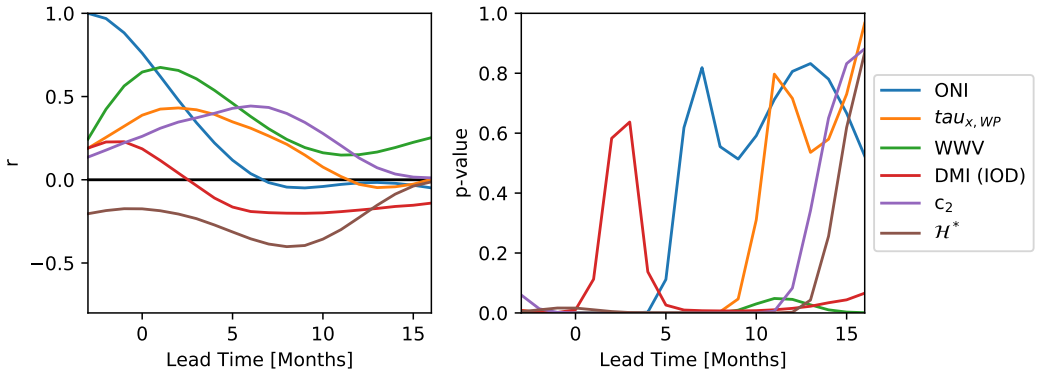
(a) 1962-2017



(b) 1962-1981



(c) 1982-2001



(d) 2002-2017

Figure 3.4: *The Pearson correlation coefficient of each predictor variable with the ONI is plotted against the lead time at the left side. The right side shows the corresponding p-value for an uncorrelated system producing a correlation that is at least as strong as on the one indicated on the left side.*

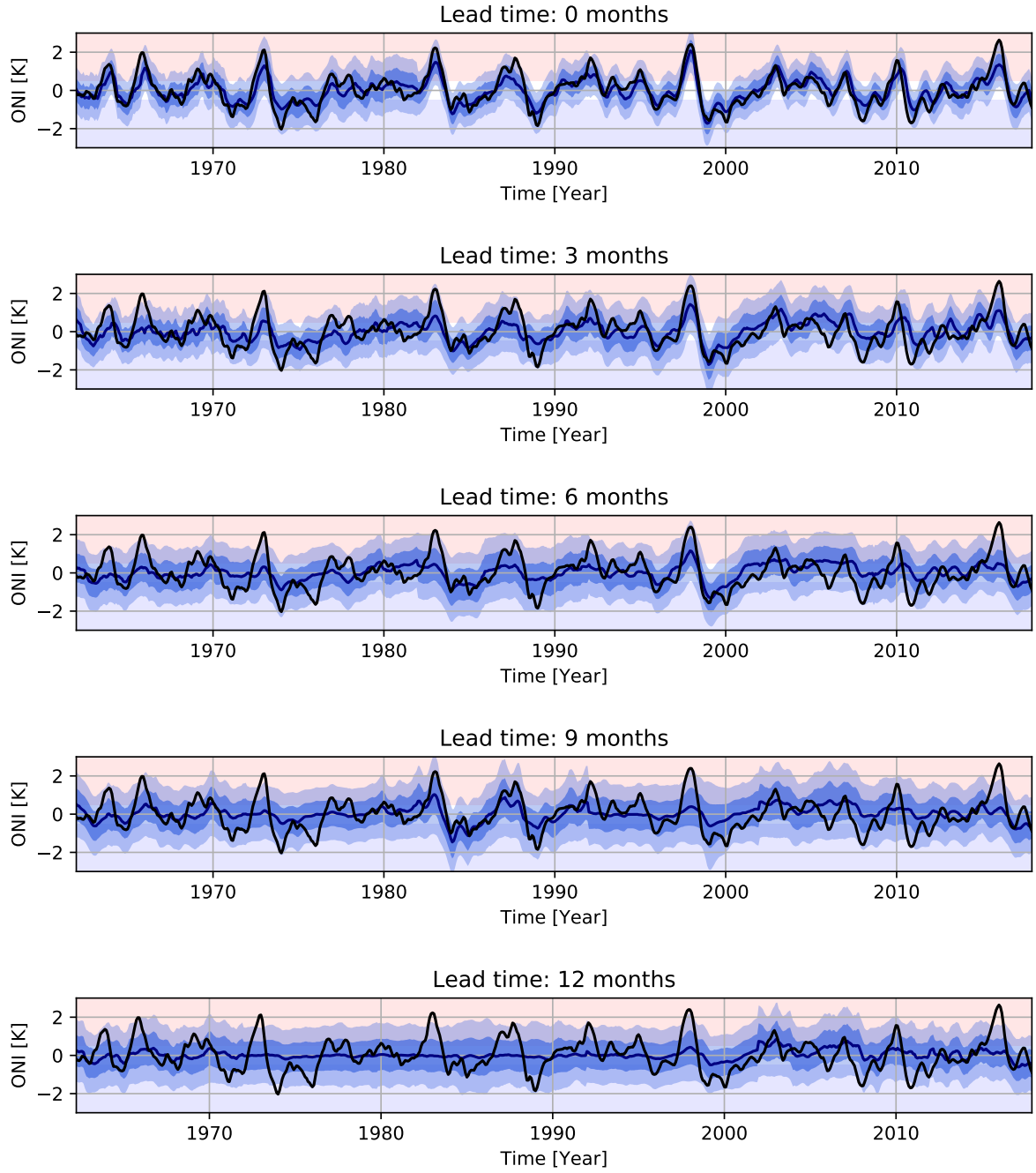
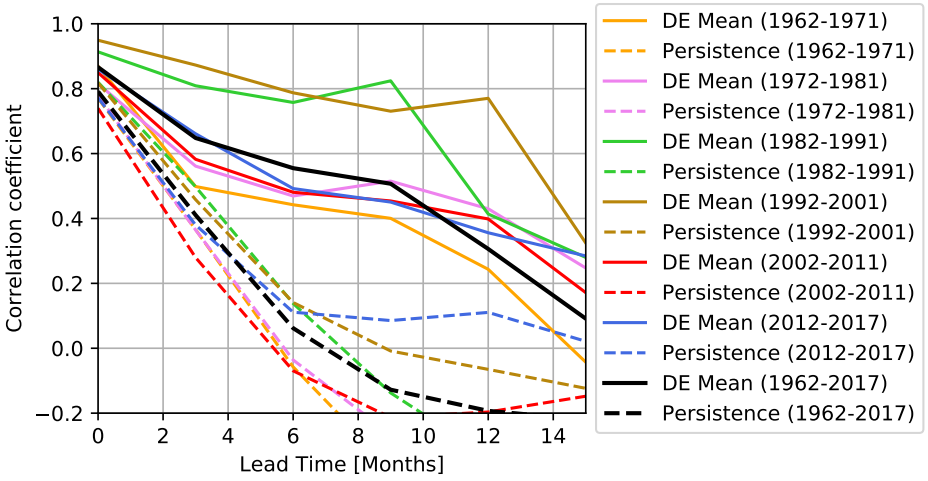
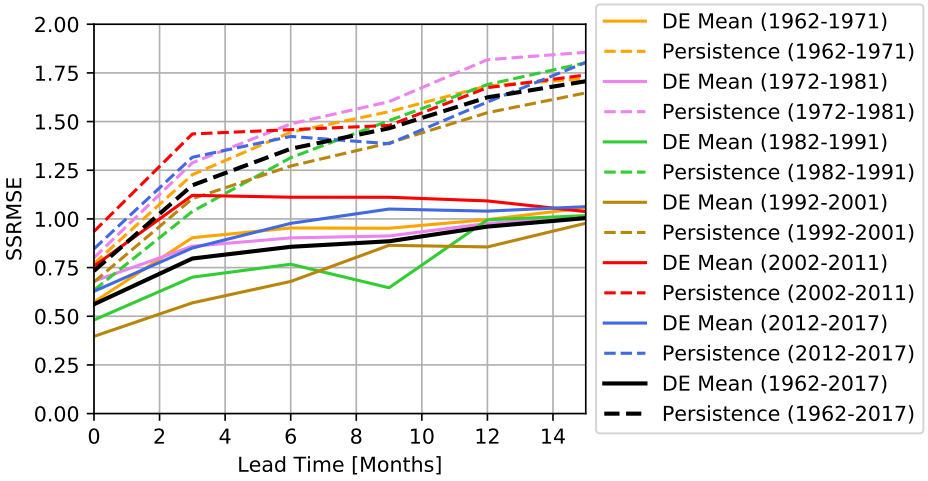


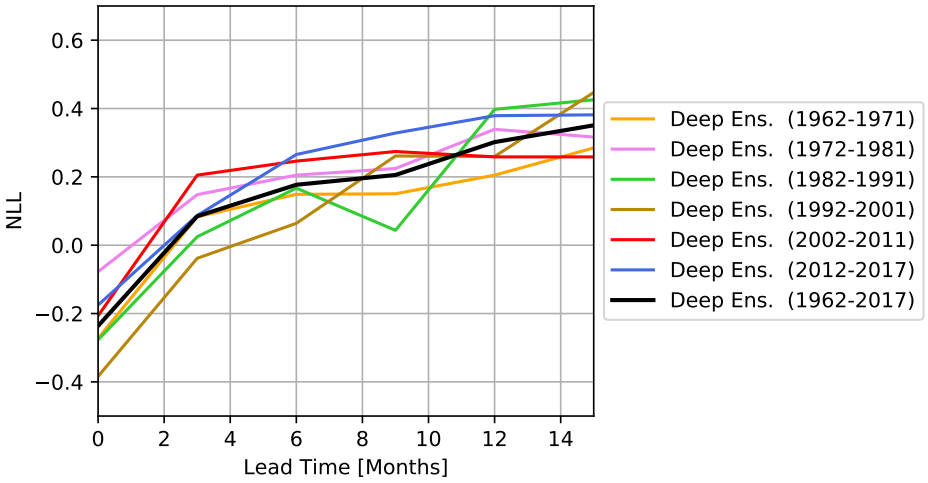
Figure 3.5: *Hindcasts of the DEs for various lead times. The solid blue line depicts the predicted mean. The darker blue shading shows the one-standard-deviation interval and the lighter shading the two-standard deviation interval.*



(a)

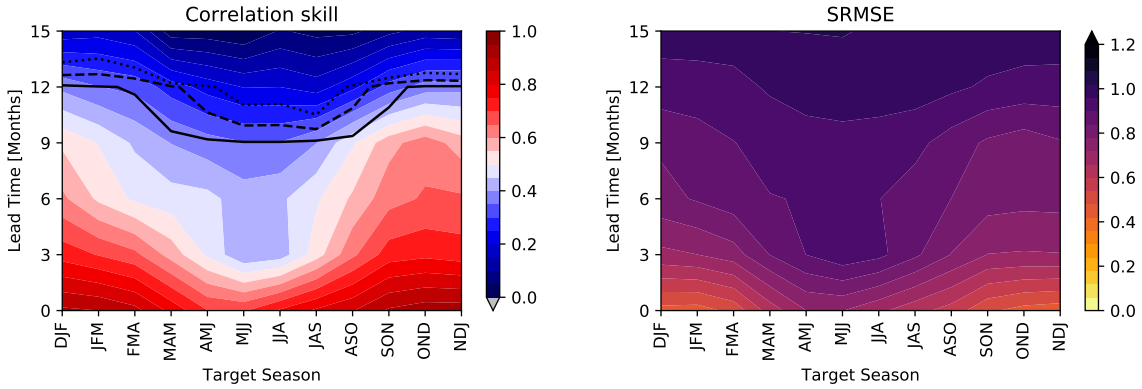


(b)

