Predicting the forecast skill for the European region with the use of machine learning models

Master's thesis in Meteorology
by

Fabian Mockert

April 2022

INSTITUT FÜR METEOROLOGIE UND KLIMAFORSCHUNG
KARLSRUHER INSTITUT FÜR TECHNOLOGIE (KIT)

Referent:	Jun.-Prof. Dr. Christian Grams
Co-referent:	Prof. Dr. Peter Knippertz

Abstract

Users of Numerical Weather Prediction (NWP) services demand reliable weather forecasts, from short-range to seasonal lead times. Weather forecasts, such as from the European Centre for Medium-Range Weather Forecasts (ECMWF), have constantly improved. The forecast skill horizon has increased by one day per decade. Therefore, the 6 day forecast in 2010 was as good as the 3 day forecast in 1980. The predictability of the atmosphere is physically limited, which we call intrinsic predictability limit. This is due to the chaotic nature of the atmosphere. Small-scale uncertainties in the initial state may grow rapidly in time. The practical predictability limit, the time-scale at which actual forecast models still provide reliable forecasts, is far from reaching the intrinsic predictability limit. Forecast models are imperfect as they can not resolve all physical processes and use parameterisations. The intrinsic and practical predictability depend on the large-scale atmospheric state which can be described by different climate modes. Preferred climate modes are the Madden-Julian oscillation (MJO), stratospheric polar vortex (SPV) and North Atlantic oscillation (NAO), and their relation to each other, also called teleconnections. Thus, the overarching hypothesis of this thesis is that knowledge about the state of these climate modes at forecast initialisation time can help to predict the practical predictability of the atmosphere. Next to the atmospheric state, the forecast system itself can provide information about the reliability of its forecast, using the dispersion of individual forecast members in an ensemble forecast, hereafter referred to as ensemble spread. Our aim is to use the climate modes and the ensemble spread, which are known a priori, to predict the forecast skill of the ECMWF forecast for the European region. Our method involves a set of three machine learning (ML) models with different network structures, a fully connected neural network (FCNN), long short-term memory network (LSTM) and convolutional neural network (CNN).

Applying the three trained ML models on the same period, the extended winter (November–March) from 2013 to 2016, for lead times up to 15 days, we find that the overall performance for all ML models is equally good. In comparison to two non-machine learning reference models, based on the climatology and the ensemble spread, our ML models are performing as good or better than the reference models. Especially at lead times of 4 to 11 days, the ML models are performing significantly better than the climatology. Using only the most confident predictions of our ML models significantly increases their skill. This effect is not visible for the climatolog-

ical reference model and therefore the gap between the performance of our ML models and the climatology is increasing. Comparing the performance of the ML models for the confident predictions, we find that the models with a more complex architecture are also more skilful than the model with the simplest architecture, especially on forecast lead times of 8 to 11 days. Analysing the importance of each predictor, we identify one predictor, the ensemble spread of the 500 hPa geopotential height field, as the most important predictor for all three models, on all lead times up to 15 days.
Good forecasts of the NWP model are characterised by a transition from a ridging to a zonal flow in the European region. Conversely, a transition from a zonal flow to a ridge for the bad forecasts. The CNN heavily relies on the ensemble spread and in some cases is mislead by it, causing wrong predictions. In these scenarios, additional information about the development of the 500 hPa geopotential height field could improve the performance of the CNN.

Previous studies have not included information given by the forecast system itself into their predictors. We show that this information can improve the performance of the ML models. Furthermore, due to the analysis of the 500 hPa geopotential height field at initialisation and lead time, we propose new ideas how to improve the performance of the ML models using a priori knowledge.

Zusammenfassung

Nutzer von numerischen Wettervorhersagediensten verlangen zuverlässige Wettervorhersagen, für kurzfristige bis saisonale Vorhersagezeiten. Wettervorhersagen, wie die vom Europäischen Zentrum für mittelfristige Wettervorhersage (EZMW), werden konstant besser. Die Vorhersagezeit verlässlicher Vorhersagen verlängert sich um einen Tag pro Dekade. Somit war die 6-Tage-Vorhersage im Jahr 2010 so gut wie die 3-Tage-Vorhersage im Jahr 1980. Die Vorhersagbarkeit der Atmosphäre ist physikalisch limitiert. Dieses Limit nennen wir intrinsische Vorhersagbarkeit, es entsteht aufgrund der chaotischen Natur der Atmosphäre. Kleinskalige Unsicherheiten im Ausgangszustand der Atmosphäre können im Laufe der Zeit schnell anwachsen. Das Limit der praktischen Vorhersagbarkeit, also das Limit für welches tatsächliche Vorhersagemodelle eine verlässliche Vorhersage bereitstellen, ist noch weit von dem intrinsischen Vorhersagelimit entfernt. Vorhersagemodelle sind nicht perfekt, da sie zum Beispiel nicht alle physikalischen Prozesse auflösen und Parameterisierungen benutzen. Die intrinsische und praktische Vorhersagbarkeit hängt von dem großskaligen Atmosphärenzustand ab. Der globale Zustand der Atmosphäre kann mithilfe von Klimamoden beschrieben werden. Wichtige Klimamoden sind zum Beispiel die Madden-Julian Oszillation (MJO), der stratosphärische Polarwirbel (SPV) oder die Nordatlantische Oszillation (NAO), und deren Verbindung miteinander, auch Telekonnektion genannt. Deshalb ist die Grundhypothese dieser Arbeit, dass Wissen über den Zustand dieser Klimamoden zum Initialisierungszeitpunkt der Wettervorhersage nützlich für die Vorhersage der praktischen Vorhersagbarkeit der Atmosphäre ist. Neben dem Zustand der Atmosphäre kann auch das Vorhersagesystem selbst Informationen über die Zuverlässigkeit der Wettervorhersage geben, indem die Streuung von einzelnen Vorhersagen der Ensemble Vorhersage betrachtet wird. Das Ziel der Masterarbeit besteht darin, mithilfe der Klimamoden und der Streuung des Ensembles, welche vor dem Eintreten der Vorhersage bekannt sind, die Vorhersagegüte der EZMW Wettervorhersage für den europäischen Raum vorherzusagen. Unsere Methoden beinhalten einen Satz von drei verschiedenen maschinellen Lernmodellen (ML-Modellen) mit unterschiedlichen Netzwerkstrukturen, ein vollständig verbundenes neuronales Netzwerk (im Englischen: fully connected neural network, FCNN), ein Netzwerk mit einem langen Kurzzeitgedächtnis (im Englischen: long short-term memory, LSTM) und einem gefalteten neuronalen Netzwerk (im Englischen: convolutional neural network, CNN).

Die ML-Modelle werden auf dem selben Testzeitraum, erweiterer Winter (November-März) von 2013 bis 2016, für Vorhersagezeiten von bis zu 15 Tagen angewandt. Die allgemeine Performance ist für alle ML-Modelle gleich gut. Im Vergleich zu zwei Referenzmodellen ohne maschinellem Lernen, das eine benutzt die Klimatologie und das andere die Streuung der Ensemble Vorhersage, ist die Perfomance unserer ML-Modelle für alle Vorhersagezeiten vergleichbar oder besser. Besonders bei Vorhersagezeiten von 4 bis 11 Tagen sind die Ergebnisse der ML-Modelle signifikant

besser als die Klimatologie. Wenn nur die zuverlässigsten Vorhersagen der ML-Modelle betrachtet werden, dann steigt die Vorhersagegenauigkeit der Modelle signifikant an. Dieser Effekt ist für das klimatologische Referenzmodell nicht zu beobachten und dementsprechend wächst der Abstand der Performance zwischen den ML-Modellen und dem klimatologischen Referenzmodell an. Beim Vergleich der drei ML-Modelle untereinander wird ein Unterschied bei den zuverlässigen Vorhersagen festgestellt. Die beiden ML-Modelle mit einer komplexeren Netzwerkstruktur erzeugen qualifiziertere Vorhersagen als das Modell mit der einfachsten Struktur, besonders für den Vorhersagezeitraum von 8 bis 11 Tagen. Bei der Analyse der Wichtigkeit jedes einzelnen Prediktors wird festgestellt, dass ein Prediktor für alle ML-Modelle am entscheidensten ist, die Streuung der Ensemble Vorhersage für die 500 hPa geopotentielle Höhe. Dies bezieht sich auf den Vorhersagezeitraum von bis zu 15 Tagen, in welchem die ML-Modelle eine bessere Performance als die Klimatologie aufweisen.
Gute numerische Wettervorhersagen zeichnen sich durch einen Übergang von einem Höhenrücken zu einer zonalen Strömung im europäischen Raum aus. Umgekehrt dazu zeichnen sich schlechte Vorhersagen durch einen Übergang von einer zonalen Strömung zu einem Höhenrücken aus. Das CNN ist stark abhängig von der Streuung des Ensembles und wird in manchen Fällen von dieser in die Irre geführt. Dies verursacht falsche Vorhersagen des CNN. In diesen Szenarien könnten zusätzliche Informationen über die Veränderung des Geopotentialfeldes die Performance des CNN verbessern.

Vorherige Studien haben Informationen aus der Ensemble Vorhersage nicht als Prediktoren für das ML-Modell verwendet. Wir zeigen, dass diese Informationen die Performance des ML-Modells verbessern können. Desweiteren werden anhand der Analyse des Geopotentialfeldes zur Initialisierung der Vorhersage und dem vorhergesagten Zeitpunkt neue Ideen vorgeschlagen, welche die Performance der ML-Modelle verbessern können. Dabei wird beachtet, dass die Informationen vor Eintreffen der Vorhersage verfügbar sein müssen.

Contents

1 Introduction

For decision makers in many sectors, such as energy and water management, the public health and emergency services or the agricultural sector, reliable weather forecasts, especially on the medium-range and subseasonal time-scale, are of great importance (White et al., 2017). To illustrate the importance of accurate forecasts, we take a look at an example in the agricultural sector (similar to Rodwell et al. (2018)): A farmer wants to harvest his crop in a few weeks, but prior to the desired day of harvesting, weather forecasts predict a heavy rainfall event with a probability of 20%. The farmer needs to know, how reliable this forecast is, as such an event could destroy his crop. One way of retrieving the reliability of this forecast is to analyse all past events where heavy rainfall was predicted with a 20% probability and count in how many situations heavy rainfall actually occurred. If the percentage is higher than the predicted 20%, the farmer knows that the forecast is unreliable and can take action such as harvesting parts of his fields prior to the expected rainfall event to reduce the potential loss of all his harvest.
Users of Numerical Weather Prediction (NWP) services demand reliable forecasts beyond 14 days and desire the skill of these forecast to be as good as the skill of a five day weather forecast (Mariotti et al., 2020), but to date this is impossible. Buizza and Leutbecher (2015) explain that the forecast skill horizon for NWP with a 'deterministic' approach, based on a single numerical integration, lies at 10 days. By shifting to a probabilistic approach with ensembles of numerical integrations to estimate a probability distribution function of forecast states, they show that the forecast skill horizon can be extended beyond two weeks. Exact values depend on the spatial and temporal scales of the predicted phenomena, the variable used and the season looked at.
Forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) have improved by 15% for 3 and 5 day forecasts in the time from 1980-2010, and nearly by 20% for 7 day forecasts (Lillo and Parsons, 2017). To express this in another way, the forecast lead time improves by one day per decade, making the 6 day forecast in 2010 as good as the 3 day forecast in 1980.

Even though the skill of forecasts is constantly improving, there are physical limits of the predictability of the atmospheric state. Lorenz (1969) explains that non-periodic oscillations have a finite predictability and thus, even with an optimum procedure, there is an intrinsic predictability limit. Small perturbations in the initial condition, no matter how small they are, cause an error in the forecast which grows in size over time. Lorenz describes the theory of atmospheric instability with a flap of a sea gull's wing which changes the future course of the weather. The phenomena is later referenced as the „butterfly effect“ (Buizza and Leutbecher, 2015).
Though the forecast skill horizon has improved by one day per decade, there are still occasions when the day-6 high resolution forecast of European 500 hPa geopotential height (Z500) field has a strong drop in performance. Rodwell et al. (2013) call these scenarios „forecast busts“, which

are events during which the anomaly correlation coefficient (ACC) is less than 40% and the root mean squared error (RMSE) is larger than 60 m for the day-6 forecast. They state that in 1990, 70 busts occurred in the ECMWF model per year and that the number of busts has reduced to roughly five per year in 2011. Their work shows that explanations can be found in the flow regime, teleconnections and instabilities of the flow. Similar to the bad predictability during forecast busts on the medium-range, a low or high predictability on the subseasonal to seasonal (S2S) range is also dependent on the atmospheric flow configuration, with stable conditions increasing the predictability (Ferranti et al., 2015). There are many different climate modes, such as weather regimes, the Madden-Julian oscillation (MJO), the Quasi-Biennial oscillation (QBO) or the North Atlantic oscillation (NAO), that importantly modulate the forecast skill in the Euro-Atlantic region.
Favourable atmospheric conditions, that may lead to enhanced skill are called windows of opportunity (Mayer and Barnes, 2020a,b). To know which processes and circulation patterns increase predictability is relevant for the interpretation of forecasts (Mariotti et al., 2020). One approach to detect the windows of opportunity at the time of initialisation of the forecast is to use machine learning (ML) models such as Mayer and Barnes (2020a) have shown.

This Master's thesis addresses the overarching question, if there is „a way to know (a priori) which flow configurations lead to a more predictable state and therefore to a more accurate forecast?“ as questioned in Ferranti et al. (2015). For this thesis, a priori knowledge involves the state of climate modes across the globe, atmospheric field variables or the spread of ensemble forecasts at the lead time being considered. The research questions that will be answered in this Master's thesis are listed below:

1. **Is it possible to improve the prediction of the forecast skill using ML models?**
2. **Are predictions of the forecast skill improving with an increasingly complex architecture of ML models?**
3. **On which time-scale are ML models able to skilfully predict the forecast skill and which predictors are driving the decision making process?**
4. **When and why are ML models failing to predict the correct forecast skill?**

The Master's thesis is organised as follows: In Chapter 2, information about weather forecasting in general is given and known sources of predictability are introduced. The „Data and Methods“ part introduces data which are used in the thesis and explains the different methods that are applied, such as forecast skill scores and the architecture of the ML models. The results are presented in Chapters 4-6, starting with an evaluation of the performance of the ML models in Chapter 4. In Chapter 5, the meteorological situation for different forecast skill scenarios is analysed. Chapter 5 serves as a foundation for the next chapter, Chapter 6, where the meteorological situation for different forecast skill categories of the convolutional neural network are analysed and a possible application of the ML model in an operational sense is discussed. The final chapter, Chapter 7, concludes the results gained in this thesis and relates them to the literature. Further, the results are critically discussed and suggestions for future work are made.

2 Theoretical background

To ease the understanding of this Master's thesis, some meteorological terms and atmospheric relationships need to be explained. In this chapter, we build a meteorological foundation that helps to explain our basic hypothesis: A priori knowledge of climate modes and ensemble spread allows to predict the forecast skill. An essential part of this thesis is based on the ensemble forecast of the European Centre for Medium-Range Weather Forecasts (ECMWF). We first introduce the concept of numerical weather prediction models and ensemble forecasts. In a following step, we investigate the forecast skill of the ECMWF ensemble forecast and introduce potential sources of predictability for the European region. In a final step, we summarise results of other studies that have worked on predicting the forecast skill.