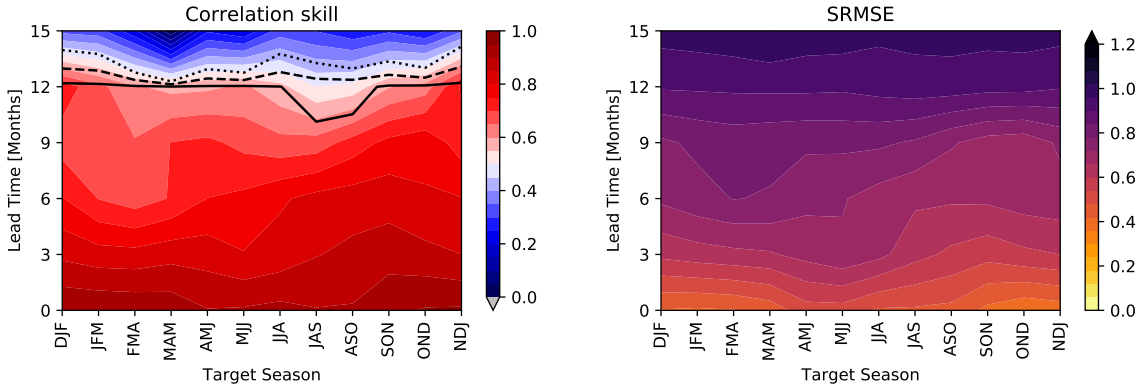


(c)

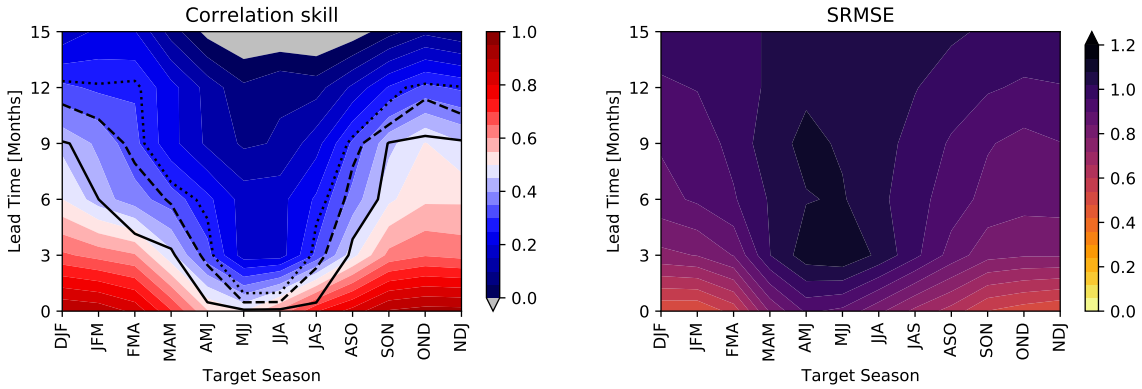
Figure 3.6: *Different skill measures for the DE (solid) in comparison to the persistence (dashed) for various lead times calculated for the entire time series (black) and different decades (colours).* ***(a)*** *shows the Pearson correlation coefficient between the predicted mean and the observed ONI.* ***(b)*** *shows the SSRMSE and* ***(c)*** *the negative loss likelihood.*



(a) 1962-2017



(b) 1982-2001



(c) 1962-1981, 2001-2017

Figure 3.7: *Seasonal prediction skill of the predicted mean from the DEs as functions of target season and lead time. On the left side, the correlation coefficient (contours) for a correlation between predicted mean and the observed ONI is depicted. The solid (dashed) black line shows the 99% (95%)-significance level for the correlation. The right side shows the SSRMSE (contours).*

interdecadal changes for ANN models in prediction skill are also described in other papers of the UBC research group (Tangang et al., 1998a,b).
Moreover, Barnston et al. (1994) shows that the prediction skill for various dynamical and statistical models was relatively strong during the 1980s and worse for the 1960s and 1970s. This is similar to what was observed by Balmaseda et al. (1995). They find a strong spring predictability barrier for the 1970s but not for 1980s. Balmaseda et al. (1995) account this to the strong phase-locking of the ENSO during the 1970s and the weaker locking in the 1980s. Furthermore, as already discussed, Barnston et al. (2012) reported the poor prediction skill of most ENSO forecast models during the 2002-2011.
In Chapter 1 it was hypothesized that this change in ENSO predictability might be connected to the change in the background state of the equatorial Pacific. The background state might strongly alter the importance of the recharge oscillator mechanism over more stochastic effects such as the convective ocean-atmosphere instability. Because of this decadal change, the PC_1 $SSTA_{5\,year}$ was included as a predictor variable for the DEs. It was expected that the DEs would generally estimate larger (smaller) uncertainties for periods in which the PC_1 $SSTA_{5\,year}$ indicates a La Niña-like (El Niño-like) background state. However, the estimated uncertainties of the DEs are not changing strongly throughout different decades. The likely reason for this is, that the DEs which had an El Niño-like decade as test period (1982-1991, 1992-2001) are trained on a rather unbalanced data set considering the changing background state. This means, the majority of the remaining data set was actually coming from more La Niña-like (1962-1981, 2002-2018) and just a small part from El Niño-like periods (either one of the decades 1982-1991 or 1992-2001). A stronger weighting of the importance of the El-Niño-like periods during training probably could have solved this issue partly. However, generally more data from various El Niño-like decades is needed to truly test if the DEs would assign different uncertainties depending on the background state.

3.4.2 Model comparison

For the period from 2003 onwards, it is possible to make a direct comparison of hindcast results of the DE with the forecasts of different statistical and dynamical models for the ENSO prediction. Thanks to the international research institute (IRI) for climate and society, it is easy to access past forecasts from various models[6]. In their data set, predictions are available for a maximum lead time of 8 months for the ONI. However, some models only provide a prediction for smaller lead times, e.g. the CFS(v2) model usually provides forecasts for lead times up to six months. For a brief description of the different models which are used in this comparison, see the Supplement of Barnston et al. (2012).
Fig. 3.8 shows a comparison of the forecasts of various models with the DE for the 0, 3 and the 6-months lead time. The prediction of the DE is, in general, more conservative than the ones from other models. For the 2009/2010 and 2015/2016 El Niño, most of the other models predict an ONI which is larger than the predicted mean of the DE (for all lead times). Furthermore, for the 2010/2011 La Niña, the other models mostly predict an ONI which is smaller than the predicted mean of the DE. However, this more conservative approach of the DE has the advantage that the model has less false positives regarding El Niño/La Niña events (prediction of an El Niño/La Niña that is not observed). This is, for instance, pronounced for late 2003, where the NASA GMAO model falsely predicts a La Niña, while observed conditions are neutral as correctly indicated by the DEs. Similarly, the NASA GMAO (and other dynamics models) falsely predicts El Niño conditions for winter 2011/2012 and 2012/2013, whereas weak La Niña and neutral conditions are observed, respectively. For these two periods, the DE indicates both times neutral conditions.
The comparison of the hindcasts of the DE with the forecasts of other models in Fig. 3.8 indicates that the DE probably could gain some more predictive skill if the predicted distribution would be skewed, i.e. a skewed Gaussian. As for now, a DE needs to find the most likely prediction in terms of a Gaussian. However, a relatively large uncertainty estimation by a DE (large standard deviation) directly forces the predicted mean to be rather neutral. Otherwise it would estimate some events that have not been observed berfore as rather likely[7]. In contrast, for a skewed distribution and the same uncertainty estimation, the predicted mean could better indicate the extreme events by skewing the distribution negatively (positively) for an El Niño (La Niña) event.
A more quantitative comparison of the model skills is given in terms of the Pearson correlation in Fig. 3.9 and the SSRMSE in Fig. 3.10. It generally appears that dynamical models (solid lines) show a better

[6] `https://iri.columbia.edu/~forecast/ensofcst/Data/`

[7] E.g. when the DE would predict a mean of 2.25 K and a relatively large standard deviation of 0.75 K, the DE would estimate that with a probability of about 16% an ONI is going to be observed with a value of 3 K or higher. Which is a very high probability for an event which was never observed up until now.

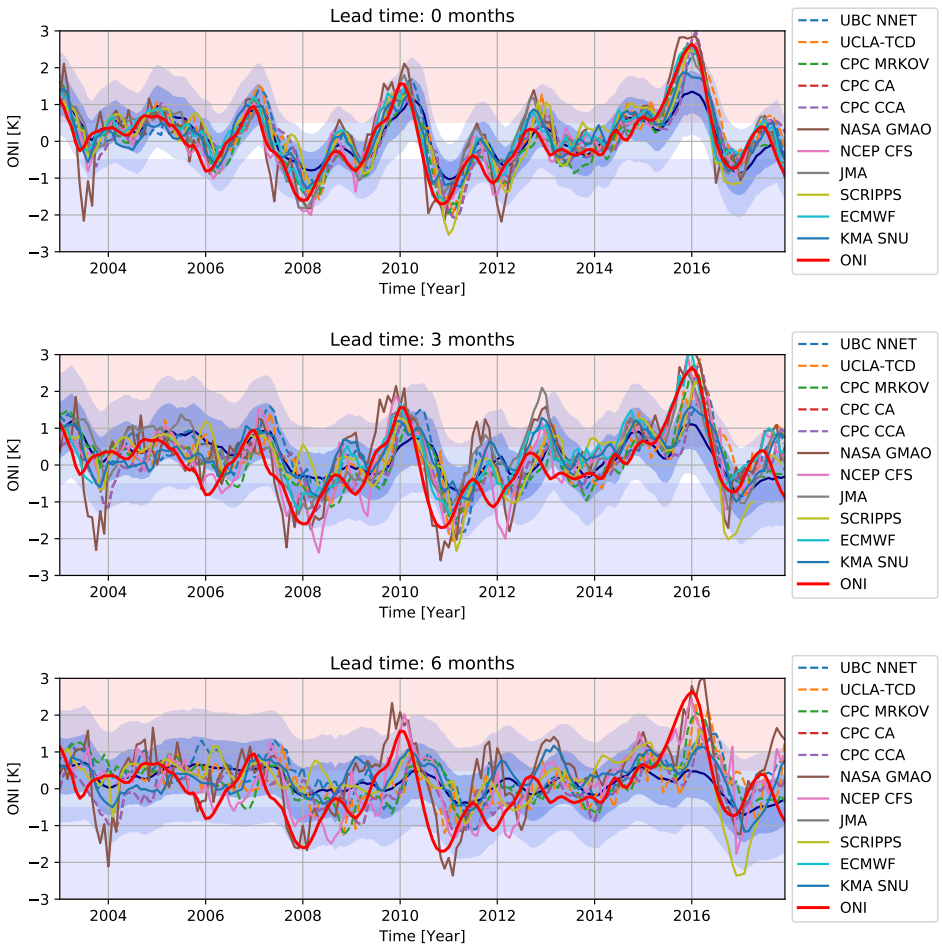


Figure 3.8: *Comparison of forecasts by statistical (dashed) and dynamical models (solid) with the DE for various lead times. Again, the solid blue line refers to the mean of the DE, the darker shading to the 1-standard deviation and the lighter shading to the 2-standard deviation interval around the predicted mean. The observed ONI is indicated by a solid red line.*

prediction skill in comparison to the DE (and other statistical models). Furthermore, the DE does not stand out concerning other statistical models (dashed lines) considering both skill measures and the entire test period (2003-2017, Fig 3.9a and 3.10a). For the sub-period from 2003 till 2011, the DE shows a stronger correlation skill for longer lead times than the other statistical models (see Fig 3.9b). However, for the period 2012-2017 the DE is one of the weaker statistical models considering the correlation skill (see Fig 3.9c). This can be attributed to the poor skill of the DE to predict the 2015/2016 El Niño for longer lead times (see 6 month lead time in 3.8). Considering the SSRMSE, the DEs do not stand out from other statistical models during the considered sub-periods (see Fig. 3.10b and 3.10c).

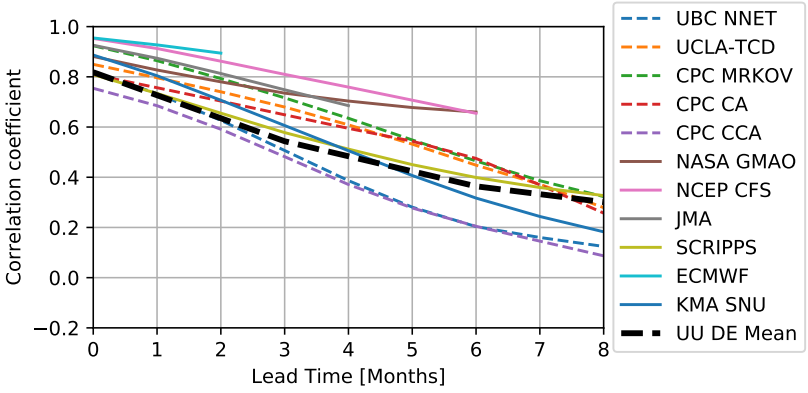
3.4.3 Comment on sensitivity

In the study by Nooteboom et al. (2018), a sensitivity analysis regarding the (hyper)parameter setting is performed to prove that the results are not simply a "lucky shot". However, I do not see the necessity and the value of such an investigation for this study.
For the methodology of the model training in Nooteboom et al. (2018) it is actually necessary to perform such an analysis. This is related to the point, that Nooteboom et al. (2018) are just splitting their data set into a train and a test data set but are not using a validation data set to search for the best model. Hence, the optimization of the (hyper)parameters selects the model that performs best on the test data set after training on the training data set. Therefore, the results on the test data set could really be just a lucky choice of (hyper)parameters that perform by chance good on the test data set.
In contrast to Nooteboom et al. (2018), in this thesis the full times series is split into a train, validation and test data set (see Fig. 3.3). Then, by using a random search in the hyperparameter space, the

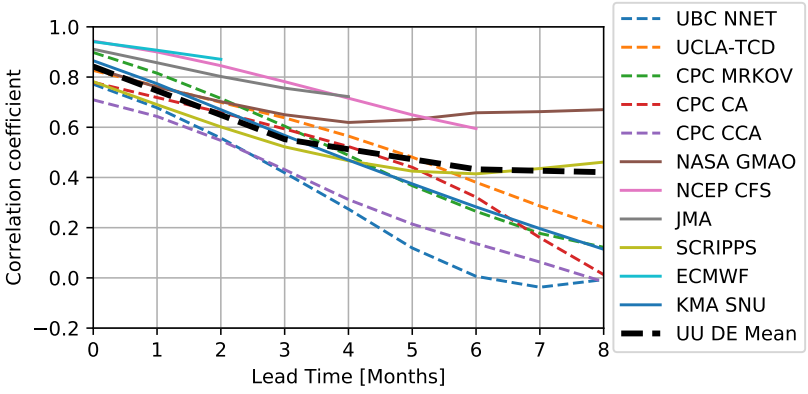
model that had the best results on the validation data set was chosen to be evaluated on the test data set. Given the fact that the hyperparameters are selected by a random search and the model selection is based on the validation data set, it would be rather surprising if the eventually chosen model with its (hyper)parameters would be a “lucky shot” for the test data set.

Table 3.2: *Decadal correlation skills for the 6-months lead time for various ANN models. All target variables are 3-months running-means. The values which are given with the approximate sign "≈" are read off from a graph. The numbers in the parenthesis indicate the exact period for which the correlation skill was computed in the respective study.*

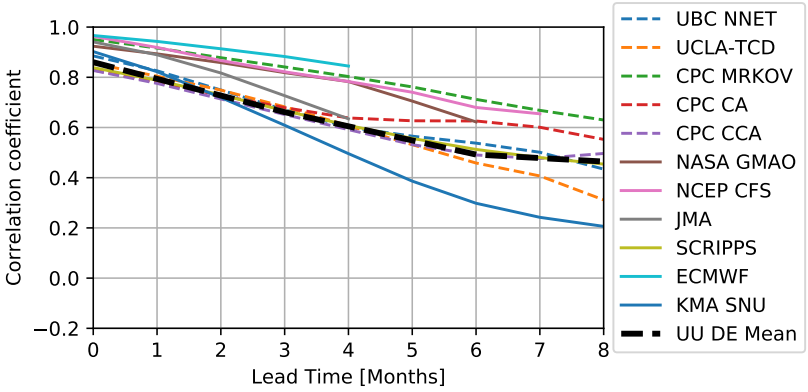
Study	**Target**	**1950s**	**1960s**	**1970s**	**1980s**	**1990s**	**2000s**
This study	ONI	-	0.44 (62-71)	0.47 (72-81)	0.76 (82-91)	0.79 (92-01)	0.48(02-11)
Tangang et al. (1997))	ONI	0.02 (52-62	0.34 (63-72)	0.25 (73-82)	0.70 (83-92)	-	-
Tangang et al. (1998a) SLP model	Niño3.5	≈ 0.7 (52-61)	≈ 0.45 (62-71)	≈ 0.55 (72-81)	≈ 0.65 (82-91)	-	-
Tangang et al. (1998b)	Niño4	≈ 0.75 (52-61)	≈ 0.6 (62-71)	≈ 0.5 (72-81)	≈ 0.75 (82-91)	-	-
Wu et al. (2006)	Niño3	$\approx 0.3 - 0.4$ (55-59)	$\approx 0.4 - 0.48$	$\approx 0.48 - 0.6$	$\approx 0.6 - 0.75$	$\approx 0.46 - 0.6$ (90-95)	-



(a) 2003-2017

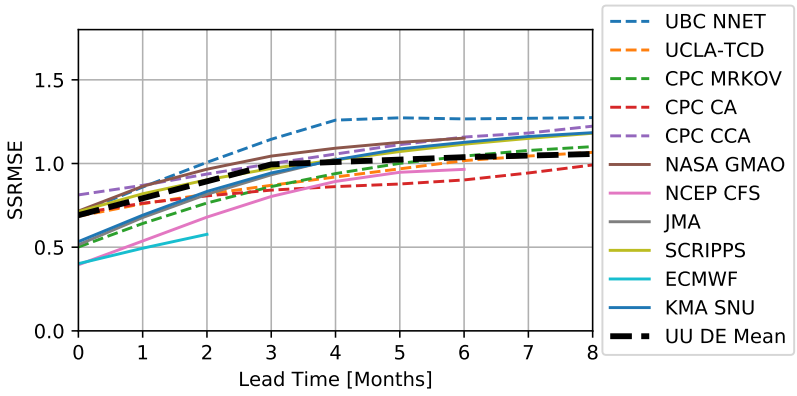


(b) 2003-2011, Similar period as in Barnston et al. (2012).

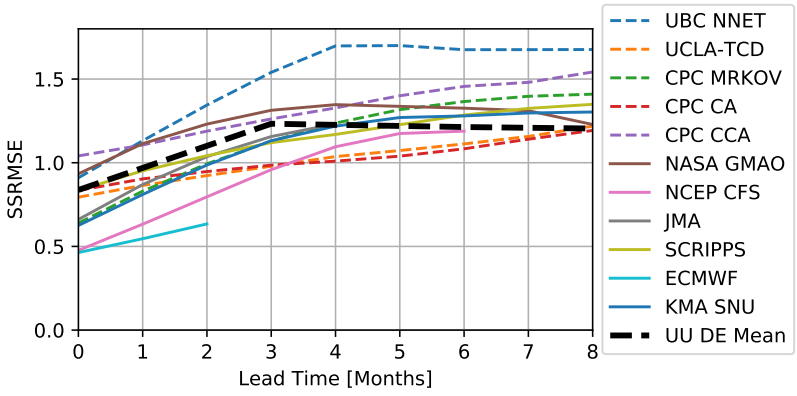


(c) 2012-2017

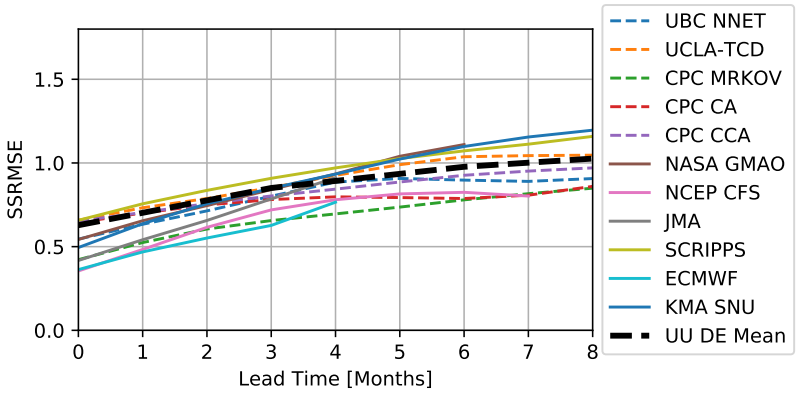
Figure 3.9: *Comparison of the correlation skill of various statistical (dashed) and dynamical models with the correlation skill of the predicted mean from the DE (black, dashed) for the period* ***(a)*** *2003-2017,* ***(b)*** *2003-2011 and* ***(c)*** *2012-2017.*



(a) 2003-2017



(b) 2003-2011, Similar period as in Barnston et al. (2012).