2.1 Deterministic numerical weather prediction

Bjerknes et al. (1904) introduce the idea that the prediction of the state of the atmosphere could be treated as an initial value problem of mathematical physics. They propose that weather in the future can be determined by knowledge of the observed weather and the usage of partial differential equations. These partial differential equations are summarised as prognostic equations and include the Navier-Stokes equations, the mass continuity equation, the first law of thermodynamics and the ideal gas law. The prognostic equations describe the temporal change of wind, pressure, density and temperature in the atmosphere. An analytical solution of these equations is due to their mathematical intractability not possible. It is possible to solve the equations numerically, using a spatial and temporal discretisation. Due to the discrete representation of the atmosphere in space and time, not all atmospheric processes are explicitly resolved. Unresolved processes of motion, processes that occur on a smaller spatial scale than the distance between two grid points or on a smaller temporal scale than one time step, enter the equations through source terms for mass, momentum and heat. Their interaction with resolved processes is taken into account by parameterisations in the equations (Bauer et al., 2015).

The predictability of the atmospheric state, using a numerical weather prediction (NWP) model, is limited by incorrect representations of physical processes and parameterisations in the forecast model and uncertainties in the initial conditions. The ability to predict the future state with NWP models is referred to as practical predictability. Limiting factors are the current capability of observations, data assimilation, modelling and computing. Improving the models to perfection does not increase the predictability to an infinite range. Intrinsic predictability describes the extent to which prediction is possible if an optimum model is used (Lorenz, 1969; Melhauser and Zhang, 2012; Bauer et al., 2015). Lorenz (1969) describes the variations of the atmosphere as a superposition of periodic and non-periodic oscillations. Periodic oscillations, such as the annual and diurnal variations, are infinitely predictable. Non-periodic oscillations have a finite predictability. Small

perturbations in the initial condition, no matter how small the errors are, change the forecast of a non-linear system. The error eventually becomes much larger than the initial error and can not be reduced below a certain limit no matter how good the forecast model is. Even when reducing the initial error, the error growth remains (as long as the error is not equal to zero). The finite limit of intrinsic predictability depends on the spatial and temporal scale of the motion considered and can be flow-dependent (Lorenz, 1996).
Zhang et al. (2019) state that the practical predictability limit, using deterministic forecasts, for mid-latitude instantaneous weather is around 10 days. It is estimated that a further reduction of the initial condition errors by a factor of 10 would increase the practical predictability limit of the current-generation, state-of-the-art NWP models by up to 5 days. Thus, their study concludes that current NWP models are still far from reaching the ultimate limit of predictability. Buizza and Leutbecher (2015) state that in the past 25 years (from 2015), there has been a shift in the approach for numerical weather prediction from a deterministic to a probabilistic one. Forecasts are no longer based on single numerical integrations but rather on ensembles of numerical integrations estimating the probability distribution function of forecast states. This advance in ensemble techniques, together with advances in simulations of relevant physical processes, improved data-assimilation methods, increased computational power and more accurate estimates of the initial condition make it possible to expand the forecast skill horizon beyond two weeks, which was thought to be the limit by Lorenz (1969).

2.2 Ensemble forecasts

The atmospheric state at initial condition of an NWP model, representing the real atmosphere, will always slightly differ from the real atmosphere. These slight differences, or errors, in the initial condition will grow for lead times sufficiently far into the future. Results of the NWP model are going to be different than the real atmosphere. Ensemble forecasts are a solution to account for the initial condition errors and the analytic intractability of sufficiently detailed stochastic dynamic equations (Wilks, 2011b). For ensemble forecasting, a finite amount of samples from the probabilistic distribution describing the uncertainty of the initial state of the atmosphere are selected (shown in Figure 2.1 at initial time). This selection of initial conditions is called the ensemble of initial conditions. Each member (circle) represents a plausible initial state of the atmosphere (ellipse), consistent with the uncertainties in observation and analysis. The NWP model is computed for each member of the ensemble of initial conditions separately (arrows). At initialisation time, all ensemble members are very similar to each other. The ensemble at a future time approximates how the initial probability distribution would have been transformed by the physical laws in the NWP model. The growth of the errors depends on the flow itself (Leutbecher and Palmer, 2007). Errors in the NWP model do not only arise from the uncertainties in the initial condition, but also from model deficiencies. Model deficiencies have various reasons such as model uncertainties due to a limited resolution or boundary condition uncertainties due to an insufficient detail of the sub-grid scale orography. To deal with these uncertainties, perturbations are continually inserted into the ensemble members throughout the execution of the forecasts (ECMWF, 2018).
The trajectories of the ensemble members can diverge significantly over time (see Figure 2.1 between the intermediate and final forecast lead time). At the final forecast lead time, multiple groups

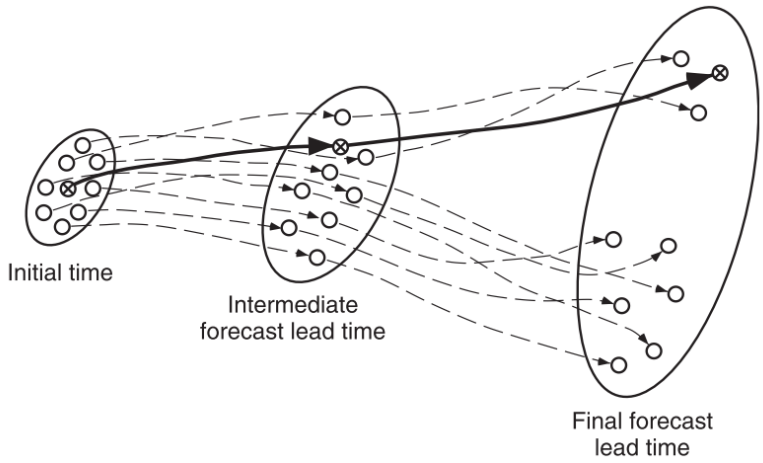


Figure 2.1: A schematic illustration of the ensemble forecasting in an idealised two-dimensional phase space. Each ensemble member is represented as a circle, with the single best initial condition including an X in the circle. The ellipses indicate the probabilistic distribution. The arrows indicate the trajectory of the ensemble members forecasting at different lead times. This schematic illustration is adapted by Wilks (2011b), Figure 7.24.

of similar forecasts can be obtained. The distribution of uncertainty is rather small at initialisation time and can increase substantially till the final lead time (see growth of the ellipses in Figure 2.1). The dispersion of the ensemble members is indicative of the uncertainty of the forecast, caused by initial analysis errors and model deficiencies. This information would not be available for a single forecast starting from the best initial condition (solid arrow).

Each ensemble member requires an own complete rerun of the dynamical model, which is computationally expensive. Operational forecast centres need to decide how many ensemble members they want to compute and which spatial resolution of the model they want to use. They also use different methods to choose the initial ensemble members. A detailed description of these methods is out of the scope of this Master's thesis. The ECMWF uses singular vectors, producing a control forecast and 50 perturbed forecasts members. Further information and cited articles are available in Wilks (2011b).

2.2.1 Forecast skill

It is important to quantify how well a forecast model describes the future state of the predicted variable. For this purpose we introduce the forecast skill.

Wilks (2011a) defines the forecast skill as the relative accuracy of a set of forecasts with respect to some set of standard reference forecasts. Reference forecasts can be for example climatological values of the predictand, a persistence forecast, a random forecast using the climatological distribution or the forecast with a different model. The forecast skill is represented as a skill score, which can generally be interpreted as a percentage improvement over the reference forecast. To do this, the agreement of the forecast with the observation must first be determined. In this Master's thesis we determine the agreement by applying the root mean squared error (RMSE). The RMSE makes a statement about the error between the forecast and observation. The error of the forecast model is then compared to the error of the reference model to determine the forecast skill. The RMSE and the skill score which is applied in this Master's thesis are explained in more detail in

Chapter 3.
When we speak of forecast skill in the following without specifying it further, we refer to the forecast skill for the Euro-Atlantic region.

2.2.2 Ensemble spread and spread-error relationship

The dispersion of the ensemble members can be used to indicate the uncertainty of the forecast and therefore the potential error between the forecast and observation. A larger (smaller) ensemble dispersion implies more (less) uncertainty in the forecast (Hopson, 2014). ECMWF (2020b) uses the standard deviation with respect to the ensemble mean as a measure for the ensemble dispersion, hereafter referred to as ensemble spread.

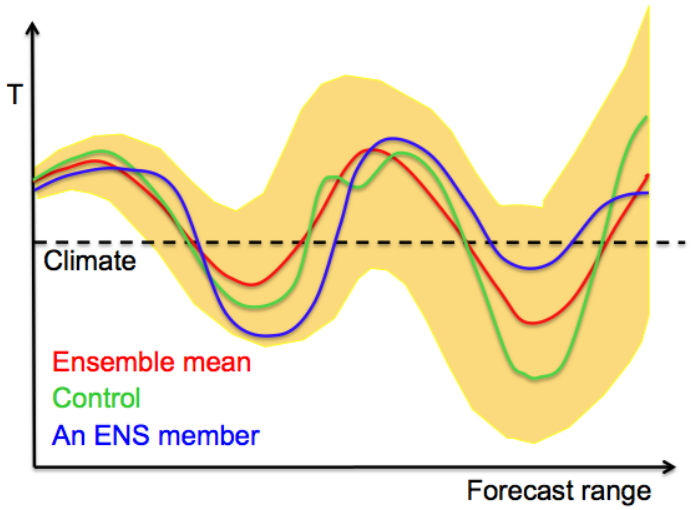


Figure 2.2: The schematic plume shows the relation between the standard deviation of the ensemble members (yellow shaded area), an individual ensemble member (blue line), the control member (green line) and the ensemble mean (red line). Figure adapted from ECMWF (2018), Figure 5.2.

The ensemble forecast consists of an unperturbed control forecast and individual perturbed forecast members. With the help of these, an ensemble mean and the ensemble spread can be determined (illustrated in Figure 2.2 on a range of forecast lead times). With an increasing forecast range, the ensemble spread usually also increases. The ensemble mean forecast is smoother than the individual members and tends to average out the less predictable atmospheric scales.
Rodwell et al. (2013) explain that in a perfectly calibrated ensemble the spread, as defined above, is an indicator of the distance of the truth from the ensemble mean, hence of the ensemble mean error. If averaged over many forecast start dates, the mean ensemble spread should match the mean ensemble error. The ensemble spread contains useful information about variations of the width of the distribution of the ensemble mean error. On a forecast range of 5-10 days, the average RMSE is fairly accurately predicted by the ensemble standard deviation (Leutbecher and Palmer, 2007). This relationship is known as the „spread-error relationship“. Ferranti et al. (2015) analyse the spread-error relationship for the 500 hPa geopotential height field (Z500) in Europe (Figure 2.3). They conclude that the ensemble of the ECMWF does indeed exhibit a good spread-error relationship for the European region. The good spread-error relationship suggests that the ensemble spread is a potential predictor of the forecast skill.

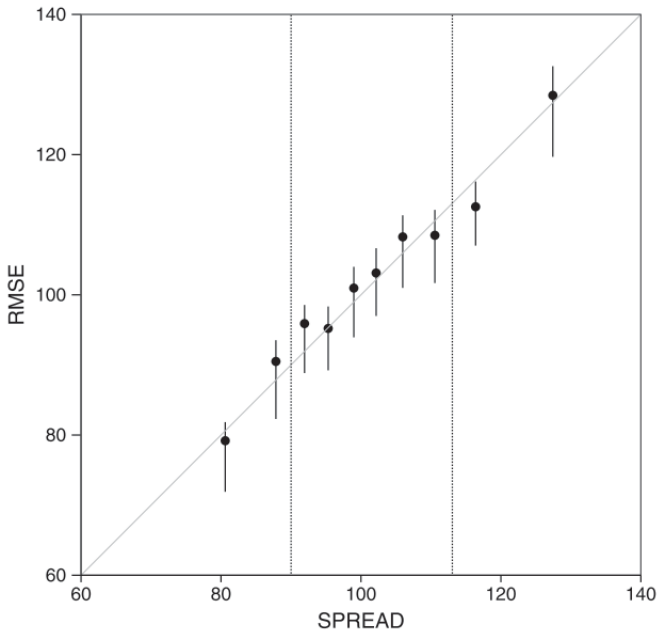


Figure 2.3: Scatterplot of RMSE versus the ensemble spread for day 10 forecasts for Europe (35.0 °N–75.0 °N, 12.5 °W-42.5 °E). Here the 500 hPa geopotential height field is analysed. The vertical dotted lines represent the upper and lower fifth values of the ensemble spread distribution. The ensemble spread distribution is binned into ten categories, each involving the same sample size. In order to remove noise, the RMSE is averaged over these bins. Sampling uncertainties, computed by re-sampling the data for each bin-average of the y-axis, are represented by a thin solid black line. As spread and error have the same physical dimension, a perfect ensemble forecast should produce points lying along the 45° line. The figure is adapted by Ferranti et al. (2015), Figure 8.

Whitaker and Louche (1998) note that statistical considerations suggest that the ensemble spread is likely to be most useful as a predictor of skill when it is „extreme“ in comparison to its climatological mean value, either very large or very small.

Weather regimes to estimate forecast skill

The atmospheric variability of the Euro-Atlantic region can be expressed by using weather regimes based on the 500 hPa geopotential height field. The forecast skill of weather regimes directly correlates with the forecast skill of the European region in general and therefore weather regimes can be used as an estimate of the forecast skill in this region.

Weather regimes describe quasi-stationary, persistent and continent-size circulation patterns located over the extratropical North Atlantic and European area (Vautard, 1990). There are several methods to classify the variability in weather in the Euro-Atlantic region (Grams et al., 2017a). One classification is the bimodal North-Atlantic Oscillation (NAO), with a positive and negative phase, resembling a cyclonic or blocked regime over Iceland and Greenland, respectively. A more detailed classification is possible with the definition of four or seven weather regimes. We briefly introduce the two weather regime classifications and describe the differences in the forecast skill of the various regimes.