(c) 2012-2017

Figure 3.10: *Comparison of the SSRMSE skill of various statistical (dashed) and dynamical models with the SSRMSE skill of the predicted mean from the DE (black, dashed) for the period* ***(a)*** *2003-2017,* ***(b)*** *2003-2011 and* ***(c)*** *2012-2017.*

Chapter 4

Encoder-Decoder neural networks for ENSO forecasting

4.1 Introduction

Typically, predictor variables for the ENSO forecasting by statistical models are selected by physical reasoning and experimentation. Hence, one needs to have a reasonable understanding of which processes determine the future state of the ENSO. This usually means, one spends a lot of time during the research on finding a working set of predictor variables. This search for good predictors is referred to as *feature engineering*. Up until now, feature engineering for ENSO predictions was mostly done manually and is therefore, for a large part, dependent on human decisions that can be largely biased. Hence, automating parts of the feature engineering process may reduce the human bias on this process and unravel hidden sources of predictability.
In this study, a method inspired by Autoencoders (AEs) will be used. AEs are ANNs that are usually used to reduce the dimensionality of data (Hinton and Salakhutdinov, 2006). They have the same number of neurons in the input and output layer but contain a hidden layer with a considerably lower amount of neurons that acts as a bottleneck (see Fig. 4.1). The feature and the label data of an AE are the same. That is why AEs are a semi-supervised learning method. The part of the ANN from the input layer to the bottleneck is called the *Encoder* because it compresses information to a lower dimension. The part from the bottleneck to the output layer is called the *Decoder*.
Tang and Hsieh (2003) use AEs to perform a nonlinear PCA (NLPCA) to analyze the ENSO. The great advantage of NLPCAs is that one nonlinear PC (NLPC) can account for various spatial anomaly patterns (see Fig. 6 in Tang and Hsieh, 2003), whereas for a linear PCA the spatial anomaly pattern is a constant eigenvector. With the NLPCA they find that the spatial pattern of the first NLPCA mode differs quite significantly between El Niño and La Niña. Hence, the amplitude of the first mode of the NLPCA contains significantly more information than the amplitude of the EOF1 of the linear PCA.
In this study, an AE architecture is used for the ENSO prediction. This is done by introducing a time lag between the feature and label (supervised learning). Hence, the network is forced to derive a low dimensional representation of the future state based on the current state. The described model is called *Encoder-Decoder (ED)* to clearly distinguish it from AEs for which there is no time lag between feature and label.

4.2 Methods and Data

Data sources and some initial preprocessing of the data is already described in Chapter 2. Here, the SSTA field computed from the ERSSTv5 between 30°S to 30°N and 120°E to 80°W on a $2.5° \times 2.5°$ grid is chosen as the feature and the label. Before the data is fed through the ED, each feature is normalized by its standard deviation. After the data is fed through the ED, the output is rescaled again.
The ED is a feed-forward neural network with two hidden layers where the second hidden layer is the bottleneck layer. The exact number of neurons in the hidden layer is later found by a random search. The first hidden layer has the ReLU as activation function and the second bottleneck layer a linear function. The output layer has again the dimension of the input and uses a linear function as activation. In contrast to a "normal" Autoencoder structure, the ED in this study does not contain a hidden layer

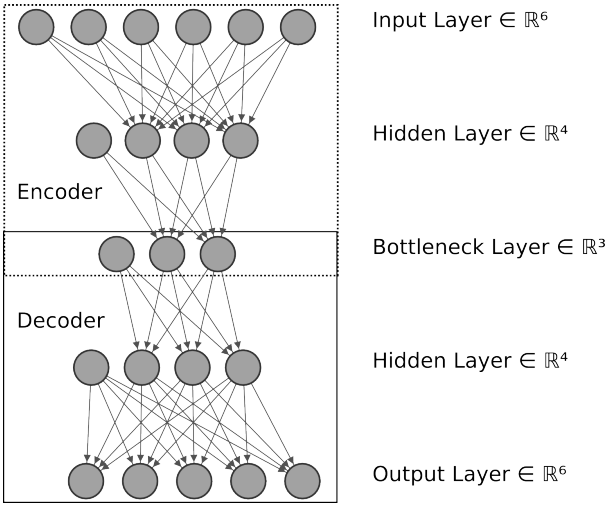


Figure 4.1: *An example of an AE architecture with six input, four hidden layer neurons and three neurons in the bottleneck layer and five output neurons. Note, that in all layers but the output layer one neuron accounts for the bias term in the subsequent layer by having a constant output equal to 1. For the hidden layers and the bottleneck layer, this neuron is not connected to the previous layer.*

in the decoder part. Hence, the decoder just can linearly combine the outputs from the bottleneck layer to generate one of the outputs. This is done to avoid overfitting that could be generated by the decoder hidden layer by allowing for any nonlinear combination of the bottleneck outputs.
Again, the ensemble method from the previous chapter is used (see Fig. 3.3). The prediction of the ensembles is the average of the predictions by the members. In contrast to the DE, the loss function is the mean squared error (MSE). The ED is regularized by Gaussian noise layers which are installed after the input and the output layer and by the L_1 and L_2 penalty terms that are applied the to weights. Next to this, dropout is applied for the first hidden layer. Finaly, Early Stopping is applied such that training is stopped when the MSE on the validation segment starts to increase again.
Hyperparameters are again optimized using a random search. This is done by a random uniform choice for the number of neurons in the first hidden layer (integer between 32 and 512), the number of neurons in second hidden layer (integer between 8 and 64), the dropout rate (float between 0 and 0.2), the standard deviation of the Gaussian noise (float between 0 and 0.5), the magnitude of the L_1 and L_2 penalty terms (float between 0 and 0.001) as well as the learning rate (float between 10^{-3} and 0.01). A member of the ED is trained by backpropagation using the Adam optimizer (Kingma and Ba, 2014).

4.3 Results

Some example 6-months lead time hindcasts for the winter season (NDJ) for selected El Niño/La Niña events are shown in Fig. 4.2. Fig. 4.2a shows the hindcast and the observation for the strong El Niño event in 1997/1998. Weak anomalies are predicted by the ED ensemble. Moreover, the spatial pattern of the hindcast by the ED is more a CP type whereas the oberved type is an EP El Niño. A qualitatively better hindcast is made for the subsequent La Niña (see Fig. 4.2b). Only in the southern EP, the ED predicts cold anomalies while in reality warm anomalies are present. A very poor hindcast by the ED is made for the El Niño in 2009/2010 (see Fig. 4.2c). Here, the ED even predicts cold anomalies in the equatorial Pacific. The subsequent La Niña is, again, predicted with better accuracy (see Fig. 4.2d).
Fig. 4.3 shows the correlation coefficients for a correlation between the predicted and the observed SSTAs. For the 0-months lead time, the correlation coefficients are high throughout the equatorial Pacific. However, for longer lead times, the correlation skills rapidly drops with the strongest decrease in the EP. For long lead times of 9 months and longer, the coefficients just remain relatively high in the WP ($r \approx 0.5 - 0.6$). On the one hand, during El Niño-like periods, correlation skills stay high for the CP, whereas they mostly drop for the rest of the Pacific (see Fig. 4.4a). On the other hand, during La-Niña-like periods, the correlation skills remain high for long lead times in the WP, whereas the ED has no prediction skill in the EP and CP.
The predicted ONI is shown in Fig. 4.5. The ED prediction just shows clear signals for El Niño and La Niña events for lead times up to 6 months. Afterwards, the prediction just slightly fluctuates around

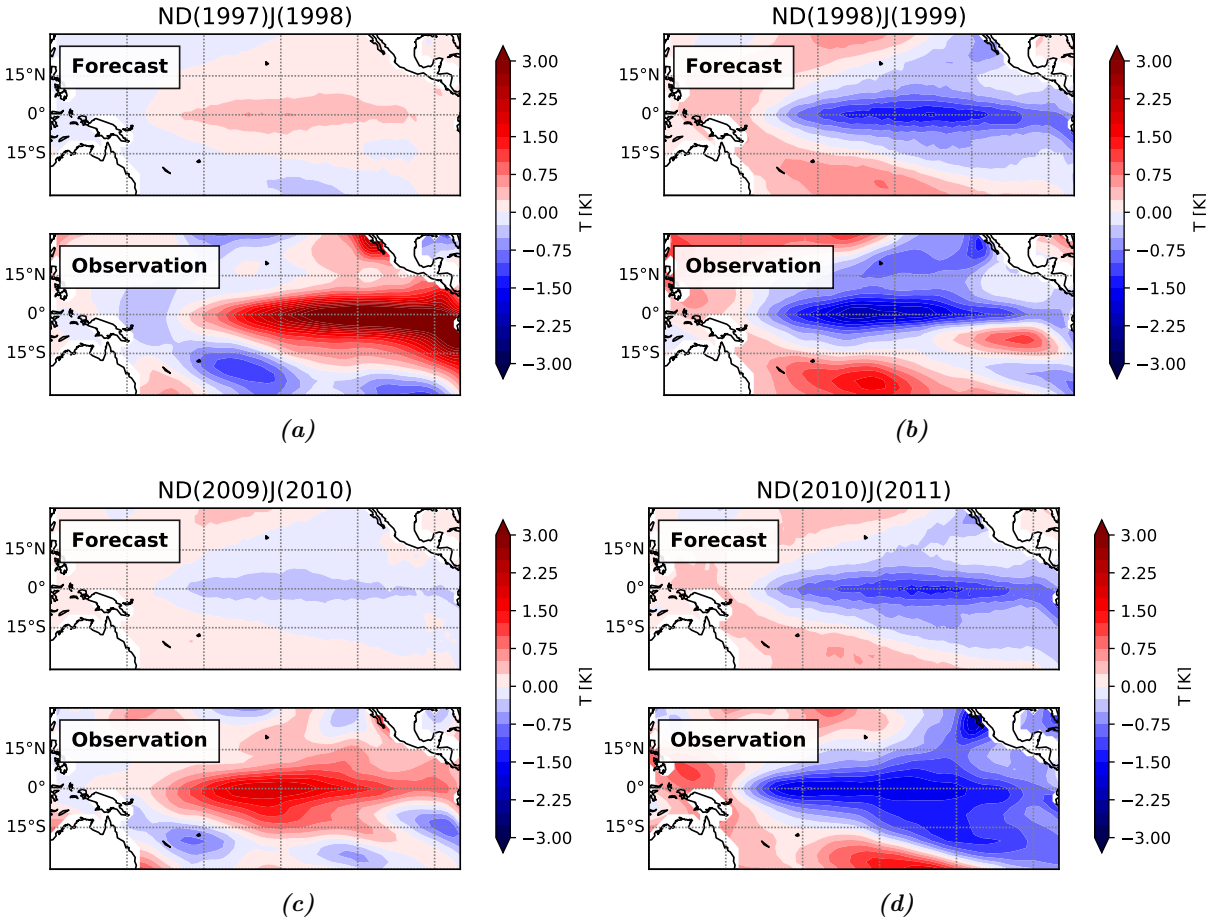


Figure 4.2: *Example predictions for the 6-months lead time by the ED in comparison to the observations for the winter period (NDJ) for some extreme events. Plots are made for* ***(a)*** *the El Niño event in 1997/1998,* ***(b)*** *for the subsequent La Niña in 1998/1999,* ***(c)*** *the El Niño in 2009/2010 and* ***(d)*** *again for the subsequent La Niña in 2010/2011.*