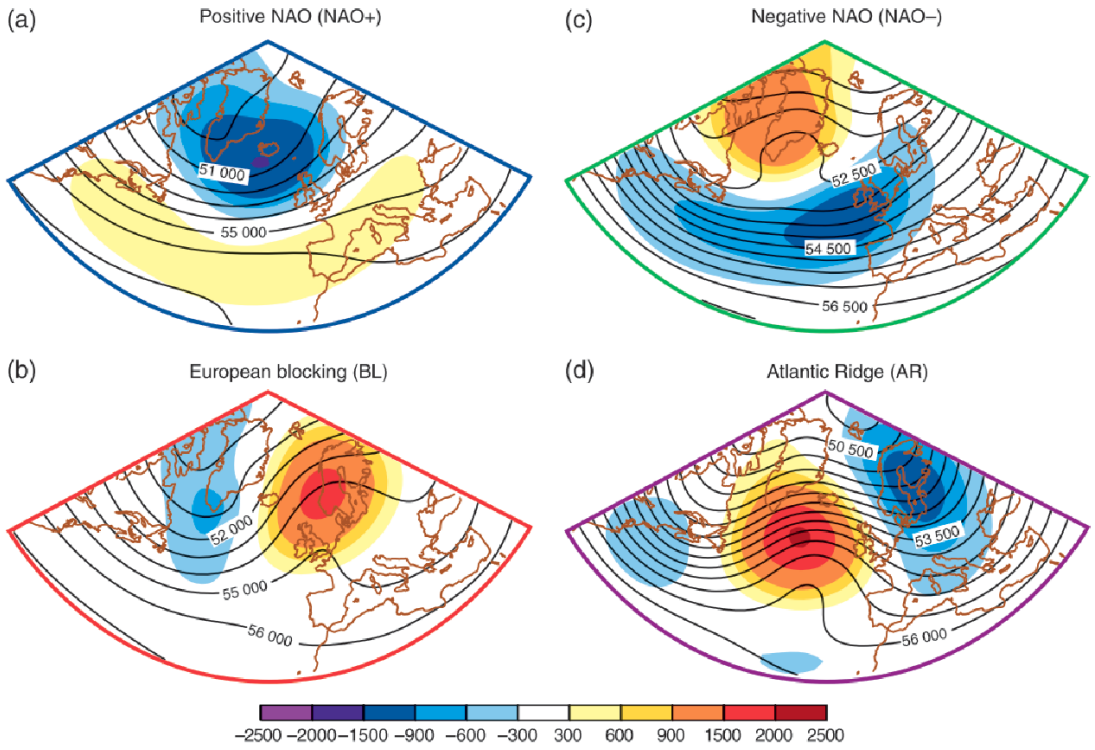


Figure 2.4: Geographical patterns of the four Euro-Atlantic climatological regimes (both anomalies and full fields) for the October to April cold season. The geopotential anomalies (colour shading) and geopotential (contours) at 500 hPa are shown for the positive NAO (a), European blocking (b), negative NAO (c) and Atlantic Ridge (d). Figure adapted by Ferranti et al. (2015), Figure 1.

Vautard (1990) and Ferranti et al. (2015) introduce the four weather regimes (Figure 2.4) as follows: The positive NAO regime (NAO+, Figure 2.4a) has a reversed dipole structure with a low centred over northern Europe and a parallel flow across the Atlantic. The negative NAO regime (NAO–, Figure 2.4c) has a strong positive anomaly over Greenland and a more zonally symmetric southern part (Vautard, 1990). The European Blocking (BL, Figure 2.4b) has a positive geopotential height anomaly centred over Scandinavia and a negative anomaly to the west over the Atlantic Ocean. The Atlantic Ridge (AR, Figure 2.4d) consists of a positive anomaly over the Atlantic Ocean and a negative anomaly over Scandinavia (Ferranti et al., 2015).

Forecasts initiated in the BL or AR regime show less skill than forecasts initiated in the NAO– or NAO+ regimes for day 9–13 forecasts (Figure 2.5, Ferranti et al. (2015)). Instability processes of the large-scale flow play an important role in the development of blocking anomalies and in the growth of errors during blocking transitions. Therefore, a better understanding of these instability processes is necessary to improve the forecasts initiated during a blocking regime (Ferranti et al., 2015). The highest probabilistic skill is observed during NAO– regimes. The skill is higher, if the NAO– regime persists longer. This increased predictability for a long weather regime is only observed for the NAO– regime and not for the three other regimes (Matsueda and Palmer, 2018)

The most frequent transitions between different weather regimes (for the four regime definition) are from the AR to the BL and from the BL to the NAO+. The preferred circuit of weather regimes starts with NAO+ transitioning to the AR, then to the BL and back to the NAO+ regime (Matsueda and Palmer, 2018). The least skilful forecasts are related to the missing of transitions to a blocking

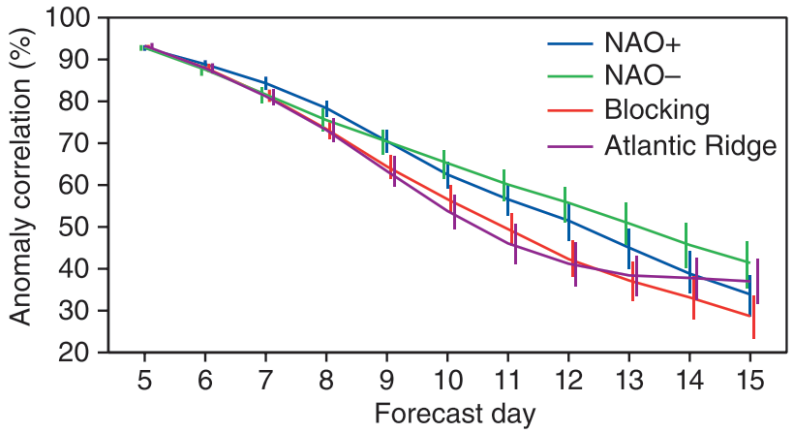


Figure 2.5: Anomaly Correlation of the ensemble means over Europe (35.0 °N–75.0 °N, 12.5 °W-42.5 °E) for the four weather regimes, NAO+ (blue), NAO– (green), European Blocking (red), Atlantic Ridge (purple). The bars indicate the 95% confidence intervals, generated with a bootstrap method using 1000 subsamples. Good forecast skill is represented by large values of the anomaly correlation. Figure adapted by Ferranti et al. (2015), Figure 3.

regime circulation and forecasts of BL initialised in the AR regime. Also, forecasts underestimate the blocking persistence, overestimate the persistence of zonal flows and over-represent transitions to the NAO+ (Ferranti et al., 2015).

Pasquier et al. (2019) mention that seasonal weather regime definitions like the four weather regimes are struggling in identifying robust patterns in the transition seasons and they suggest to solve this problem by introducing an extended seven weather regime definition which is valid for the variability in large-scale flow patterns in all seasons.
Grams et al. (2017a) explain that these seven regimes represent the winter and summer patterns and they mention that as one of the regimes (Greenland Blocking) is similar in all seasons, only seven and not eight regimes are found.
The seven weather regimes (shown in Figure A.1) are separated into cyclonic and blocking regimes. The cyclonic regimes (Atlantic Trough (AT), Zonal regime (ZO), Scandinavian Trough (ScTr)) have a predominant negative Z500 anomaly and are more frequent in winter months. The blocking regimes (Atlantic Ridge (AR), European Blocking (EuBL), Scandinavian Blocking (ScBL), Greenland Blocking (GL)) have a strong positive Z500 anomaly and are more frequent in the summer months. In Grams et al. (2017b), the seven weather regimes are described in more detail.
Büeler et al. (2021) analyse the forecast skill for the seven weather regime definition. As for the four weather regime definition (Ferranti et al., 2015; Matsueda and Palmer, 2018), they find that the blocking regimes, in specific the European and Scandinavian Blocking, have a decreased forecast skill, especially in winter time (December to February, DJF, Figure 2.6). All intervals that do not meet the conditions for one of the seven weather regime categories are designated as „no regime“. The forecast skill for the no regime in winter is lower than for the other seven weather regimes for most lead times up to 30 days (Figure 2.6). This highlights the difficulty in predicting phases lacking in persistence and do not fit into one of the distinct large-scale patterns. Periods of no regime could be identified as windows of low sub-seasonal predictability.

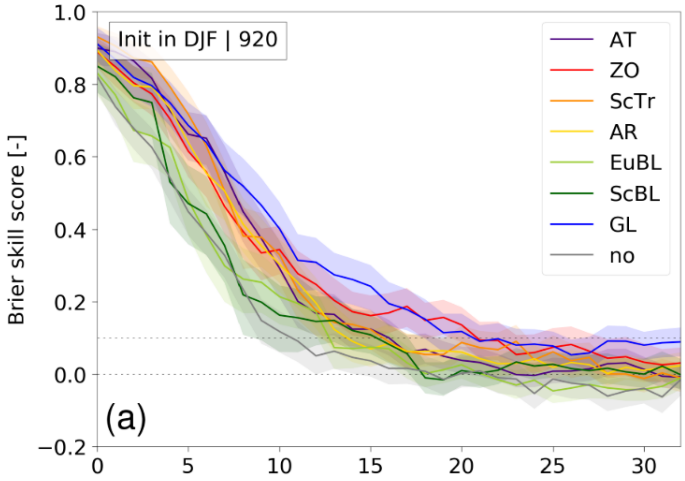


Figure 2.6: Winter (DJF) single-category Brier Skill Score for individual weather regimes (life cycle) as a function of lead time. The available number of forecasts in the winter season is indicated as 920 forecasts. The Brier Skill Score (BSS) has a negative orientation, values close to 1 indicate skilful forecasts. In the work of Büeler et al. (2021), the forecast skill horizon is defined as the forecast day where the BSS is below 0.1. The figure is adapted by Büeler et al. (2021), Figure 11a.

2.3 Sources of predictability

The forecast skill in the European region is influenced by large-scale atmospheric circulation patterns, which we refer to as climate modes. The interaction of geographically separated climate modes is called teleconnections (Holton, 2002). Climate modes and their teleconnections can be considered as sources of predictability. In this section, we introduce climate modes that influence the forecast skill in the European region and describe the relationship between these climate modes and the forecast skill.

2.3.1 Madden-Julian oscillation

The Madden-Julian oscillation (MJO) is the dominant intraseasonal mode of organised convective activity in the Tropics (Vitart, 2017), with a typical period of 20-90 days (Madden and Julian, 1971, 1972, 1994). Through excitation of quasi-stationary Rossby waves, the MJO is modulating the mid-latitude circulation days to weeks after the MJO activity (Mayer and Barnes, 2020a). Madden and Julian (1972) have described a typical life cycle of the MJO, it begins with a negative pressure anomaly over East Africa and the Indian Ocean, accompanied by increasing, large-scale convection over the Indian Ocean. Both, the pressure anomaly and the centre of the large-scale convection move eastward across Indonesia. Rising pressure over the Indian Ocean causes the convection to weaken and from the Western Hemisphere towards Africa, no enhanced convection is indicated, only a system of wind and pressure anomalies. The MJO can be split into active and inactive periods, and into eight different phases, describing the spacial position of the strongest convection. More details about the phases can be found in the „Data and Methods“ chapter (Chapter 3).

An active MJO has a positive impact on the forecast skill score for weekly forecasts (12-18, 19-25 and 26-32 days) for the European region, looking at the 500 hPa geopotential height, the 850 hPa

temperature and total precipitation (Vitart and Molteni, 2010). Vitart and Molteni (2010) mention that the impact is especially important for weekly forecasts at 19-25 days, as the models have close to no skill for inactive MJO periods for this specific lead time. The MJO can be used to decide if the monthly forecast at 19-25 days should be trusted or not. The forecast skill of predicting the MJO itself has increased significantly in the past years for the ECMWF model. These improvements are partly responsible for the increase of forecast skill of some teleconnections of the MJO (Vitart, 2014).

Baggett et al. (2017) mention that low-frequency variability, which is associated with phenomena such as the Quasi-Biennial oscillation (QBO) are likely to modulate the nature of MJO teleconnections.
Baldwin et al. (2001) describe the QBO as the dominating variability in the equatorial stratosphere (≈16–50 km). They explain that the oscillation can be seen as downward propagating easterly (EQBO) and westerly (WQBO) wind regimes and has a variable period of approximately 28 months. The onset of easterly and westerly wind regimes occur mainly during the Northern Hemisphere late spring at the 50 hPa level. Effects of the QBO influence the stratospheric flow, the variability in the mesosphere near 85 km, the chemical constituents due to circulation changes and the Earth's surface through the polar vortex.
Mayer and Barnes (2020b) have shown an enhanced impact of the MJO on the prediction skill, from the Pacific to Europe, during strong QBO periods. During EQBOs, the impact of the MJO on the extratropics is enhanced. At WQBOs, not only the impact is enhanced, but also the overall prediction skill compared to the neutral QBO (NQBO). They consider the daily mean Z500 forecasts from two operational forecast centres, the ECMWF and the National Centers for Environmental Prediction (NCEP). For both forecasts, they find significantly improved forecast skill during an active MJO phase with a westerly QBO phase, from the Pacific to Europe with lead times of up to 4 weeks.

2.3.2 Warm conveyor belts

Warm conveyor belts (WCB) are coherent warm and moist air streams, originating in the planetary boundary layer of an extratropical cyclone's warm sector. Air rises ahead of the cold front and across the warm front into a region of quasigeostrophic forcing, which results in precipitation below the WCB air stream. Condensation and therefore strong latent heat release contributes to the upward motion (Wandel et al., 2021). WCBs are contributing to upper-level ridge building and the formation of blocking anticyclones. The ridge formation in the North Atlantic due to the WCBs can result in uncertainties of forecasts in the European region (Wandel et al., 2021; Rodwell et al., 2018).

2.3.3 Rossby wave packets

Grazzini and Vitart (2015) describe Rossby wave packets as mid-latitude, atmospheric, synoptic-scale disturbances which propagate predominantly as downstream-developing waves. They travel in coherent wave packets, developing from baroclinic conversion downstream from a pre-existent disturbance and decay barotropically upstream, where wave breaking is also more frequently ob-

served. The reason for their formation can be manifold: Diabatic heating in the mid-lower troposphere due to pre-existing synoptic disturbances such as extratropical cyclones, bursts of organised tropical convective systems associated with MJO propagation, flow distortion from orography or recycling from previous waves in the jet-stream waveguide.
Grazzini and Vitart (2015) state that higher than average medium-range forecast skill in the European region is often associated with the presence of long-lasting Rossby wave packets from the west Pacific. Contrary, bad medium-range forecast skill is often associated with shorter Rossby wave packets originating from the United States of America or the western Atlantic.
The ECMWF model has difficulties in representing the convective environment over the central United States at the time of initialisation for forecast busts, occasions when the day-6 high resolution forecast of European 500 hPa geopotential height field has a strong drop in performance (Rodwell et al., 2013). These small perturbations first generate errors in small-scale moist convective systems and the errors grow upscale (Lillo and Parsons, 2017; Zhang et al., 2019) and propagate downstream along the jet stream (Rodwell et al., 2018), decreasing the forecast skill in the European region (Parsons et al., 2019).

2.3.4 Stratospheric polar vortex

An explanation of the stratospheric polar vortex (SPV) is given by Lee (2021). In winter time (end of August till April), the temperature difference between the poles and the tropics is stronger than during summer, resulting in a strong meridional temperature gradient, strongest in the stratosphere. Due to a balance between the vertical wind shear and the temperature gradient, the polar night jet stream encircles the cold air over the winter pole. Combining the polar night jet and the cold air which it encircles results in the cyclonic stratospheric polar vortex.
Sudden stratospheric warmings (SSWs) are described by Vitart (2014). The polar vortex with westerly winds in the winter hemisphere abruptly slows down or even reverses its direction, in the course of a few days. This is accompanied by the rise of stratospheric temperature by several tens of Kelvins. Sudden stratospheric warmings are considered as a potential source of state-dependent forecast skill in the Northern Hemisphere in winter (Vitart, 2014; Tripathi et al., 2015; Büeler et al., 2021).
In the ERA-Interim reanalysis, the probability of the NAO– regime increases after an SSW event. In the ECMWF forecasting system, the probability diminishes, the model under-represents the impact of the stratosphere on the troposphere (Vitart, 2014). Tripathi et al. (2015) show that both scenarios, anomalously weak and strong stratospheric polar vortices compared to climatology, enhance the predictability of the surface circulation. Büeler et al. (2021) present that strong states of the SPV are often followed by relatively persistent large-scale states, resembling the NAO+ regime, and anomalously weak states tend to be followed by the NAO– regime. The coupling between the stratosphere and troposphere enhances the sub-seasonal forecast skill for the NAO regimes (Baldwin et al., 2001; Tripathi et al., 2015).

2.4 Previous studies predicting the forecast skill

As Lorenz (1969) and other studies state, the predictability of atmospheric circulation has an intrinsic limit. Due to this limit, predicting the forecast skill will always be an issue, even if the average performance of numerical models increases (Tennekes et al., 1986).
Tennekes et al. (1986) summarise this issue by saying that „no forecast is complete without a forecast of forecast skill“. They mention that many forecasters believe that there is a relation between the predictability and the circulation type of the atmosphere.
Grönaas (1985) suggests that a subjective statement about the quality of the forecast in terms of an average, below average and above average quality, should be given.

Kalnay and Dalcher (1987) use the dispersion between members of an ensemble of forecasts from five different analyses to predict the forecast skill of an NWP model. When using regional verification fields in the Northern Hemisphere (North America, Europe, North Atlantic and North Pacific), they are able to provide a good prediction of the quality of the forecast given by an NWP model for lead times up to 5 days.

In the study of Palmer and Tibaldi (1988), they use different sets of predictors to predict the forecast skill of the 500 hPa geopotential height field in 12 regions in the Northern Hemisphere. One of their conclusions is that the RMSE of the day-1 forecast of the current forecast can provide information about the forecast error for later lead times.

Wobus and Kalnay (1995) predict the regional forecast skill using an ensemble of forecasts provided by four different operational centres. They apply a linear regression scheme on the previous 60 days to train the model. The correlation between the observation and their predicted skill for most mid-latitude regions and seasons is for 3–4 day forecasts at 0.4 to 0.7, using the anomaly correlation coefficient. Values above 0.4 are considered to be skilful. They suggest that the forecast skill improves if a larger ensemble of forecasts is used.

Albers and Newman (2019) use a linear inverse model (LIM) to identify, a priori, the expected extratropical subseasonal forecast skill for the mean sea level pressure and the 500 hPa geopotential height field. With the use of the signal-to-noise ratio of the LIM, they are able to identify a subset of higher forecast skill for the LIM model, but also for the forecasts given by the ECMWF and NCEP. On forecast lead times of 3–4 weeks (5–6 weeks), the LIM model identifies 20–30% (10%) of the forecasts with a relatively higher forecast skill compared to the remainder of the forecasts. Their a priori identification of usable subseasonal forecasts can be a guidance for users of NWP services.

Scher and Messori (2018) predict the weather forecast uncertainty for the European region, using a convolutional neural network. Their study has many similarities to this Master's thesis, but also differences such as the metric used as a predictand or the predictors. Their approach to predict the uncertainty of a weather forecast includes only information of the large-scale atmospheric state at initialisation time. Their goal is to find a method predicting the reliability of a forecast prior

to the ensemble forecast model run, thus information given by the ensemble forecast, such as the ensemble spread, is not used as a predictor field for the ML model. The focus of their study is on day-6 forecasts. Their approach using an ML model to predict forecast errors is not as skilful as the approach using the ensemble spread of the ensemble forecast. At many lead times, the ML model outperforms two other approaches which have been proposed in the literature, clustering by weather type and persistence in the phase space. Their method is not supposed to replace NWP models, but rather complement them, providing additional guidance for the confidence measure of ensemble forecast systems at initialisation time.
The evaluation of the ML model by Scher and Messori (2018) is performed with all predictions that the ML model offers. ML models can provide next to the prediction itself also a confidence of their prediction. The skill of the ML model by Scher and Messori (2018) could be increased if only the most confident predictions are selected for the evaluation. The research question by Mayer and Barnes (2020a) is different to the question asked in this Master's thesis or by Scher and Messori (2018), but they use the confidence of their ML model predictions to identify forecasts of opportunity, periods of atmospheric conditions that lead to an enhanced predictability. Their ML model is trained to predict the sign of the 500 hPa geopotential height anomaly in the North Atlantic (at 40 °N, 35 °W) at a lead time of 22 days, using the outgoing longwave radiation in the Indian-Pacific region (from 20 °S to 30 °N and 45 °E to 210 °E). The 10% most confident predictions are defined as forecasts of opportunity and with that selection, their accuracy increases from 56% for all predictions to an accuracy of 79% for the most confident predictions.

2.5 Wording: weather forecasting and neural network predictions

Studies investigating the forecast skill often use the terminology „forecasting the forecast skill". They do not differentiate between the terms „forecast" and „prediction", as for them there is no conflict with other scientific disciplines. Introducing machine learning and statistics to weather forecasts, this is no longer the scenario. In publications about machine learning, the word „prediction" is used to describe the outcome of the neural network, the prediction. In order to avoid confusion, we strictly separate these two terms and use the term „forecast" in context of weather forecasts and „prediction" in context of the outcome of the neural network. Terminologies such as „numerical weather prediction model" are not changed and also we want to state that „predictability" and „prediction" are used in different ways: Our neural networks predict the predictability for the large-scale weather situation in Europe. The handling of these two terms has been adopted from Scher and Messori (2018).

3 Data and methods

3.1 Data

3.1.1 ERA-5 reanalysis

Analysis data provides the best possible estimate of the current state of the Earth system. It is retrieved by combining the latest observations with a short-range weather forecast, constrained by previous observations. The process leading to analysis data is called Earth system data assimilation. If using the same data assimilation system for an extended period reaching back decades, the analysis is called reanalysis of past weather and climate.
Observations of the Earth system are unevenly distributed and come with errors. Reanalyses can fill the gap in the observational record and provide a data set which is spatially complete and consistent in time. Due to their consistency in time and space, they are helpful to understand climate change, current weather extremes and to evaluate the quality of forecasts (ECMWF, 2020a,c).

The ERA-5 reanalysis is produced by the ECMWF within the Copernicus Climate Change Service (C3S). ERA-5 covers the period of 1950 till the present. In 2016, it has replaced the previous reanalysis of ECMWF, ERA-Interim, which covers the period of 1979 till 2019. The ERA-5 reanalysis is based on the Integrated Forecasting System (IFS) Cy41r2 which uses a 4-dimensional variational analysis and was operational in 2016. It has a horizontal resolution of 31 km, 137 levels in the vertical up to 1 Pa and an hourly output. These and further details about the ERA-5 reanalysis are documented in Hersbach et al. (2020). In this thesis, we reduce the horizontal resolution to a grid spacing of 2.5° and use the following variables from the ERA5 reanalysis data: Geopotential height field at 500 hPa and 50 hPa, zonal wind at 850 hPa, 200 hPa and 100 hPa.

3.1.2 Subseasonal to Seasonal Prediction Project Database

Vitart (2017) gives a technical summary of the Subseasonal to Seasonal (S2S) Prediction Project Database and point out the importance and goals of this project. The main focus of the S2S database is to cover the gap between the medium-range forecast models and the seasonal forecast models to improve the forecast skill and enhance the understanding of processes on this time range. Medium-range forecast models are valid up to 15 days after initialisation and are considered as atmospheric initial-condition problems, the seasonal forecast models are ranging from three to six months after initialisation and depend on slowly evolving components of the Earth system, such as the sea surface temperature. The S2S gap is defined as the period from two weeks till two months after initialisation. The database is modelled on The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE) database for medium-range forecasts and Climate-System Historical Forecast Project (CHFP) for seasonal forecasts. In the S2S database, models by eleven different operational centres are involved. In this Master's

thesis, we are focusing on the model by the ECMWF and therefore describe their model in more detail.
The ECMWF model includes forecast lead times up to 46 days, with a horizontal resolution of $0.25° \times 0.25°$ (longitude$\times$latitude) for 0-10 days of lead time and $0.5° \times 0.5°$ for lead times larger than 10 days and a vertical resolution of 91 vertical levels. The real-time forecast ensemble has a size of 51 members and is initialised twice weekly. The reforecasts are generated on the fly, producing a reforecast set twice a week, starting at the same day and month as the next real-time forecast but for the past 20 years. Reforecasts are used to calibrate the real-time forecast. Reforecasts from the ECMWF include an ensemble of 11 members. The model involves ocean coupling but no sea ice coupling. The reforecasts of the 500 hPa geopotential height field are of interest for this work. For the generation of the root mean squared error, the full horizontal resolution of the field is used. In the generation process of the Z500 ensemble spread, the horizontal resolution is reduced to a grid spacing of 2.5°.

3.1.3 Predictors

The predictors for ML models, also called features, can be of any dimension in time and space. In this thesis, two sets of predictors are used. One set of predictors has no spacial extent, for each date there is one scalar value per feature. Features in this set are for example climate indices, describing the state of a climate mode. The other set of predictors has two dimensions in space for each time step. These features are mainly atmospheric field variables for a specific region on the globe. The different predictors and their sources are explained in the following.

Real-time multivariate MJO series

Wheeler and Hendon (2004), and later on Vitart and Molteni (2010), explain how information about the MJO phase and strength is retrieved via the real-time multivariate MJO index. The index is based on a pair of empirical orthogonal functions (EOFs) of the combined fields of near-equatorially (between 15 °N and 15 °S) averaged 850 hPa and 200 hPa zonal wind and satellite-observed outgoing long-wave radiation data. Projecting the forecast or analysis onto the EOFs, with the annual cycle and components of inter-annual variability removed, results in principal component time series, also called Real-time Multivariate MJO series 1 (RMM1) and 2 (RMM2).

The state of the MJO can be visualised in a two dimensional phase space using the RMM1 and RMM2 (Figure 3.1). The location of the enhanced convective signal can be identified by the position in the phase space, as it is divided into eight sections, each section representing a specific phase of the MJO. During phases 2 and 3 enhanced tropical convection associated with the MJO is located over the Indian Ocean. Conversely, phases 6 and 7 are characterised by enhanced tropical convection over the western North Pacific. The amplitude of a data point (distance to the centre of the phase space) represents the strength of the MJO activity. The MJO is considered to be active for an amplitude ($\sqrt{RMM1^2 + RMM2^2}$) larger than one (solid circle). Data is downloaded from the Australian Bureau of Meteorology (2022).

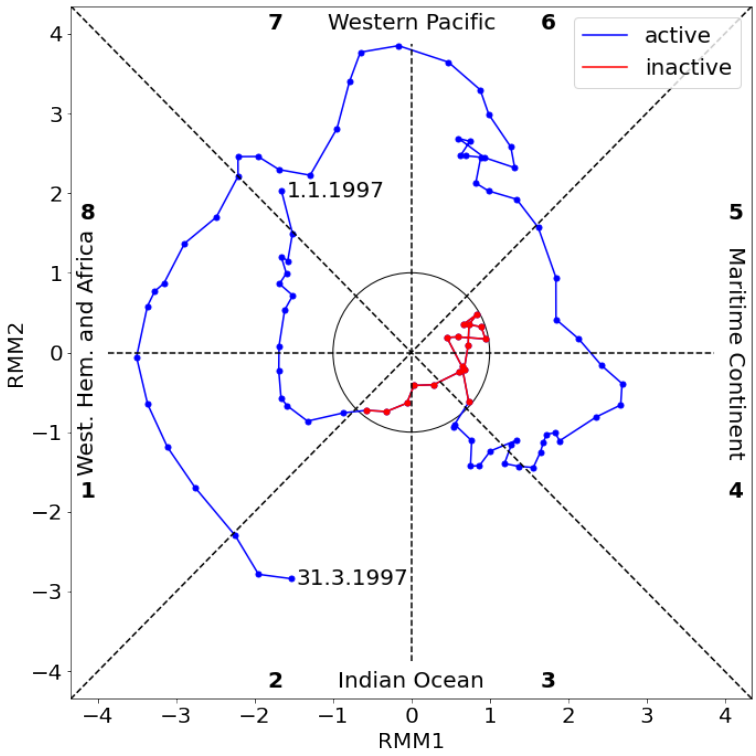


Figure 3.1: Phase space diagram of the real-time multivariate MJO index for the time from the 1 January 1997 till the 31 March 1997. There are two active MJO periods visible, from 1-14 January and from the 2 February till the 31 March (blue) and one inactive MJO period in between (red).

Quasi-Biennial oscillation index

A 3-month running averaged monthly zonal-mean zonal wind anomaly at 50 hPa over the tropics (10 °S-10 °N) defines the QBO index (Yoo and Son, 2016). The index differs between EQBO winter and WQBO winter, the former is reached when the standard deviation is above 0.5, the latter when it is less than -0.5, which corresponds to approximately $5\,\mathrm{m\,s^{-1}}$(Lim et al., 2019). Values in between are considered as neutral QBO (NQBO) events (Mayer and Barnes, 2020b). Data is computed from ERA-Interim and with the approach described by Yoo and Son (2016).

Ocean Niño index

There are different indices describing the El Niño-Southern oscillation. Trenberth and National Center for Atmospheric Research Staff (EDS) (2020) explain the Ocean Niño index (ONI), which is used as operational definition by the National Oceanic and Atmospheric Administration (NOAA). It is also used in this Master's thesis. The ONI is based on sea surface temperature (SST) anomalies (compared to the climatology between 1971-2000) averaged over the region 5 °N-5 °S and 170 °W-120 °W. A 3-month running mean is applied and an El Niño or La Niña event is classified if the anomalies are exceeding 0.5 °C or −0.5 °C, respectively, for at least five consecutive months. The 3-monthly running mean ONI is downloaded from NOAA (2022c) and linearly interpolated to daily values.

North Atlantic oscillation index

The dipole structure between the two quasi-permanent pressure systems, subtropical anticyclone near the Azores and subpolar low pressure system near Iceland, is leading to the construction of a two point NAO index (NAOI), defined as the standardised sea level pressure (SLP) difference between the stations of Ponta Delgada on the Azores and Reykjavik on Iceland (Wanner et al., 2001). Data is downloaded from NOAA (2022b).

Arctic Oscillation index

The AO is defined as the leading EOF of the monthly sea level pressure fields north of 20 °N. The AO index (AOI) is defined as the first principal component time series of the SLP fields north of 20 °N (Wanner et al., 2001). Data is retrieved from NOAA (2022a).

Pacific-North American pattern index

The Pacific-North American pattern (PNA) index is generated by projecting the daily 500 hPa geopotential height anomalies over the Northern Hemisphere (0-90 °N) onto the second leading mode of a rotated empirical orthogonal function analysis of monthly mean 500 hPa geopotential heights during the period of 1950-2000 (NOAA, 2021). Data is downloaded from NOAA (2021).