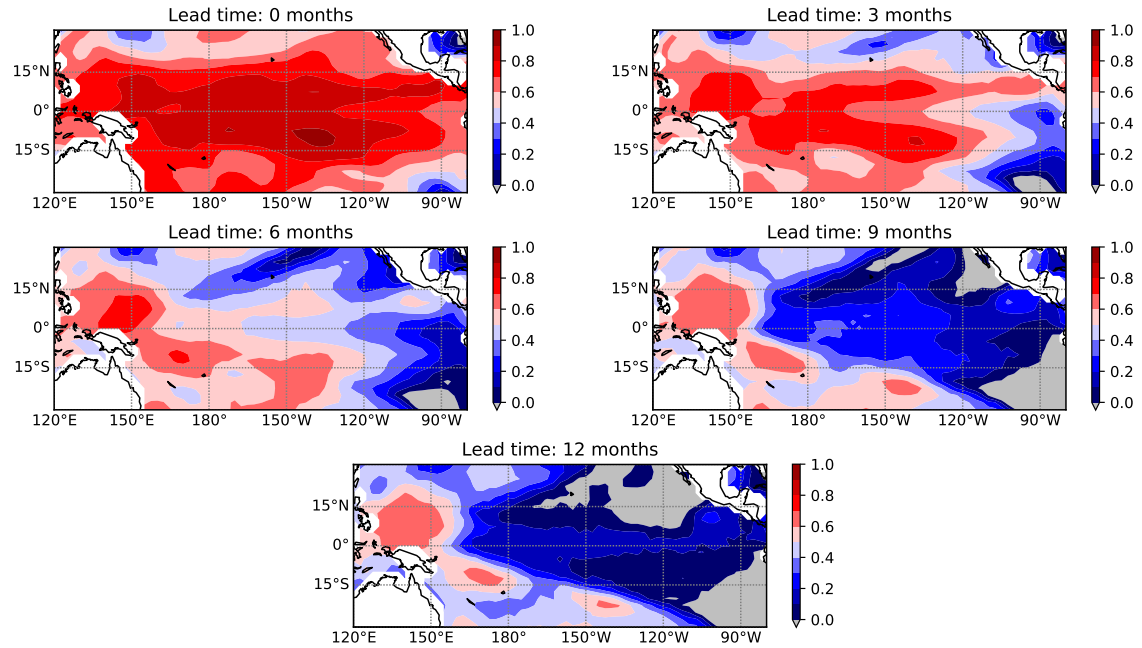
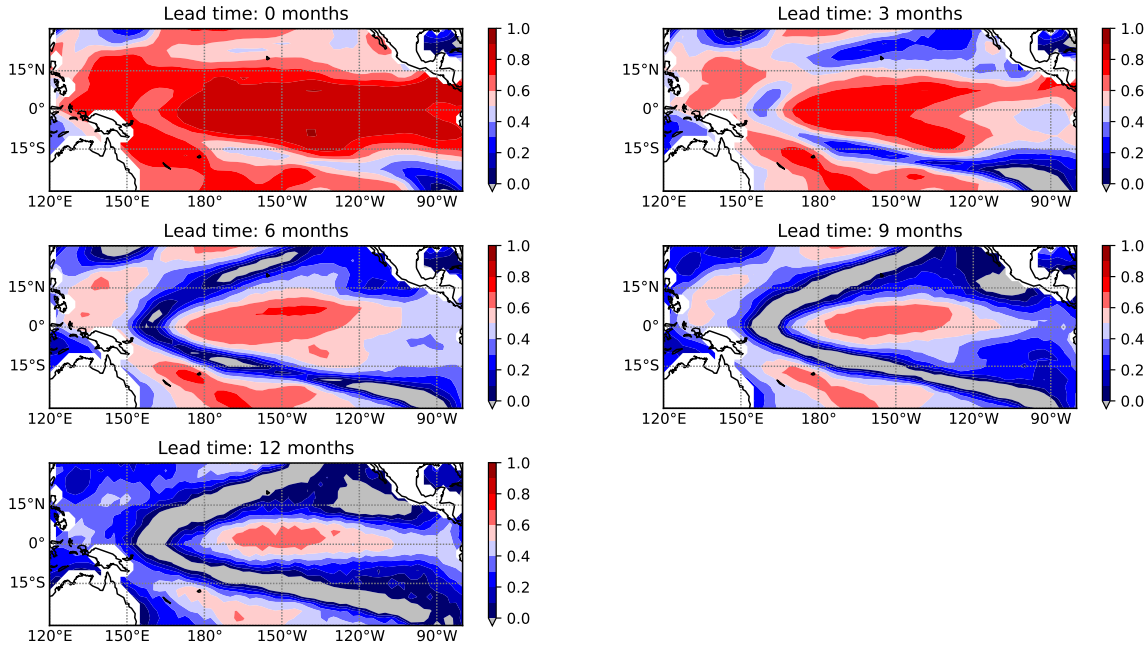
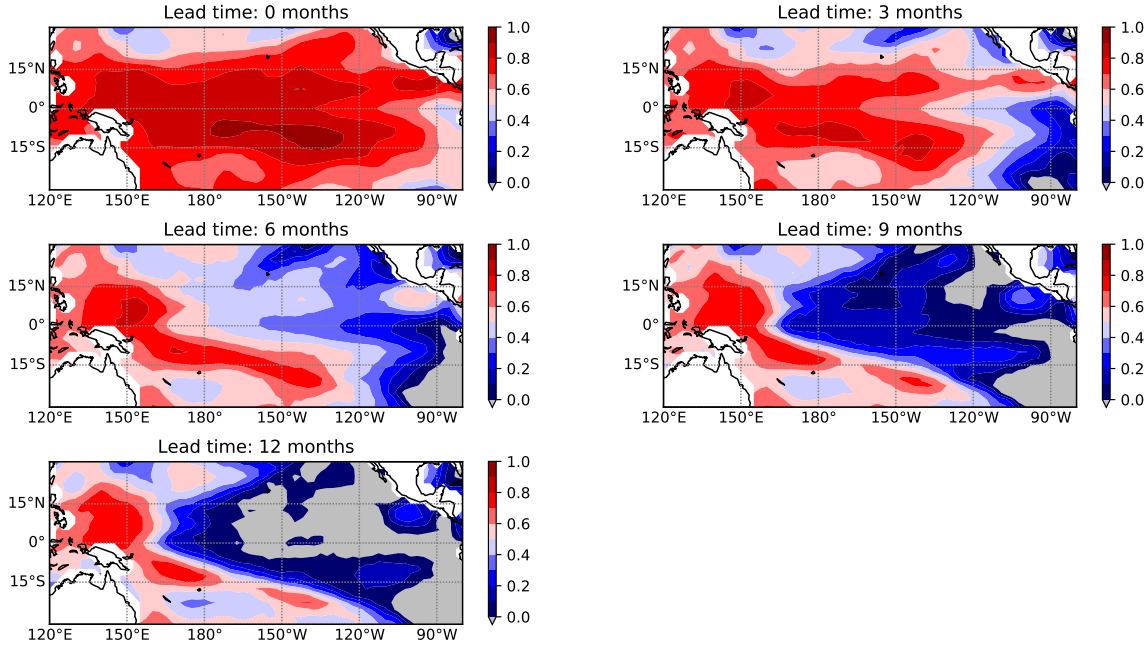


Figure 4.3: *Pearson correlation coefficients for correlations between predicted and observed SSTAs for various lead times.*



(a) 1982-2001



(b) 1962-1981, 2001-2017

Figure 4.4: *As Fig. 4.3 but for* ***(a)*** *El Niño-like periods (1982-2001) and* ***(b)*** *La Niña-like periods (1962-1981, 2001-2017).*

0 K. The correlation skill (see Fig. 4.7a) and the SSRMSE (see Fig. 4.7b) in terms of the ONI still show that the ED beats the persistence forecast for all lead times. The prediction skills of the ED are overall worse than for the one for DE model from Chapter 3. Interestingly, the same characteristic regarding the decadal change in prediction skill can be seen. While better prediction skills are archived for the period between 1982-2002 (12-months lead, $r \approx 0.6$, SSRMSE ≈ 0.8), prediction skills are considerably worse for the other decades ($r \approx 0.0 - 0.3$, SSRMSE $\approx 1 - 1.5$). Note, although there is a considerable correlation for longer lead times for the period 1982-2002, the actual prediction showed a nearly constant prediction of and ONI of 0 K (see Fig. 4.5). Hence, although the correlation coefficient is rather high, the actual prediction made by the ED is not very useful for long lead times.
The seasonal prediction skills are shown in Fig. 4.7. As for the DE, forecast skills are lowest for the summer season and are best for the winter season. In the comparison to the DE, the ED has similar prediction skill for the summer seasons but has less skill regarding the other seasons with the biggest differences for the winter seasons (significant skill of the DE extends 3-months further).

4.4 Discussion

The results in this chapter provide a proof of concept that the ED is capable in making skillful predictions for the ENSO. And hence it proves that the bottleneck architecture of the model could keep the model from severe overfitting. The bottleneck, therefore, forced the ANN to focus on relatively strong relations between feature and label and disregard noise as a source for information. It is even rather surprising that the ED model was already able to make skillful forecasts given the little amount of data (696 months) with respect to the complexity of the prediction (1625 grid points). This indicates that the evolution of the ENSO is a relatively low dimensional problem.
Whereas El Niño events are poorly predicted by the ED, La Niña events seem to be more predictable for the ED. This is likely because observed El Niño events are often followed by La Niña events, whereas the opposite does not usually happen (Clarke and Zhang, 2019). Furthermore, the ED shows the same qualitative features regarding the interdecadal change of prediction skills as other forecast models.
Since the ED generally performed worse than the DE and the DE had an average prediction skill in comparison to other statistical models, it follows that the ED generally scores worse than most of the existing statistical models. To improve the performance of the ED, it would be interesting to investigate if pretraining the model weights on a vast amount of data from climate models could improve the results. Moreover, it might be interesting to include other data fields such as the heat content anomaly as features as well as use a convolution neural network (CNN) instead of a simple MLP for the ED.