Stratospheric polar vortex index

In literature, the strength of the stratospheric polar vortex (SPV) is measured by the zonal-mean zonal winds at 10 hPa and 60 °N (Lee, 2021). In this thesis, we measure the influence of the SPV by the first principal component of the 100 hPa geopotential height field. Using the 6 hourly data by ERA-Interim, the daily mean for all values further north than 60 °N is calculated. To remove the seasonal cycle, the daily climatology is computed and removed. The data is weighted with the square root of the cosine of the latitudinal degree and afterwards an empirical orthogonal function (EOF) is applied. The product of the instantaneous field with the leading EOF field results in the first principal component (PC) which is used as indicator of the SPV state.

Warm conveyor belt metric

The warm conveyor belt (WCB) metric for the east and west Pacific (WCB_{east}: 20 °N-45 °N, 170 °E-220 °E; WCB_{west}: 20 °N-45 °N, 120 °E-170 °E) has been configured by Quinting and Grams (2021). The metrics are based on ERA-Interim and the 5 day mean of the average WCB frequency in the specific region is used for the metric. A more detailed description and the coding source are available at Quinting and Grams (2021).

Atmospheric field variables

Most atmospheric field variables for the convolutional neural network are retrieved from the ECMWF ERA-5 reanalysis data set which has been described at the beginning of this chapter. The variables include: Geopotential height field at 500 hPa (Z500) and 50 hPa (Z50), zonal wind

at 100 hPa (U100), 200 hPa (U200) and 850 hPa (U850). The outgoing long-wave radiation (OLR) data is retrieved from the website of NOAA (Liebmann and Smith, 1996).

Ensemble spread of Z500 reforecast

A further predictor for the different ML models is the ensemble spread of the 500 hPa geopotential height field given by the 11 different members of the ECMWF reforecast, provided by the S2S Prediction Project Database (Vitart et al., 2017). The ensemble spread is retrieved by calculating the standard deviation between the eleven ensemble members (ten perturbed forecasts and one control forecast) for each grid-point available. This two dimensional field of ensemble spread data serves as a predictor field for the CNN. The FCNN and LSTM network do not require a spatial extent of their predictors. For these networks, the mean of the ensemble spread values inside the European region (see Table 3.2 for the region boundaries) is computed.

3.1.4 Weather regimes

Grams et al. (2017a) describe an approach to identify seven weather regimes. The approach includes an empirical orthogonal function (EOF) analysis of 10 day low-pass filtered geopotential height anomalies at 500 hPa and a k-means clustering. The spatial domain is chosen from 30° to 90 °N, 80 °W to 40 °E. The used data set is ERA-Interim by the ECMWF. The weather regime index uses the projection of the instantaneous Z500 anomalous field to the cluster mean to derive individual weather regime life cycles. A weather regime life cycle needs to fulfil the following conditions: The index value needs to be larger than the standard deviation of the life cycle index computed over the whole available data set, the period where the life cycle index is larger than the standard deviation needs to be at least 5 days long and there needs to be a local maximum with a monotonic increase (decrease) of the index during the previous (following) 5 days.

3.2 Methods

3.2.1 Root mean squared error

The root mean squared error (RMSE) is a measure for the error between the forecast of the atmospheric field variable (here the Z500 field in the European region) and the observed atmospheric field.
Prior to the generation of the root mean squared error, the 500 hPa geopotential height field is weighted at every grid point by the cosine of the latitude:

$$\mathrm{Z500'}_{\mathrm{lat,\,lon}} = \mathrm{Z500}_{\mathrm{lat,\,lon}} \sqrt{\cos\left(\mathrm{lat}\frac{\pi}{180}\right)}, \tag{3.1}$$

with the latitude (lat) and longitude (lon) of the grid point expressed in degrees. This weighting accounts for the convergence of the meridians near the pole.
The mean squared error (MSE) is compiled by summing the square of the difference between the

forecast (y_m) and the observed values (z_m) at every grid point (m) and dividing the result by the number of grid points (M):

$$\mathrm{MSE} = \frac{1}{M} \sum_{m=1}^{M} (y_m - z_m)^2. \tag{3.2}$$

In comparison to the mean absolute error, which uses the absolute error instead of the squared error, the MSE is more sensitive to larger errors and therefore to outliers. To obtain the same physical dimension as for the forecasts and observations, the root mean squared error (RMSE) is introduced: $\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$. The RMSE has a negative orientation, which means that smaller values indicate a more accurate forecast (CAWCR, 2015). Vitart and Molteni (2010) define the skill of a model to be useful until the RMSE of the ensemble mean reaches the RMSE obtained by the climatology.

3.2.2 Categorised RMSE as predictand

In this work, the ML models are trained to make a categorical prediction of the forecast skill score RMSE. The RMSE is computed by comparing the 500 hPa geopotential height field in the European region (12.5 °W to 42.5 °E and 35 °N to 75 °N) of the ECMWF reforecast for a specific lead time as prediction and the ERA-5 reanalysis as observational truth. This process is done for all dates between 1997 to 2016 where a reforecast is available (3164 (816) reforecasts in the training (testing) period).

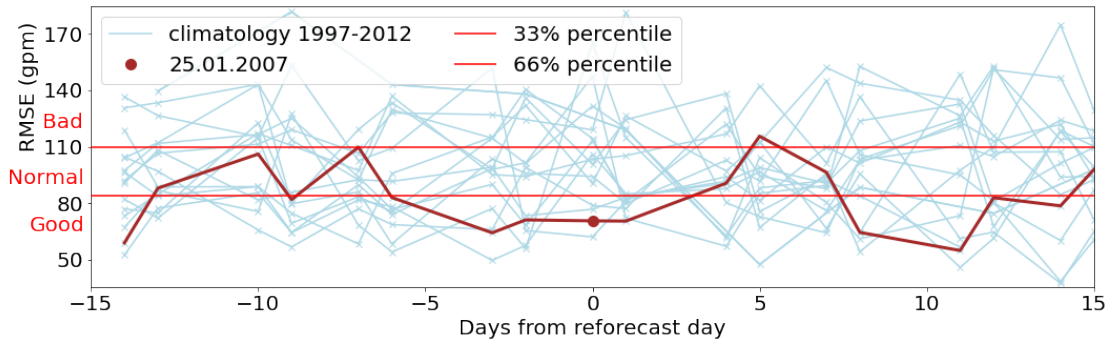


Figure 3.2: Visualisation of the generation of a categorised RMSE score. The RMSE values (blue crosses, values indicated on y-axis) for the 30 day running climatology around the reforecast date (0 on x-axis) are used to compute the forecast skill category. In this visualisation the RMSE values are computed for reforecasts in the 30 day running climatological window around the 25 January, with a lead time of 10 days. The 33% and 66% percentiles of the climatological RMSE values (red horizontal lines) serve as the boundaries for the three different forecast skill categories, $\mathrm{good}_{\mathrm{IFS}}$, $\mathrm{normal}_{\mathrm{IFS}}$ and $\mathrm{bad}_{\mathrm{IFS}}$. Comparing the RMSE value for the reforecast initialised on 25 January 2007 (red dot) with the category boundaries (red horizontal lines) shows that the RMSE for this specific reforecast is in the $\mathrm{good}_{\mathrm{IFS}}$ forecast skill category.

The generation process of a categorised RMSE score with three categories (visualised in Figure 3.2) includes the following steps: For each reforecast, the 30 day running climatological RMSE values in the training period (1997-2012) are collected (light blue crosses) and the 33% and 66% percentiles (red horizontal lines) are calculated. We explicitly exclude the testing period for the climatology as the ML models are not allowed to get information from the testing period. Including the testing period in the climatology indirectly transmits information of the RMSE in the testing

period to the ML models. The 33% and 66% percentiles of the climatological distribution serve as boundaries between the three categories. If the RMSE value of a single reforecast (red dot) is below (above) the 33% (66%) percentile, that reforecast is assigned to the „good$_{\text{IFS}}$“ („bad$_{\text{IFS}}$“) forecast skill category. RMSE values between the 33% and 66% percentiles are in the „normal$_{\text{IFS}}$“ forecast skill category. The subscript IFS stands for Integrated Forecast System, the underlying model ECMWF is using for their forecasts. In contrast to the IFS subscript, categories predicted by a ML model are either marked with „NN“ for Neural Network or more specific „FCNN“ for Fully Connected Neural Network, „LSTM“ for Long Short Term Memory or „CNN“ for Convolutional Neural Network.

3.2.3 Performance measures of the machine learning models

To quantify the quality of the predictions, different scores are applied. The accuracy and ranked probability score (RPS) consider the predicted and observed forecast skill categories. The ranked probability skill score (RPSS) compares the prediction of the ML model with a non-ML reference prediction. The equations are taken by the website of the Collaboration for Australian Weather and Climate Research (CAWCR, 2015).

Accuracy

The accuracy is the simplest score to calculate. It is the fraction between the number of correct predictions to the total number of predictions:

$$\text{Accuracy} = \frac{1}{N} \sum_{k=1}^{K} n(F_k, O_k), \tag{3.3}$$

with N as the total number of predictions, k iterating through the three categories ($K = 3$) good, normal and bad, and $n(F_k, O_k)$ the number of predictions where the prediction (F) and observation (O) are in the same category (k). The accuracy can range from 0 to 1 with a perfect score being 1.

Ranked probability score

The ranked probability score answers the question of how well the probabilistic prediction predicts the observed (here IFS) category.

$$\text{RPS} = \frac{1}{K-1} \sum_{k=1}^{K} \left[\left(\sum_{l=1}^{k} p_l \right) - \left(\sum_{l=1}^{k} d_l \right) \right]^2, \tag{3.4}$$

where K is the number of forecast categories, p_l is the predicted probability in the category l, and d_l is an indicator (0 = no, 1 = yes) for the observation in category l. The score ranges from 0 to 1 with 0 as a perfect score.

Table 3.1: Predictor variables for the three ML models: fully connected neural network (FCNN), long short-term memory (LSTM) and convolutional neural network (CNN).

FCNN	LSTM	CNN
Day of Year	Day of Year	Z500
RMM1	RMM1	Z50
RMM2	RMM2	U100
QBO	QBO	U200
ONI	ONI	U850
NAO	NAO	OLR
AO	AO	Z500 Ensemble Spread
PNA	PNA	
SPV	SPV	
WCB west	WCB west	
WCB east	WCB east	
Z500 Ensemble Spread	Z500 Ensemble Spread	

Ranked probability skill score

The ranked probability skill score (RPSS) measures the improvement of the multi-category probabilistic prediction relative to a reference prediction and is therefore strictly proper.

$$\text{RPSS} = \frac{\text{RPS} - \text{RPS}_{reference}}{0 - \text{RPS}_{reference}} = 1 - \frac{\text{RPS}}{\text{RPS}_{reference}}, \tag{3.5}$$

with the RPS as described above for the probabilistic prediction of the ML model and a reference prediction. Values can reach from $-\infty$ to 1. A score of 0 indicates no skill improvement compared to the reference prediction. For values smaller than 0, the reference prediction is performing better and for values between 0 and 1 the ML model prediction is performing better.

3.2.4 Neural networks

In this Master's thesis, three different architectures of ML models are used to predict the categorised RMSE. The models differ by the type of layers that are implemented and the predictor variables given to the network (Table 3.1). Next to these differences, the models share many details in their setup. In the following, the common details are explained and afterwards the specific setups for the different model architectures are explained.

ML models require a training and a test set which are independent of each other. At training time (1997-2012), we present the ML model the predictors and the categorised RMSE. For the training process, a k-fold cross validation method with four folds is applied. This means the training data is split into four equally sized segments and the ML model uses each segment ones as a validation set. The ML model with the best accuracy on the validation set is selected. To reduce variations in

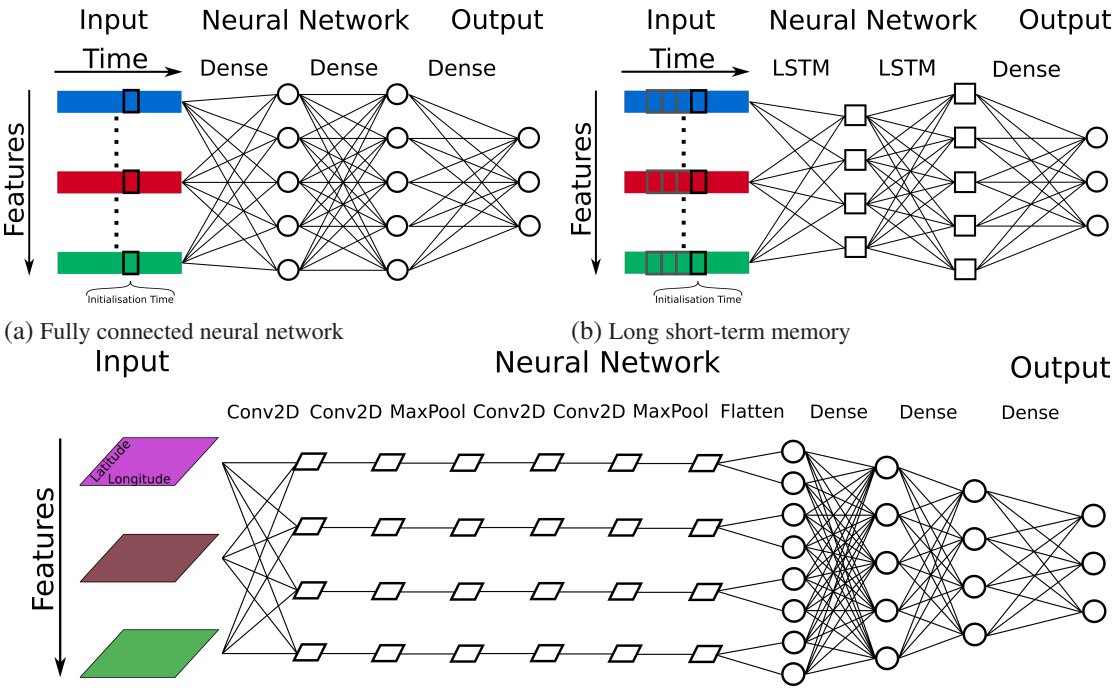
the results due to the randomly chosen initialisation parameters, this training process is repeated 10 times, each time with a different set of random initialisation parameters. To avoid ML models which do not properly learn on the training data, we introduce a minimum accuracy which the model needs to achieve on the training data. ML models which do not achieve an accuracy of at least 35% on the training data are discarded. The remaining ML models are used for making a probabilistic prediction of the categorised RMSE on the testing data set (2013-2016). The final probabilistic prediction is generated by computing the mean of the probabilistic predictions of all remaining ML models.