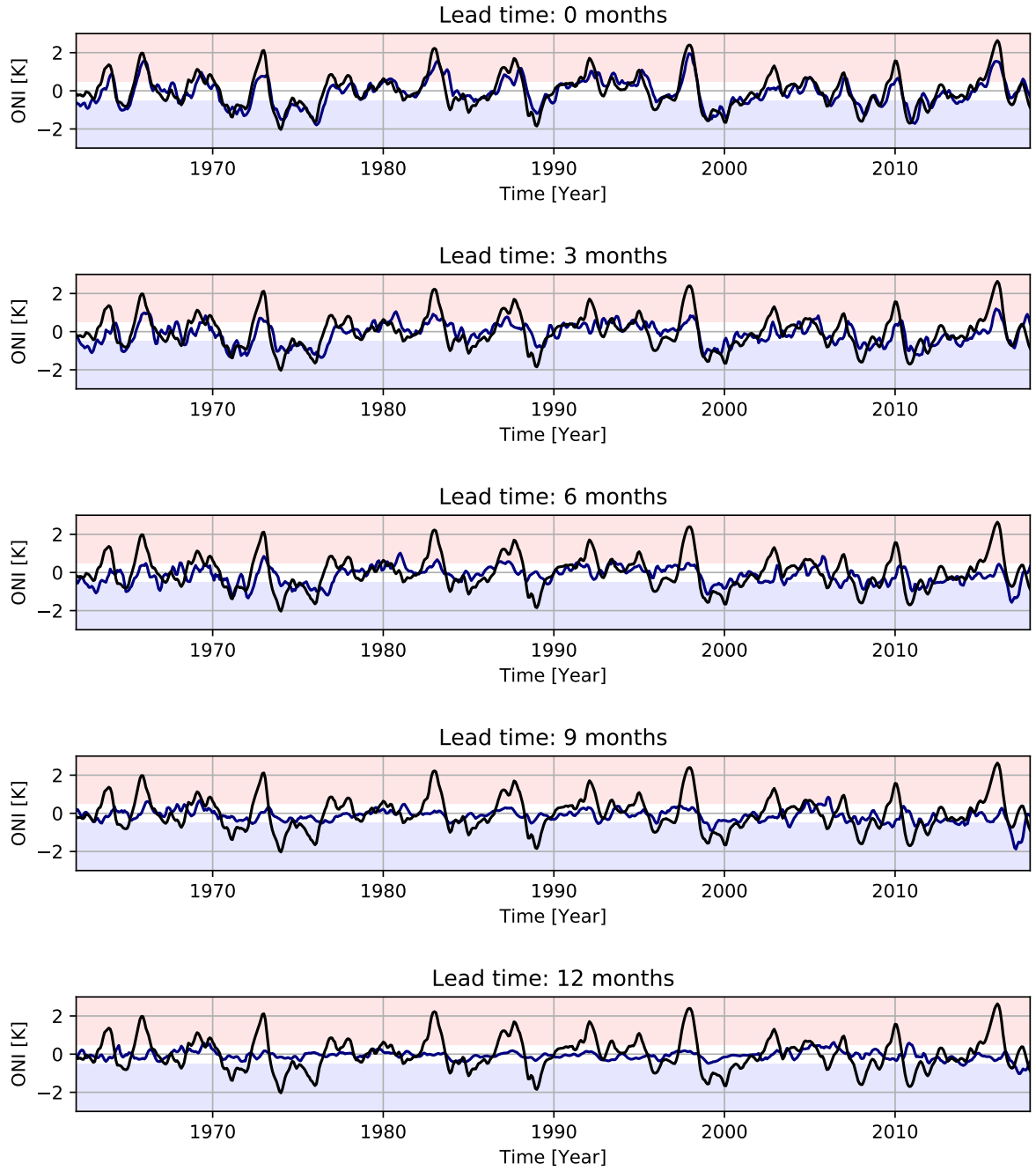
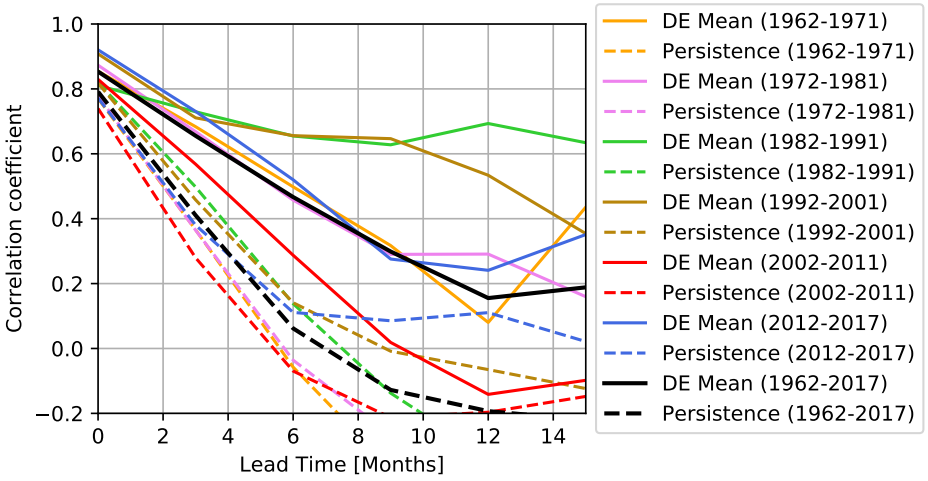
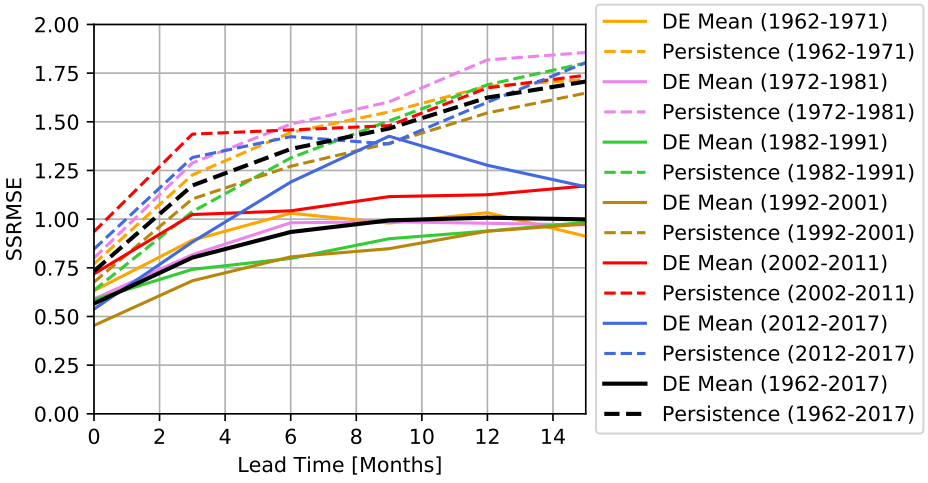


Figure 4.5: *Predictions of the ED for various lead times. The solid blue line depicts is the predicted ONI and the black line the observed value.*

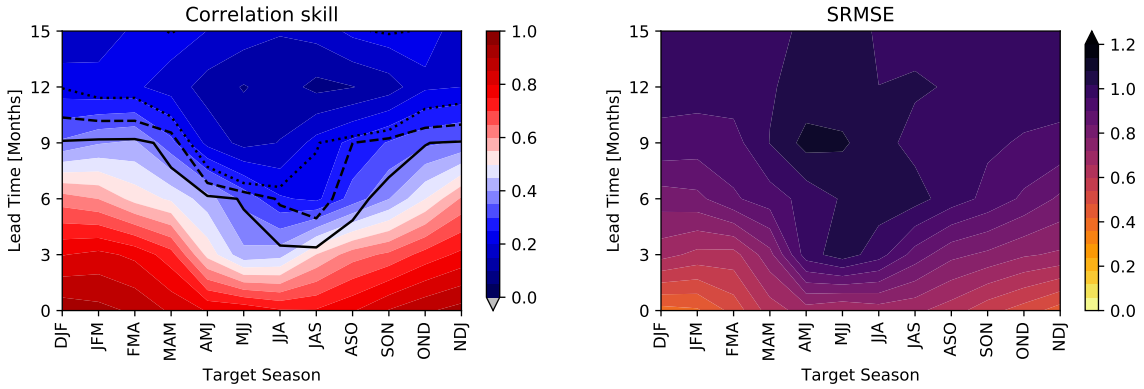


(a)

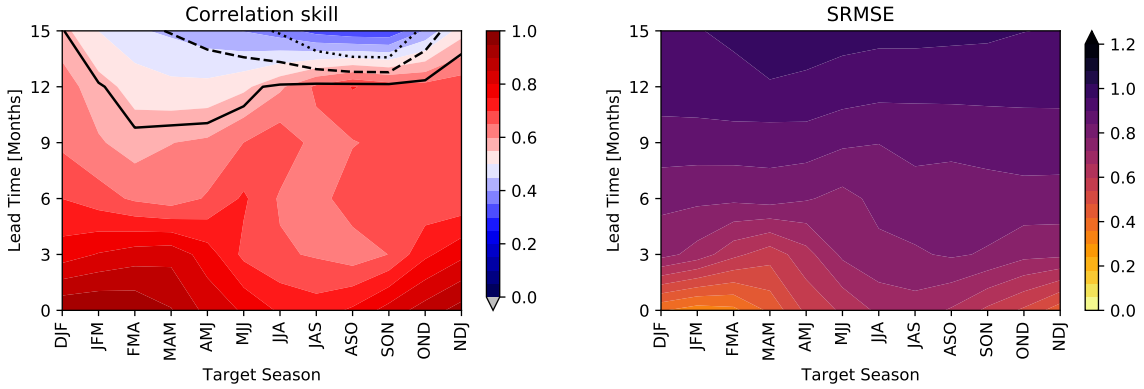


(b)

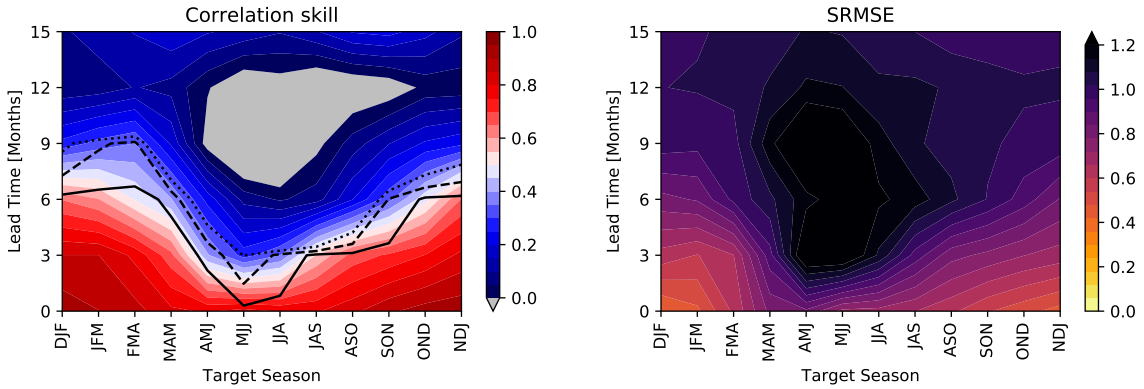
Figure 4.6: *Different skill measures for the ED (solid) in comparison to the persistence (dashed) for various lead times calculated for the entire time series (black) and different decades (colors). The plot in* ***(a)*** *shows the Pearson correlation coefficient between the predicted mean and the observed ONI. The plot in* ***(b)*** *shows the SSRMSE.*



(a) 1962-2017



(b) 1982-2001



(c) 1982-2001

Figure 4.7: *Seasonal prediction skill of the predicted ONI of the ED as functions of target season and lead time. On the left side, the correlation coefficient (contours) for a correlation between predicted mean and the observed ONI is depicted. The solid (dashed) black line shows the 99% (95%)-significance level for the correlation. The right side shows the SSRMSE (contours).*