As in related studies, such as Scher and Messori (2018), the setup of each ML model architecture is determined on a trial-and-error basis. ML models are trained for a maximum of 60 epochs (except the LSTM model with a maximum of 100epochs). An early stopping function stops the training process if the validation loss is not improving for 15 consecutive epochs. The learning rate has an initial value of 0.001 and is reduced by a factor of 0.1 if the validation loss is not improving for 7 consecutive epochs.
The ML model structures are sequential models. All ML models use a dense layer, with three neurons and a softmax activation function, as the output layer (Figure 3.3 and in more detail for a lead time of 6 days in Figure A.2). The softmax activation function provides a probability distribution for each of the three neurons. These three neurons represent the three different forecast skill categories: good, normal and bad. The Adam optimiser is used with a sparse categorical cross-entropy loss function and the accuracy as a metric.
The hyperparameters, which are described in the next sections for each ML model individually, have been tuned separately for lead times of 5, 10 and 15 days. The best hyperparameter setup for the lead time of 5 days (15 days) is used for all ML models with a lead time up to 7 days (larger than 12 days). The best hyperparameter setup for the lead time of 10 days is used for ML models with a lead time larger than 7 but smaller than 12 days. Each input feature (b) is normalised using the „MinMaxScaler“ function by the „scikit-learn“ package. The minimum (maximum) value of the training set is set to 0 (1):

$$b' = \frac{b - \min(b_{\text{train}})}{\max(b_{\text{train}}) - \min(b_{\text{train}})}. \tag{3.6}$$

Fully connected neural network

The fully connected neural network (FCNN) is the simplest network used in the thesis. It uses a set of twelve scalar predictors (climate mode indices, WCB metrics, day of the year at initialisation time and the mean Z500 ensemble spread anomaly over the European region at the given lead time, Table 3.1). As these predictors have no spatial or temporal extend, the FCNN is also referred to as a zero dimensional ML model.
The FCNN consists of two dense layers with 128 neurons each and the ReLu activation function. The ReLu function transforms negative values into zero values: $e' = \max(0, e)$. The first dense layer is followed by a dropout layer with a dropout rate of 0.2 for lead times up to seven days and



(a) Fully connected neural network

(b) Long short-term memory

(c) Convolutional neural network

Figure 3.3: A simplified view of the three different ML models used in this Master's thesis: fully connected neural network (a), long short-term memory (b), convolutional neural network (c). The dropout layers are excluded and the number of nodes is reduced.

0.1 for lead times larger than seven days. The batch size is 64 for lead times up to 7 days and 32 for longer lead times. The dropout layer randomly sets 20% (for a dropout rate of 0.2) of the input values to 0. All other input values are scaled up by $\frac{1}{1-\text{rate}}$ that the sum over all inputs is unchanged. The dropout layer helps to prevent the neural network from overfitting (Keras, 2022).

Long short-term memory

The long short-term memory (LSTM) neural network uses the same set of predictors as the FCNN (Table 3.1), but instead of using only values at initialisation time, it also uses values prior to initialisation time. The input data of each feature consists of a time series with scalar values and therefore it is considered to be an one dimensional ML model.

The LSTM network also consists of two layers, but now using the LSTM layers given by Keras instead of the dense layers. The first LSTM layer has 32 (64, 128) nodes for a lead time of up to 5 (6-12, 13-15) days including a dropout layer with the factor 0.2 for lead times larger than 12 days. The second LSTM layer has 64 (64, 192) nodes for a lead time of up to 5 (6-12, 13-15) days, with no dropout layer but a recurrent dropout of 0.1 for lead times up to 7 days. The batch size is for all lead times 32. Whilst training, the length of the input time series has been varied between 1 and 14 days. The best evolution of the validation accuracy and the validation loss during training has been achieved when taking time series of 3 consecutive days up to initialisation time.

Table 3.2: Spatial extend of the different CNN input regions. A graphical visualisation can be found in Figure 3.4.

Region	Longitude min [°E]	Longitude max [°E]	Latitude min [°N]	Latitude max [°N]
Europe	−12.5	42.5	35	75
Euro-Atlantic	−90	42.5	30	75
Euro-Pacific	−165	60	0	80
Extended NH	−180	180	−30	90
Global	−180	180	−90	90

Convolutional neural network

For the convolutional neural network (CNN), a different set of predictors is used. The CNN in this work requires a two dimensional input field. Here, the two dimensions are the longitude and latitude and the features are either atmospheric field variables or the Z500 ensemble spread of the ECMWF reforecast for the specific lead time (Table 3.1). In total seven features are used as predictor fields. Similarly to the length of the time series in the LSTM model, the size of the input region is one parameter which needs to be fixed beforehand. Figure 3.4 shows the different regions that have been tested to determine the region where the CNN learns best. Regions covering a full latitude circle have been extended on both sides with additional 90° on each side to reduce the effect of a hard border at the dateline. The grid size has been reduced to a 2.5°×2.5° grid. A detailed overview of the extend of the different regions can be found in Table 3.2. The region with the best evolution of the validation accuracy and validation loss during the training process is the Euro-Atlantic region.

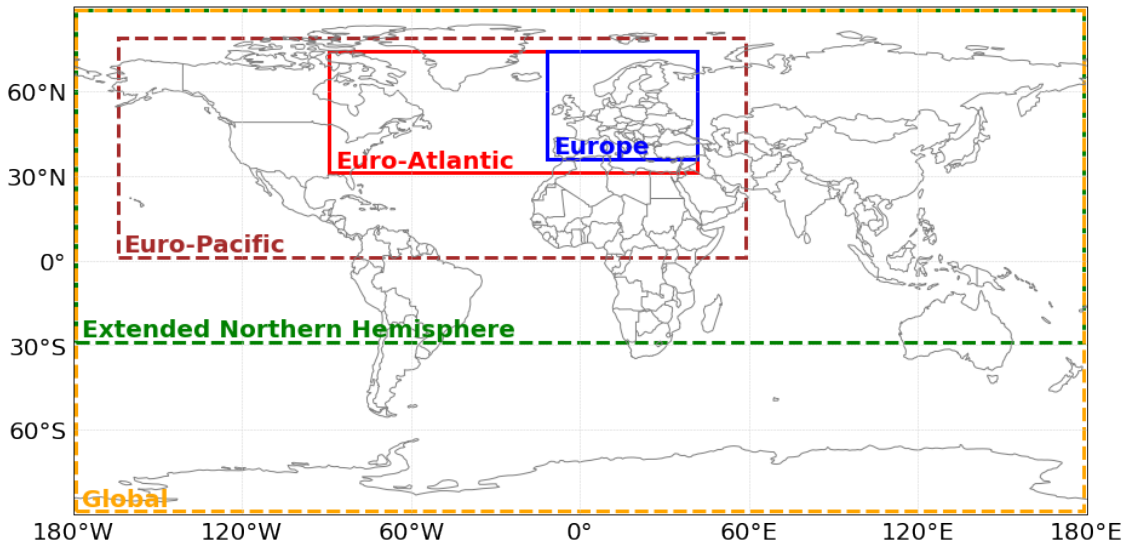


Figure 3.4: Different regions that are used for the predictor fields in the CNN. The CNN sufficiently learns for the European and Euro-Atlantic regions (solid blue and red lines). For the other regions (dashed lines), the CNN does not properly learn.

The CNN models for different lead times share the same network architecture. The CNNs are set up with two consecutive convolutional layers, followed by a max-pooling and dropout layer. This set of four layers is repeated once. After the second set of four layers, a flatten and two dense layers, separated by a third dropout layer are added to the CNN. All convolutional layers have a kernel size of 3×3, use the ReLu activation function and use the padding „same“. With the „same“ padding, the output shape of the convolutional layer equals the input shape. This is achieved by adding a frame of zero values around the two dimensional input data. The first (last) two convolutional layers are using 64 (32) filters. Both max-pooling layers have a pooling size of 2×2 and a stride of 2. The three dropout layers have a dropout rate of 0.3, 0.3 and 0.1, respectively. The two dense layers use the ReLu activation function and 512 (128) nodes for the first (second) layer.

3.2.5 Non-machine learning reference predictions

To calculate the RPSS, a probabilistic reference prediction is necessary. We use two different reference models. One reference model is based on the climatological distribution of the forecast skill categories and the other model is based on the Z500 ensemble spread. Both approaches do not include any machine learning techniques. We refer to them as the climatological reference model and the spread reference model.

Climatological reference model

Creating a fair climatological reference requires the usage of only the training data ranging from 1997 to 2012. For each day of the year, a 30 day moving window is applied on the training years. The fractions between the sum of each forecast skill category occurring in the moving window and the total amount of reforecasts in the moving window represent the probabilistic prediction of each category for the selected day of the year. The probabilistic prediction is transformed into a deterministic prediction by choosing the category with the highest probability, for each reforecast date separately.

Spread reference model

The generation of the spread reference prediction is similar to the generation of the RMSE categories in Section 3.2.2. For each lead time, the 33% and 66% percentiles of the 30 day moving window of the Z500 ensemble spread are calculated. The actual ensemble spread of the reforecast is then compared to the 33% and 66% percentiles of the 30 day running climatology (using only values from the training period). As known by the spread-error relationship, values in the upper (lower, middle) tercile are allocated to a bad (good, normal) forecast skill. This approach leads to a deterministic prediction of the different forecast skill categories. There is no information in the deterministic prediction that could help to create a probabilistic prediction. For the transformation to probabilistic values, fixed probability values are set for each category. A probability of 35% is assigned to the predicted category and categories that are not predicted obtain a confidence of 32.5% each. The fixed values were determined after testing different probability values for the predicted category from 100% down to 35%, comparing the RPS to each other.

3.2.6 Feature importance

Gaining an insight into the relevance of features for the confidence of an ML model can be achieved by using interpretable machine learning methods. One of these methods is the feature importance (König et al., 2020). Though different approaches of feature importance exist, we focus on the (relative) permutation importance. To determine how important one feature is for the trained ML model, the values of this feature are randomly shuffled. The random re-ordering of one feature should cause less accurate predictions as the shuffled data no longer corresponds to the observation (Kaggle, 2019). To evaluate the difference between the shuffled data set and the original data set, a performance metric, here the RPS, of both data sets is calculated and brought into relation:

$$\mathrm{RFI}_{feature} = \frac{\mathrm{RPS}_{\mathrm{feature}} - \mathrm{RPS}_{original}}{\mathrm{RPS}_{original}}, \tag{3.7}$$

with the subscript „feature“ indicating which feature has been shuffled.
Each feature is shuffled 50 times to reduce the effect of the random shuffling on the performance. The complete process is repeated for every feature in the data set. The higher the RFI values the more important is a certain feature for the ML model prediction. Comparing the RFI among different features can indicate the most important feature for the decision making process.

3.2.7 Class activation map

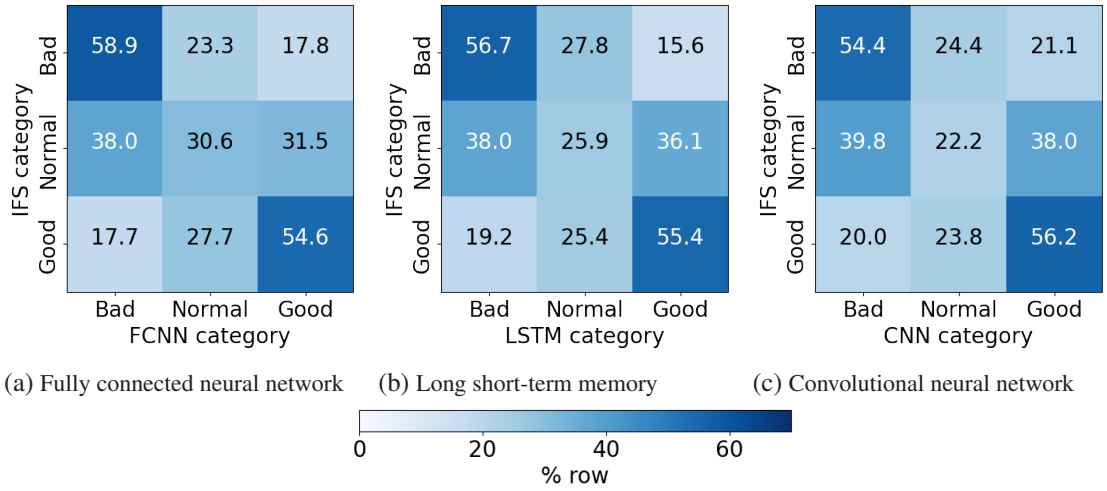
ML models are often seen as „black boxes“, but this is not necessarily true. For example, representations learned by CNNs can be visualised with a variety of techniques (Chollet, 2018). We use the concept of the gradient-weighted class activation mapping (Grad-CAM). This technique is useful to understand which parts of a given „image“ lead to its final classification decision. In this Master's thesis, we do not use classical RGB images, but each of the seven features used for the CNN can be understood as one colour of the image, resulting in an image consisting of seven different colours instead of three. Grad-CAM highlights regions in the two dimensional input data which are important to predict one of the forecast skill categories. To achieve this, it uses the gradient information which is flowing into the last convolutional layer and produces a coarse localisation map (Selvaraju et al., 2020). Overlaying the coarse localisation map with the features indicates areas of particular importance.

4 Evaluation of machine learning models

We use a set of three different types of ML models (see Chapter 3). The first step in analysing these models is to compare their performance, measured by their accuracy and RPSS, to each other and to the non-ML reference models. The distribution patterns (confusion matrices) between the predicted (ML model) and true (Integrated Forecast System) forecast skill categories are similar for different lead times. Therefore, representative for all lead times, only the models for day-6 reforecasts are analysed in full detail. After the detailed analysis of day 6 lead time, zooming out to lead times from 0 to 15 days visualises the performance of the ML models for the whole time range where they are skilful, compared to the non-ML reference models. At the end of the chapter, the importance of each feature/predictor to the model decision is investigated. Results from the feature importance specify the direction of the further analysis in Chapters 5 and 6.
When training the ML models on data throughout the full year, the models only learn seasonal effects, which is good forecast skill in summer time and rather bad forecast skill in winter time. This effect persists even after removing the seasonality of the forecast skill and after including the day of the year as a predictor. To avoid models only learning the seasonal effect, the data is restricted to the extended winter period (NDJFM, November to March). Training is performed for the period January 1997 to December 2012 and the testing data includes the extended winter periods from January 2013 to December 2016. In the testing data set, 328 reforecasts are available. The predictions of the ML models for this subset of reforecasts are evaluated in the following analysis.
The three observed categories for the test data are not equally distributed. 130 reforecasts (40%) belong to the good, 108 (33%) to the normal and 90 (27%) to the bad forecast skill category. The convolutional neural network produces skilful predictions with the European and Euro-Atlantic region as an input area. With the Euro-Atlantic region, the most skilful predictions are found and therefore, only the CNN model with the Euro-Atlantic region is further analysed.

4.1 Confusion matrices and performance measures

Confusion matrices (Figure 4.1) visualise which percentage of the forecast skill categories have been predicted correctly (diagonal) and incorrectly (off-diagonal) by a model. As the IFS forecast skill is not equally distributed on the three categories, we treat each IFS category separate. The percentage values in each IFS category (row) sum up to 100%. A perfect model would only include values of 100% on the diagonal from the top left to the bottom right. The values on this diagonal represent the percentage of correct predictions in this category. The weighted mean of this diagonal equals the accuracy measure.