Chapter 5

Conclusion and outlook

Two artificial neural network (ANN) models were examined for the ENSO prediction in this study. The first model is a so-called Deep Ensemble (DE), which is a feed-forward neural network that predicts a Gaussian distribution. Hence, the forecast for the ENSO comes with an estimation of the predictive uncertainty. This is a novel feature for statistical ENSO prediction. In general, there was no clear indication that the DE can overcome the spring predictability barrier in a meaningful way. This means, although some extreme events such as the El Niño in 1997/1998 showed up with a signal in the hindcast already for the 12-months lead time (i.e. forecast issued in autumn 1996), the forecast was still accompanied by large predictive uncertainty. Moreover, the prediction skill of the DE does not stand out from other statistical models. However, it was noticeable that the forecasts by the DE are generally more conservative than the forecasts by other statistical models in the prediction of extreme events. Changing the predicted distribution to a skewed distribution might considerably improve the skill of the model.
The second model that was investigated is an Encoder-Decoder (ED) model which is inspired by the architecture of Autoencoders. The advantage of the ED is that the predictive scheme can take spatial information into account as well as predict the entire sea surface temperature anomaly (SSTA) field. Despite the vast amount of input data (SSTA values from 1625 grid points) and the comparably low amount of observed periods (696 months), the ED was still able to make skillful predictions. This can be attributed to the effective regularization of the network as well as to the presence of the bottleneck layer in the architecture of the ED. This bottleneck forces the ANN to reduce the dimensionality of the input. Hence, the ANN is forced to focus on strong relations between the input and the output field and disregard noise. This avoids that the ANN starts to use noise to explain the target variables, which is the usual reason for the overfitting of ANN models.
As other models for the ENSO prediction, the DE and the ED show a strong interdecadal change of prediction skill. In a warm (cold) background state the predictive skills are higher (lower) and the spring predictability barrier is weaker (stronger). An extended literature analysis led to a hypothesis on how to explain the decadal variations of the ENSO predictability: A relatively warm background state of the Pacific (El Niño-like periods) causes a weak self-limitation of the ENSO. This, in turn, enables the deterministic recharge/discharge oscillator to effectively determine the evolution of the ENSO. In contrast, self-limitation is strong when the background state is relatively cold (La Niña-like periods) which limits the influence of the recharge oscillator and makes the ENSO dynamics more stochastic and less predictable. A brief data analysis indicated that the decadal variations of the background state of the Pacific showed that equatorial convection over the warm pool is shifted towards the East during warm periods. This causes a weaker self-limitation following arguments from Hameed et al. (2018). It was, unfortunately, beyond the scope of this study to prove this hypothesis more rigorously. To prove this hypothesis, extended climate model simulations are needed. First, the hypothesis of Clarke (1994) about convectively generated westerlies must be quantified more precisely. This means one needs to prove that although convection mostly excites atmospheric Rossby waves with westerly wind anomalies at the equator, atmospheric Kelvin waves are excited as well. Moreover, these Kelvin waves need to be strong enough in a way that an effective self-limitation is possible. Afterwards, one needs to show that the larger (smaller) fetch of atmospheric Kelvin waves over the equatorial Pacific would increase (decrease) the self-limitation and that this increased (decreased) self-limitation is accompanied by a lower (higher) predictability of the ENSO.

Appendix A

NinoLearn - A research framework for statistical ENSO forecasts

During the investigation for this study, the research framework NinoLearn was developed that is available on GitHub (`https://github.com/pjpetersik/ninolearn`). In this repository, one can find the specific scripts that were used in this study in the subdirectory `research/Master_thesis`. The subdirectory `ninolearn` contains the actual NinoLearn package. The documentation with an installation guideline (for Linux) and some tutorials on how to use NinoLearn can be found under the link `https://pjpetersik.github.io/ninolearn`.
NinoLearn comes as a Python package. It was developed to facilitate research for statistical ENSO forecasts. Moreover, Ninolearn makes research more transparent, reproducible and reusable as well as comparable with other research in statistical ENSO prediction. Lastly, NinoLearn aims to establish conventions, i.e. lead time definition as in Barnston et al. (2012), and best practices for statistical ENSO forecasts, i.e. data split into train, validation and test data set as well as splitting the data into connected time series. To sum up, NinoLearn aims to standardize and accelerate the production steps which are involved in the build-up of a statistical model for the forecast of the ENSO. In the following, the core functionalities of NinoLearn are introduced.

A.1 Download

Each research for the development of a new statistical model for ENSO forecasts starts with the download of data. Most often researchers use the same data sets for their research. However, for each of the data sets, they need to write own downloading scripts. This is time-consuming work. With NinoLearn researchers can easily download data from the common sources (e.g. NCEP/NCAR reanalysis) as well as add a new data source if it is not included into NinoLearn.

A.2 Read raw data

After the data is downloaded. One always needs to figure out how to open the data set and what are the specifics of the data set:

- How to open the specific format (.csv, .nc, .mat, etc.)?
- How many header lines does the csv-file have?
- What is the exact name of the variable one is interested in?
- Do I need to write special functions to extract the data I am interested in?
- etc.

To get the data from a raw file into a format one can easily work with can be a time-consuming and tedious work. NinoLearn already contains a collection of reading routines for specific data sets. If no reading routine is available for particular data sets, researchers can write a new routine and easily share this routine through NinoLearn afterwards with the public.

A.3 Prepare and preprocess data

To make the raw data useful for the statistical models, often one wants to clean the raw data, harmonize it with other data sets (e.g. same time axis format) and preprocess the data. Writing code for preprocessing routines such as evolving complex networks (ECN) can be exhausting work. Hence, NinoLearn also aims here to serve as a pool where researchers can collectively work on writing these routines.

A.4 Statistical models

A lot of models have been proposed in the past for statistical ENSO prediction. However, the model code is often not publicly available. If it is available, the code is sometimes not written in a standardized way, which makes is technically difficult to fastly reuse the code during own research. NinoLearn aims to standardize the build-up of a statistical model. This is archived by demanding a certain coding structure, i.e. models need to be Python classes that need to have functions that follow naming convections:

- `.set_parameters(*args)` - the function to set the hyperparameters of the model
- `.fit(X,y)` - the method to train the model
- `.predict(X)` - the method to make predictions with the trained model
- `.evaluate(true, pred)` the method to evaluate the prediction of the model
- `.save(path)` - save the model
- `.load(path)` - load the model

A.5 Evaluation

When statistical models were evaluated in the past, often researchers used different methods to evaluate them which make it difficult to compare the results. With NinoLearn also this process shall be standardized in the future in such a way that the performance of the model is evaluated onto a common period e.g. 1962 - 2017 as well as 1962 - today using cross-testing scheme introduced in Chapter 3 (see Fig 3.3). Furthermore, if researchers want to introduce a new evaluation method, they can easily compare it to the skill of other modelsbecause with NinoLearn they have these models and the data sets that these models use easily available.

Appendix B

Acronyms

AE Autoencoder

ANN artificial neural network

ARIMA autoregressive integrated moving average

BNN Bayesian neural network

CN complex network

CP central Pacific

DE Deep Ensemble

DMI Dipole Mode Index

ECN evolving complex network

ED Encoder-Decoder

EEOF Extended Empirical Orthogonal Function

ENSO El Niño-Southern Oscillation

EOF Empirical Orthogonal Function

EP eastern Pacific

IOD Indian Ocean Dipole

ML machine learning

MLP Multilayer Perceptron

MSE mean-squared error

NLL negative-log-likelihood

NLPCA non-linear principal component analysis

OLR outgoing longwave radiation

OLRA outgoing longwave radiation anomaly

ONI Oceanic Niño Index

PCA Principal Component Analysis

PDO Pacific Decadal Oscillation

ReLU Rectified Linear Unit

RMSE root-mean-squared error

SC seasonal cycle

SLP sea level pressure

SLPA sea level pressure anomaly

SRMSE standardized root-mean-squared error

SSH sea surface height

SSHA sea surface height anomaly

SSRMSE seasonally-standardized root-mean-squared error

SST sea surface temperature

SSTA sea surface temperature anomaly

UBC University of British Columbia

WP western Pacific

WWB westerly wind burst

WWV warm water volume

Acknowledgment

This master thesis was a very exciting journey for me into the world of machine learning and ENSO research. First of all, I want to thank my supervisor Henk Dijkstra for his great supervision during this research project. As for my Complex Systems Research Project, I enjoyed the productive meetings that we had and his way of supervision that always encouraged me to follow my ideas. Furthermore, I want to thank the Potsdam Institute for Climate Impact Research (PIK) for having me as a guest for two weeks in May 2019. I want especially thank Reik Donner and Jonathan Donges for making my stay at the PIK possible. During these two weeks, I got great inspiration for my research project, especially regarding the CP and EP El Niño types, and enjoyed lots of discussions with researchers from the PIK. Moreover, I want to thank Markus Abel from the company 4cast for the possibility to present my research at his company and the discussion of my research with him during which the idea for the Encoder-Decoder model was born.

Danksagung

Am Ende meines Studiums angelangt, möchte ich an dieser Stelle meiner Familie, meinen Freunden und meiner Freundin Janna danken. Viele Dinge in meinem Studium wären ohne die Unterstützung von euch nicht möglich gewesen. Angefangen bei der finanziellen Unterstützung, über die Hilfe bei den zahlreichen Umzügen hinzu der emotionalen Unterstützung in den eher schwierigen Zeiten meines Studiums. Mein Studium der Meteorologie und Klimaphysik hat mich zwar das ein oder andere Mal ganz schön herausgefordert, aber ist im Großen und Ganzen doch voll von spannenden Erfahrungen und Erkenntnissen gewesen. Vielen Dank, dass ihr auf diesem Weg meine treuen Begleiter gewesen seid.

Bibliography

An, S.-I. and Jin, F.-F. (2000). An eigen analysis of the interdecadal changes in the structure and frequency of ENSO mode. *Geophysical Research Letters*, 27(16):2573–2576.

Baawain, M. S., Nour, M. H., El-Din, A. G., and El-Din, M. G. (2005). El Niño southern-oscillation prediction using southern oscillation index and Niño3 as onset indicators: Application of artificial neural networks. *Journal of Environmental Engineering and Science*, 4(2):113–121.

Balmaseda, M. A., Davey, M. K., and Anderson, D. L. T. (1995). Decadal and Seasonal Dependence of ENSO Prediction Skill. *J. Climate*, 8(11):2705–2715.

Balmaseda, M. A., Mogensen, K., and Weaver, A. T. (2013). Evaluation of the ECMWF ocean reanalysis system ORAS4. *Quarterly Journal of the Royal Meteorological Society*, 139(674):1132–1161.

Barnston, A. G., Tippett, M. K., L'Heureux, M. L., Li, S., and DeWitt, D. G. (2012). Skill of Real-Time Seasonal ENSO Model Predictions during 2002–11: Is Our Capability Increasing? *Bull. Amer. Meteor. Soc.*, 93(5):631–651.