(a) Fully connected neural network (b) Long short-term memory (c) Convolutional neural network

Figure 4.1: Confusion Matrices for the three different ML models, FCNN (a), LSTM (b) and CNN (c), and a lead time of 6 days. The matrices indicate for each IFS category (y-axis) the distribution of the predictions made by the ML model (x-axis). The percentage values in each IFS category (row) sum up to 100%. The shading visualises the percentage values of each category which are also given in numbers.

Table 4.1: Summary of the performance measures for all ML models (FCNN, LSTM, CNN) and the reference models (climatology, ensemble spread) at a lead time of 6 days.

Model	Accuracy	RPS	RPSSclim	RPSSspread
FCNN	47.9	0.201	0.133	0.087
LSTM	46.0	0.197	0.152	0.107
CNN	44.5	0.198	0.147	0.101
Climatology	36.3	0.232	0.000	−0.053
Spread	43.0	0.220	0.050	0.000

All ML models are predicting the good and bad category well (more than 54% of these forecast skill categories are predicted correctly, Figure 4.1). But also all models lack with the correct prediction of the normal forecast skill. In order to give a possible explanation for this behaviour, we need more knowledge about the importance of each individual predictor. Therefore, we provide the explanation at the end of this chapter in Section 4.6.

A large number of correct predictions is important, but also a small number of completely wrong predictions is desirable. The model should not predict the good category, if the bad category is actually occurring and vice versa (values in the left bottom and right top corner). For all of the three ML models, these two combinations range between 15–21% and thus have the lowest percentage out of all combinations.

To consider a ML model to be skilful, it should outperform the climatological reference model in both skill measures, the accuracy and the ranked probability skill score (RPSS). Furthermore it is desirable that the ML model performs better than the non-ML spread reference model, otherwise the use of a complex machine learning method is not justified.

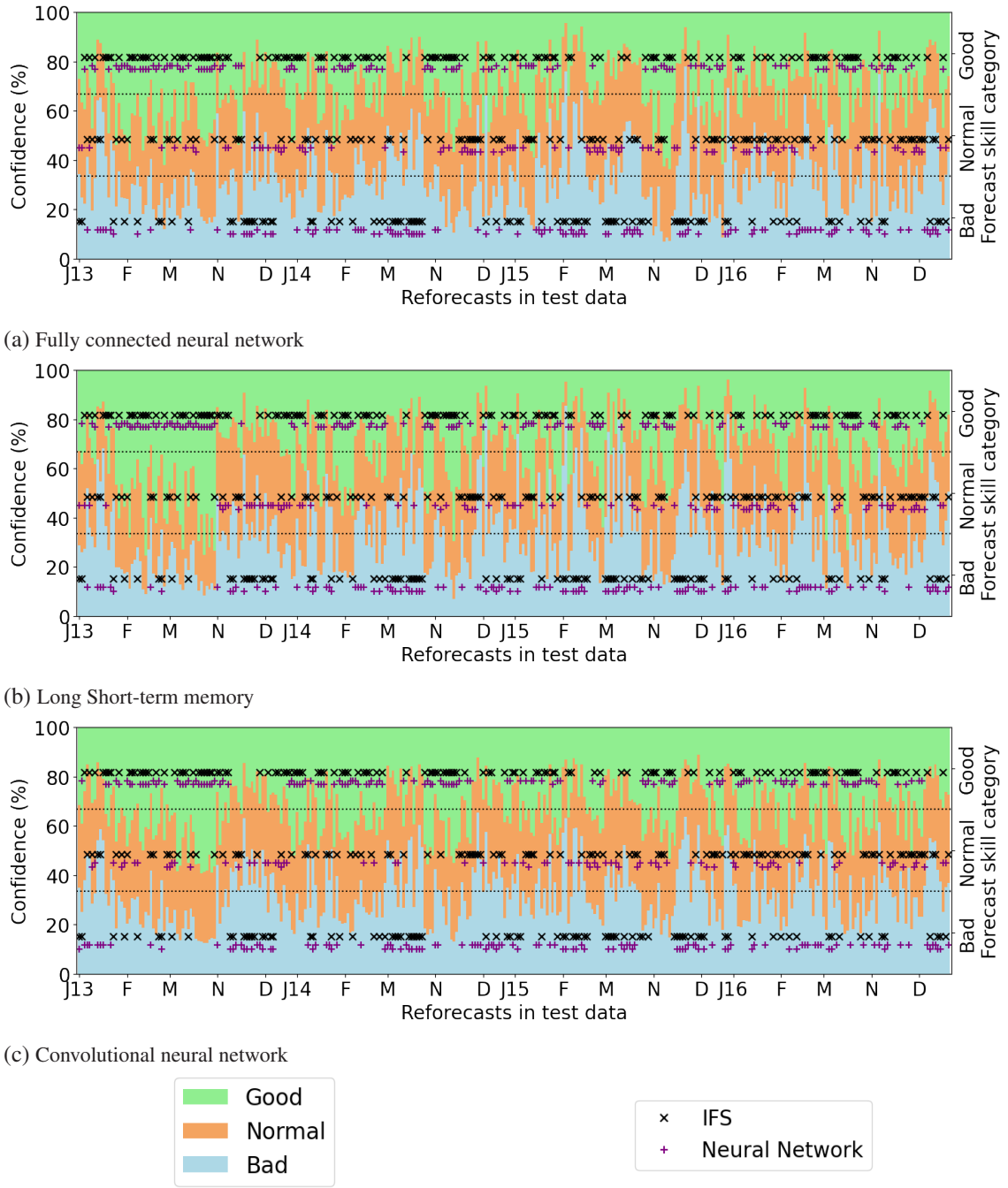
The spread reference model outperforms the climatological reference model, therefore it can be considered as skilful (Table 4.1).
There is no clear favourite among the ML models (Table 4.1) as all ML models perform similarly well. In comparison with the non-ML models, all ML models perform significantly better than the climatological reference model, in terms of all skill measures, and therefore they can be considered as skilful. The spread reference model has a similar accuracy, but the RPS and RPSS are worse than for the ML models.

These results show quantitatively that the ML models outperform simple non-ML approaches in predicting the forecast skill category. If using the predictions for all reforecasts available in the testing period, the ML models are performing equally well.

4.2 Confidence of the prediction

The given ML models provide a probabilistic prediction for each category. In the last section, we have transformed the probabilistic prediction to a deterministic prediction by selecting the category with the largest percentage. The percentage with which each category is predicted can also be interpreted as the confidence of the ML model.
The confidence of the prediction is an important indicator for the end user of a ML model to judge whether the prediction of a ML model is reliable or not. In many scenarios our ML models are confident in their prediction of the forecast skill category with a lead time of 6 days (Figure 4.2). 65–100 out of 328 reforecasts reach a confidence larger than 50% for one category.
Markers in Figure 4.2 indicate long-lasting periods during which the skill of subsequent forecasts falls into the same category. We refer to these periods as clusters. Each prediction of the ML models is independent from the previous prediction. Information about the clustering is not directly transmitted to the ML models, as this information is not available at initialisation time. Still, the models often generate these clusters, combined with a larger confidence for these predictions than on average. The formation of clusters in the IFS forecast skill categories shows that the categories persist for several days up to weeks which links back to the sources of predictability introduced in Chapter 2. The ML models are capable of detecting patterns in the predictor variables which lead to these persistent forecast categories.

The results of the study by Mayer and Barnes (2020a) suggest that with an increase in confidence in the prediction, there is also an increase in the skill of the prediction, namely in the accuracy and the RPSS. The increase in the ML model performance for more confident predictions is visible in Figure 4.2 as the predicted forecast skill categories for more confident predictions are more often correct (lowered purple plus signs in Figure 4.2).
This behaviour is quantitatively shown in Figure 4.3. The most confident predictions are selected out of all reforecast dates and the accuracy and $\text{RPSS}_{\text{clim}}$ are calculated for these predictions. The accuracy and $\text{RPSS}_{\text{clim}}$ for the spread reference model are not shown when using a selection of dates based on the confidence. This is due to the generation process of the probabilistic prediction for this model for which the dominant category always has a confidence of 35% and therefore a selection of the most confident predictions is not possible (see Section 3.2.5).



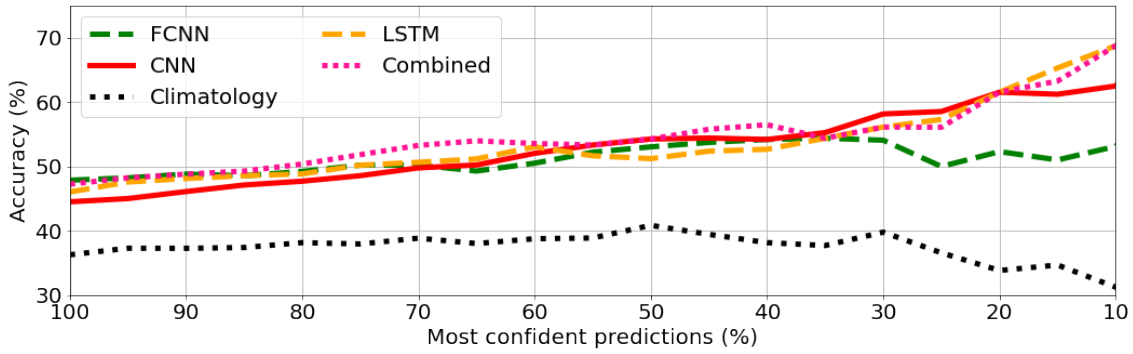
(a) Fully connected neural network

(b) Long Short-term memory

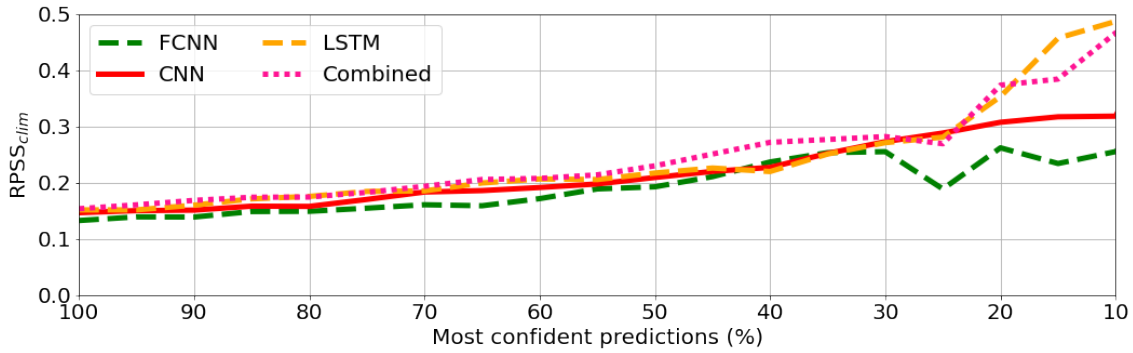
(c) Convolutional neural network

Figure 4.2: Deterministic and probabilistic predictions of the three ML models, FCNN (a), LSTM (b) and CNN (c), for a lead time of 6 days. The black crosses represent the IFS forecast skill category (right y-axis) for all reforecast dates in the testing period (x-axis). The purple plus signs indicate the predicted forecast skill category by the ML model. To identify the correct predictions more easily, the purple plus signs are lowered for correct predictions. The bar plots in the background indicate for each prediction the confidence (left y-axis) of the three different categories. The green, orange and blue bars represent the good, normal and bad forecast skill category respectively.

Up to the 40% most confident predictions, the accuracy and $\text{RPSS}_{\text{clim}}$ (Figures 4.3a and 4.3b) are similar for all ML models and the performance steadily increases with a decreasing selection of predictions. Selecting the 10–40% most confident predictions, the performance is varying for the different ML models. The FCNN reaches similar skill levels as for less confident predictions. The



(a) Accuracy



(b) RPSS$_{\text{clim}}$

Figure 4.3: Increase of the performance measures, accuracy (a) and RPSS$_{\text{clim}}$ (b), with an increase of confidence. The selection of confident predictions ranges from 100% down to 10% (x-axis). 100% includes all predictions of the ML model, 10% include the 10% most confident predictions. The combination of the ML models is computed by taking the mean of their probabilistic predictions. For the accuracy, also the climatology is shown (black dotted line).

LSTM and CNN models continue to improve their performance and the LSTM is outperforming the CNN.
We generate a combined ML model by computing the mean probabilistic prediction out of the FCNN, LSTM and CNN. If all three ML models are confident about the prediction of the same forecast skill category at the same time, we expect the prediction to be correct. The performance of the combined ML model is for most confidence levels (Figure 4.3) as good or better than the most accurate individual ML model.

These results confirm the visual impression from Figure 4.2: The performance of the ML models increases the higher the confidence is. The probabilistic predictions contain valuable information about the confidence, and therefore about the reliability of the prediction.

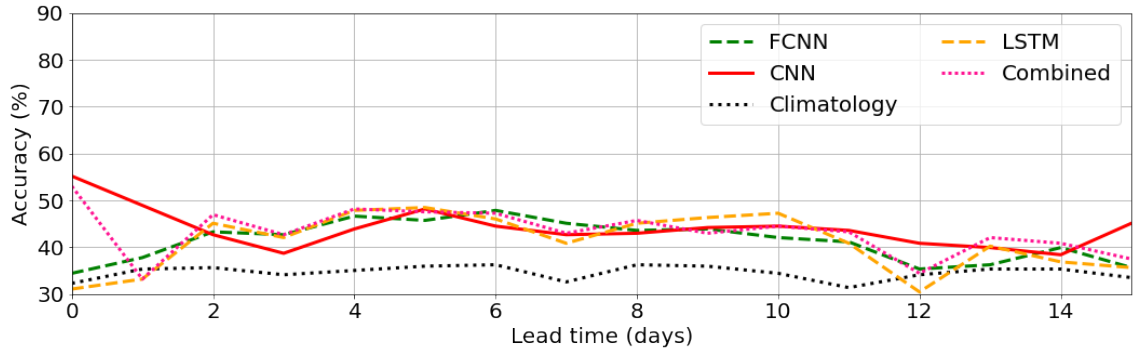
In comparison to the increase of accuracy for the ML models, the accuracy of the climatological prediction does not increase if only the most confident predictions are analysed (black dotted line in Figure 4.3a). The difference in the accuracy between the climatology and the ML models is larger than 10% for all confidence levels, with an increase of the difference for more confident predictions. A complete list of all performance measures for the 10% most confident predictions is given in Table 4.2.

Table 4.2: Summary of the performance measures for all ML models (FCNN, LSTM, CNN) and the climatological reference model at a lead time of 6 days for the 10% most confident predictions. This table is similar to Table 4.1, but only using the 10% most confident predictions.