Barnston, A. G., van den Dool, H. M., Zebiak, S. E., Barnett, T. P., Ji, M., Rodenhuis, D. R., Cane, M. A., Leetmaa, A., Graham, N. E., Ropelewski, C. R., Kousky, V. E., O'Lenic, E. A., and Livezey, R. E. (1994). Long-Lead Seasonal Forecasts—Where Do We Stand? *Bull. Amer. Meteor. Soc.*, 75(11):2097–2114.

Bjerknes, J. (1969). Atmospheric teleconnections from the equatorial pacific. *Mon. Wea. Rev.*, 97(3):163–172.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight Uncertainty in Neural Networks. *arXiv:1505.05424 [cs, stat]*. arXiv: 1505.05424.

Bunge, L. and Clarke, A. J. (2014). On the Warm Water Volume and Its Changing Relationship with ENSO. *J. Phys. Oceanogr.*, 44(5):1372–1385.

Capotondi, A., Wittenberg, A. T., Newman, M., Di Lorenzo, E., Yu, J.-Y., Braconnot, P., Cole, J., Dewitte, B., Giese, B., Guilyardi, E., Jin, F.-F., Karnauskas, K., Kirtman, B., Lee, T., Schneider, N., Xue, Y., and Yeh, S.-W. (2014). Understanding ENSO Diversity. *Bull. Amer. Meteor. Soc.*, 96(6):921–938.

Chen, D., Lian, T., Fu, C., Cane, M. A., Tang, Y., Murtugudde, R., Song, X., Wu, Q., and Zhou, L. (2015). Strong influence of westerly wind bursts on El Niño diversity. *Nature Geoscience*, 8(5):339–345.

Clarke, A. J. (1994). Why Are Surface Equatorial ENSO Winds Anomalously Westerly under Anomalous Large-Scale Convection? *J. Climate*, 7(10):1623–1627.

Clarke, A. J. (2014). El Niño Physics and El Niño Predictability. *Annu. Rev. Mar. Sci.*, 6(1):79–99.

Clarke, A. J. and Zhang, X. (2019). On the Physics of the Warm Water Volume and El Niño/La Niña Predictability. *J. Phys. Oceanogr.*, 49(6):1541–1560.

Csáji, B. C. (2001). Approximation with artificial neural networks. *Faculty of Sciences, Etvs Lornd University, Hungary*, 24:48.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Math. Control Signal Systems*, 2(4):303–314.

Diaz, H. F., Hoerling, M. P., and Eischeid, J. K. (2001). ENSO variability, teleconnections and climate change. *International Journal of Climatology*, 21(15):1845–1862.

Dijkstra, H. A. (2008). *Dynamical oceanography*. Springer Science & Business Media.

Donges, J. F., Zou, Y., Marwan, N., and Kurths, J. (2009). Complex networks in climate dynamics: Comparing linear and nonlinear network construction methods. *The European Physical Journal Special Topics*, 174(1):157–179.

Duan, W. and Wei, C. (2013). The 'spring predictability barrier' for ENSO predictions and its possible mechanism: results from a fully coupled model. *International Journal of Climatology*, 33(5):1280–1292.

Geman, S. and Geman, D. (1987). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. In Fischler, M. A. and Firschein, O., editors, *Readings in Computer Vision*, pages 564–584. Morgan Kaufmann, San Francisco (CA).

Hameed, S. N., Jin, D., and Thilakan, V. (2018). A model for super El Niños. *Nature Communications*, 9(1):2528.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507.

Horii, T., Ueki, I., and Hanawa, K. (2012). Breakdown of ENSO predictors in the 2000s: Decadal changes of recharge/discharge-SST phase relation and atmospheric intraseasonal forcing. *Geophysical Research Letters*, 39(10).

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257.

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., Menne, M. J., Smith, T. M., Vose, R. S., and Zhang, H.-M. (2017). Extended Reconstructed Sea Surface Temperature, Version 5 (ERSSTv5): Upgrades, Validations, and Intercomparisons. *J. Climate*, 30(20):8179–8205.

Izumo, T., Vialard, J., Lengaigne, M., de Boyer Montegut, C., Behera, S. K., Luo, J.-J., Cravatte, S., Masson, S., and Yamagata, T. (2010). Influence of the state of the Indian Ocean Dipole on the following year's El Niño. *Nature Geoscience*, 3(3):168–172.

Jin, F.-F. (1997). An Equatorial Ocean Recharge Paradigm for ENSO. Part I: Conceptual Model. *J. Atmos. Sci.*, 54(7):811–829.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., and Joseph, D. (1996). The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, 77(3):437–472.

Kao, H.-Y. and Yu, J.-Y. (2009). Contrasting Eastern-Pacific and Central-Pacific Types of ENSO. *J. Climate*, 22(3):615–632.

Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*. arXiv: 1412.6980.

Kirtman, B. P. and Schopf, P. S. (1998). Decadal Variability in ENSO Predictability and Prediction. *J. Climate*, 11(11):2804–2822.

Kug, J.-S., Jin, F.-F., and An, S.-I. (2009). Two Types of El Niño Events: Cold Tongue El Niño and Warm Pool El Niño. *J. Climate*, 22(6):1499–1515.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 6402–6413. Curran Associates, Inc.

Liebmann, B. and Smith, C. A. (1996). Description of a Complete (Interpolated) Outgoing Longwave Radiation Dataset. *Bulletin of the American Meteorological Society*, 77(6):1275–1277.

Ludescher, J., Gozolchiani, A., Bogachev, M. I., Bunde, A., Havlin, S., and Schellnhuber, H. J. (2014). Very early warning of next El Niño. *Proceedings of the National Academy of Sciences*, page 201323058.

Mantua, N. J., Hare, S. R., Zhang, Y., Wallace, J. M., and Francis, R. C. (1997). A Pacific Interdecadal Climate Oscillation with Impacts on Salmon Production*. *Bull. Amer. Meteor. Soc.*, 78(6):1069–1080.

McDermott, P. L. and Wikle, C. K. (2019). Bayesian Recurrent Neural Network Models for Forecasting and Quantifying Uncertainty in Spatial-Temporal Data. *Entropy*, 21(2):184.

McPhaden, M. J. (2002). Mixed Layer Temperature Balance on Intraseasonal Timescales in the Equatorial Pacific Ocean. *J. Climate*, 15(18):2632–2647.

Meinen, C. S. and McPhaden, M. J. (2000). Observations of Warm Water Volume Changes in the Equatorial Pacific and Their Relationship to El Niño and La Niña. *J. Climate*, 13(20):3551–3559.

Neelin, J. D. and Dijkstra, H. A. (1995). Ocean-Atmosphere Interaction and the Tropical Climatology. Part I: The Dangers of Flux Correction. *J. Climate*, 8(5):1325–1342.

Newman, M., Alexander, M. A., Ault, T. R., Cobb, K. M., Deser, C., Di Lorenzo, E., Mantua, N. J., Miller, A. J., Minobe, S., Nakamura, H., Schneider, N., Vimont, D. J., Phillips, A. S., Scott, J. D., and Smith, C. A. (2016). The Pacific Decadal Oscillation, Revisited. *J. Climate*, 29(12):4399–4427.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

Nooteboom, P. D., Feng, Q. Y., López, C., Hernández-García, E., and Dijkstra, H. A. (2018). Using Network Theory and Machine Learning to predict El Ni\˜no. *Earth System Dynamics*, 9(3):969–983. arXiv: 1803.10076.

Radebach, A., Donner, R. V., Runge, J., Donges, J. F., and Kurths, J. (2013). Disentangling different types of El Niño episodes by evolving climate network analysis. *Physical Review E*, 88(5).

Ren, H.-L. and Jin, F.-F. (2013). Recharge Oscillator Mechanisms in Two Types of ENSO. *J. Climate*, 26(17):6506–6523.

Rodríguez-Méndez, V., Eguíluz, V. M., Hernández-García, E., and Ramasco, J. J. (2016). Percolation-based precursors of transitions in extended systems. *Scientific Reports*, 6:29552.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.-T., Chuang, H.-y., Iredell, M., Ek, M., Meng, J., Yang, R., Mendez, M. P., van den Dool, H., Zhang, Q., Wang, W., Chen, M., and Becker, E. (2013). The NCEP Climate Forecast System Version 2. *J. Climate*, 27(6):2185–2208.

Saji, N. H., Goswami, B. N., Vinayachandran, P. N., and Yamagata, T. (1999). A dipole mode in the tropical Indian Ocean. *Nature*, 401(6751):360–363.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Suarez, M. J. and Schopf, P. S. (1988). A Delayed Action Oscillator for ENSO. *J. Atmos. Sci.*, 45(21):3283–3287.

Tang, Y. and Hsieh, W. W. (2003). Nonlinear modes of decadal and interannual variability of the subsurface thermal structure in the Pacific Ocean. *Journal of Geophysical Research: Oceans*, 108(C3).

Tangang, F. T., Hsieh, W. W., and Tang, B. (1997). Forecasting the equatorial Pacific sea surface temperatures by neural network models. *Climate Dynamics*, 13(2):135–147.

Tangang, F. T., Hsieh, W. W., and Tang, B. (1998a). Forecasting regional sea surface temperatures in the tropical Pacific by neural network models, with wind stress and sea level pressure as predictors. *Journal of Geophysical Research: Oceans*, 103(C4):7511–7522.

Tangang, F. T., Tang, B., Monahan, A. H., and Hsieh, W. W. (1998b). Forecasting ENSO Events: A Neural Network–Extended EOF Approach. *J. Climate*, 11(1):29–41.

Trenberth, K. and National Center for Atmospheric Research Staff (Eds) (2019). The Climate Data Guide: Nino SST Indices (Nino 1+2, 3, 3.4, 4; ONI and TNI). `https://climatedataguide.ucar.edu/climate-data/nino-sst-indices-nino-12-3-34-4-oni-and-tni`.

Tsonis, A. A. and Swanson, K. L. (2008). Topology and Predictability of El Niño and La Niña Networks. *Physical Review Letters*, 100(22).

Wheeler, M. C. and Nguyen, H. (2015). TROPICAL METEOROLOGY AND CLIMATE | Equatorial Waves. In North, G. R., Pyle, J., and Zhang, F., editors, *Encyclopedia of Atmospheric Sciences (Second Edition)*, pages 102–112. Academic Press, Oxford.

Wiedermann, M., Radebach, A., Donges, J. F., Kurths, J., and Donner, R. V. (2016). A climate network-based index to discriminate different types of El Niño and La Niña. *Geophysical Research Letters*, 43(13):7176–7185.

Wieners, C. E., de Ruijter, W. P. M., Ridderinkhof, W., von der Heydt, A. S., and Dijkstra, H. A. (2016). Coherent Tropical Indo-Pacific Interannual Climate Variability. *J. Climate*, 29(11):4269–4291.

Wu, A., Hsieh, W. W., and Tang, B. (2006). Neural network forecasts of the tropical Pacific sea surface temperatures. *Neural Networks*, 19(2):145–154.

Zebiak, S. E. (1986). Atmospheric Convergence Feedback in a Simple Model for El Niño. *Mon. Wea. Rev.*, 114(7):1263–1271.

Zebiak, S. E. (1990). Diagnostic Studies of Pacific Surface Winds. *J. Climate*, 3(9):1016–1031.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 301–320.