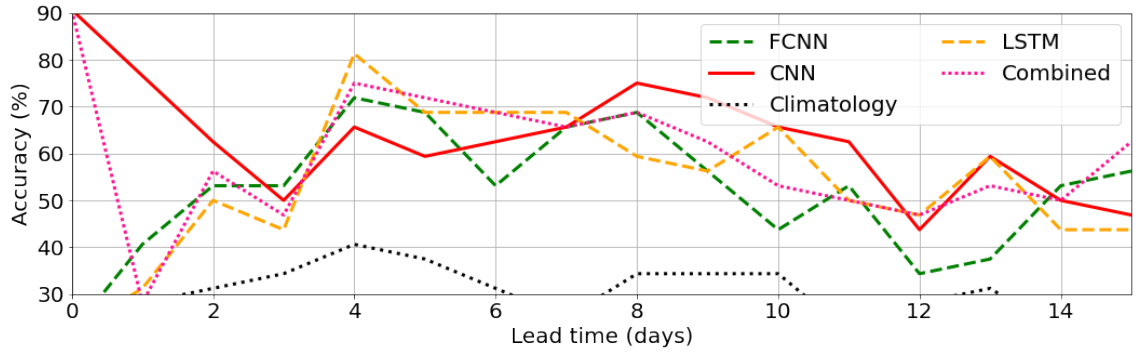
Model	Accuracy	RPS	RPSSclim	RPSSspread
FCNN	53.1	0.155	0.255	0.242
LSTM	68.8	0.121	0.486	0.457
CNN	62.5	0.148	0.31	0.324
Climatology	31.2	0.258	0.000	−0.194

4.3 Comparison of different lead times

After the intensive analysis of the performance of the ML models for a lead time of 6 days, we now analyse the performance for a lead time of 0 to 15 days. Lead times larger than 15 days are not analysed as all ML-models do not perform significantly better than climatology for longer lead times. We compare the accuracy and $\text{RPSS}_{\text{clim}}$ of all ML models with each other and with the climatological reference model for all predictions and the 10% most confident predictions.



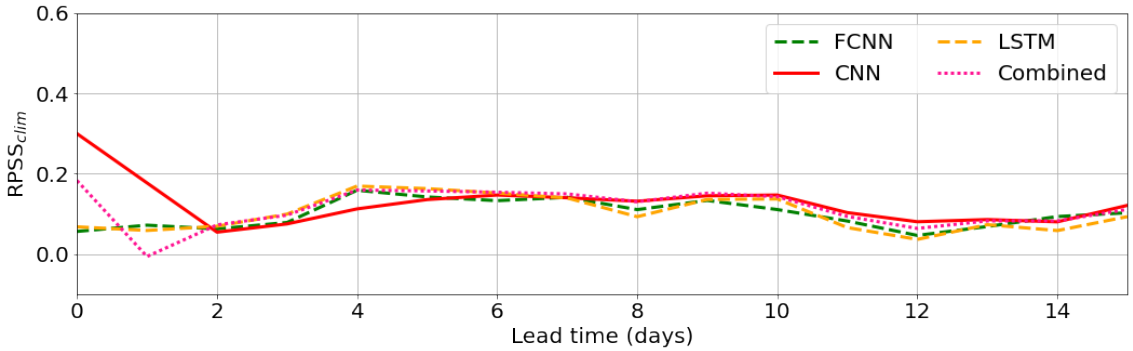
(a) Accuracy for all predictions



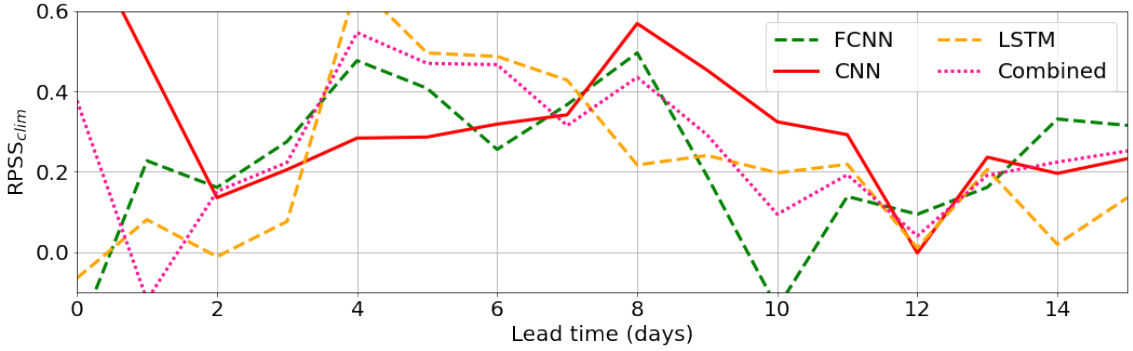
(b) Accuracy for the 10% most confident predictions

Figure 4.4: Accuracy (y-axis) for all predictions (a) and the 10% most confident predictions (b) for lead times from 0 to 15 days (x-axis). The three different ML models (FCNN, LSTM, CNN), the combination of them and the climatological reference prediction are compared.

The ML models perform best on lead times from 4 to 10 days. In this time range, the accuracy for all ML models is around 45% (Figure 4.4a) and the $RPSS_{clim}$ reaches values of 0.15 (Figure 4.5a), if all predictions are analysed. On longer lead times (11–15 days), the accuracy of the ML models decreases to 35–40%, with the CNN performing better than the FCNN and LSTM. On the longer lead times, the $RPSS_{clim}$ is reduced to values around 0.1 for all ML models. The CNN has only slightly higher $RPSS_{clim}$ values than the FCNN and LSTM. The performance of the combined ML model for all predictions is similar to that of the individual ML models, for all lead times. With 35% accuracy, the climatological reference model is inferior to the predictions of the ML models for almost all lead times.



(a) $RPSS_{clim}$ for all predictions



(b) RPSSclim for the 10% most confident predictions

Figure 4.5: $RPSS_{clim}$ for all predictions (a) and the 10% most confident predictions (b) for lead times from 0 to 15 days. The three different ML models (FCNN, LSTM, CNN) and the combination of them are compared.

If only the 10% most confident predictions of each ML model are chosen, an enormous boost in the accuracy and $RPSS_{clim}$ is observed for the lead times 4–10days. The LSTM performs best at the shorter lead time ranges (4-7 days), with an accuracy of up to 80% (Figure 4.4b) and a $RPSS_{clim}$ up to 0.6 (Figure 4.5b). On longer lead time ranges (8–11 days), the CNN outperforms the other ML models with an accuracy ($RPSS_{clim}$) up to 75% (0.55). The performance of the combined ML model predictions is approximately the mean of all three ML model predictions, therefore it performs better than the CNN on shorter lead times and better than the LSTM on longer lead times. For the confident predictions, the FCNN is performing the worst amongst the ML models. Still, the accuracy of all ML models is also for the confident predictions far better than the accuracy of the climatological reference model, which is not improving for more confident predictions.

4.4 Relative feature importance

In the previous sections, we compared the performance of the different ML model architectures, the FCNN, LSTM and CNN, with each other and with the non-ML reference models. To understand the decision making of the ML models in selecting a forecast skill category, we need to evaluate the importance of each feature (predictor) used in the ML models.
All features used for the ML models are known at the time of the forecast initialisation. All are independent of the forecast, except for the (spatial mean) ensemble spread of the Z500 field which is calculated using the 11 forecast members of the ECMWF reforecast. Generally, it would be desirable to only use features independent of the ECMWF reforecast to be able to predict the forecast skill prior to computing the ensemble reforecast. The following results show that excluding the ensemble spread, generated by the ECMWF reforecast, deteriorates the models' performance significantly.

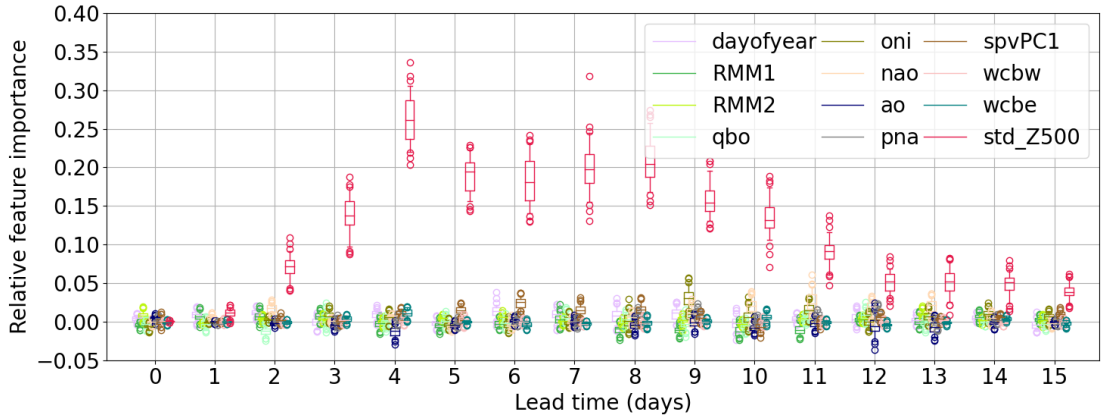
For all ML models on all lead times (excluding day 0 and day 2 predictions of the CNN model), the ensemble spread is the most important feature (Figure 4.6). Features like different climate modes for the FCNN and LSTM (Figures 4.6a and 4.6b) or atmospheric variables for the CNN (Figure 4.6c) influence the decision process only little in comparison to the ensemble spread.

The FCNN and LSTM models learn, next to the ensemble spread, little from the first principal component of the stratospheric polar vortex, the North Atlantic oscillation index and the Ocean Niño Index. For these, the RFI only ranges up to 0.05, where at the same lead times the RFI of the ensemble spread reaches values up to 0.30.

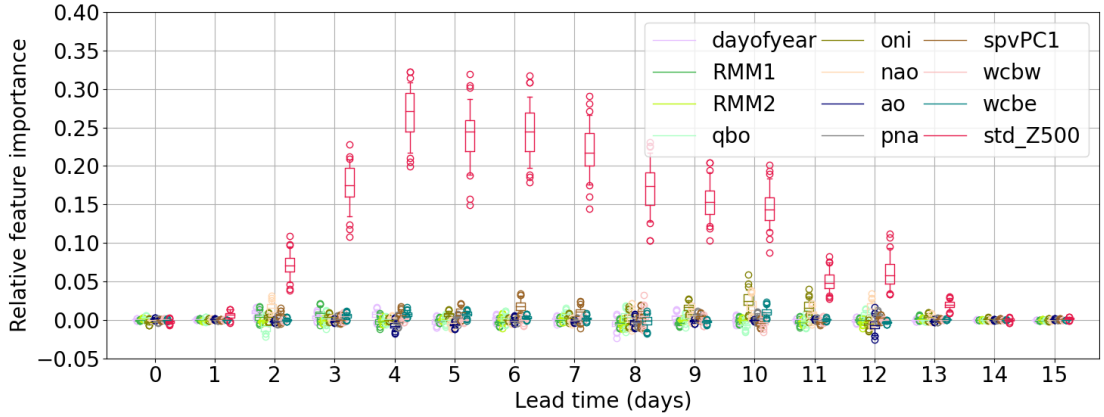
The CNN models (Figure 4.6c) learn, next to the large contribution of the ensemble spread, little by the 500 hPa geopotential height field and the zonal wind at 100 hPa. This coincides with the findings for the FCNN and LSTM, as the NAO and 500 hPa geopotential height field, and the SPV and zonal wind at 100 hPa are closely connected. The influence of the ONI can not be observed in the CNN models as the selected region (Euro-Atlantic) does not cover the area of the ONI.

The importance of the ensemble spread peaks in the middle of the analysed forecast range. For the FCNN there is a plateau from 4 to 8 days, the LSTM has a plateau from 4 to 7 days and with the CNN the peak is at 8 days. For the short and the long lead times, the importance of the ensemble spread decreases and also the models are performing worse (as seen in Figures 4.4 and 4.5). The decreased performance is explainable by the decreased ensemble spread importance. A possible explanation for this behaviour is provided in the last section of this chapter (Section 4.6).

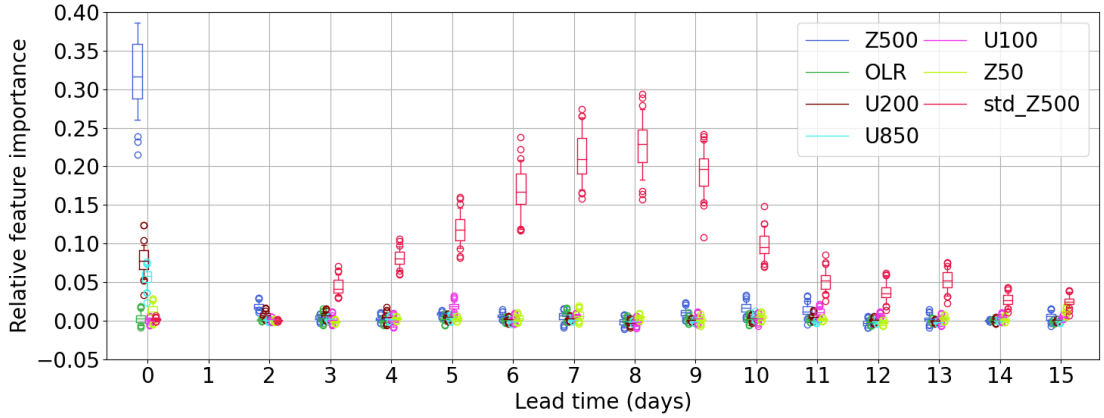
From these results we can conclude that the three ML models mainly learn from the Z500 ensemble spread, given by the eleven forecast members of the ECMWF reforecasts.



(a) Fully connected neural network



(b) Long short-term memory



(c) Convolutional neural network

Figure 4.6: Relative feature importance for each of the ML model architectures, FCNN (a), LSTM (b) and CNN (c), and lead times from 0 to 15 days. Positive values indicate that the ML models perform better with the use of the specific feature and the larger the positive value is, the more important is the particular feature for the model. The box and whisker plots represent the distribution of the feature importance for 50 independent runs, each time randomising the selected feature. The boxes (whiskers) range from the 25–75% (5%–95%) percentile. The solid line inside the boxes represents the median. If at one lead time no box and whisker plot is visible, the ML model was not able to produce skilful predictions for this lead time.

4.5 Machine learning models without the ensemble spread as a predictor

The Z500 ensemble spread is the most important predictor for all three ML models. But also the ensemble spread is the only predictor which requires information from the ensemble forecast. We are interested in how skilful the ML models are if they are trained and tested without the Z500 ensemble spread as a predictor. Operationally, this approach allows the ML models to run prior to the ensemble forecast run. The architecture of the ML models remains unchanged.

The FCNN and LSTM models learn most from the SPV, NAO and ONI predictors (Figures A.5a and A.5b in the Appendix). These are the same predictors that provide some importance next to the ensemble spread in the ML models that include the ensemble spread as a predictor (Figures 4.6a and 4.6b).
For a lead time of 6 days, the accuracy is reduced about 10% to 38.4% for the FCNN and 34.8% for the LSTM (Table A.1). The $RPSS_{spread}$ is for both ML models around 0.0 and therefore, the ML models are as skilful as the non-ML reference model using the ensemble spread. Next to a reduced accuracy and $RPSS_{spread}$, also the confidence of the models predictions is drastically lowered (Figures A.4a and A.4b).

The CNN model learns the most from the Z500 and U100 predictors (Figure A.5c). Again, these are the predictors that provide some importance for the CNN models that include the ensemble spread as a predictor (Figure 4.6c). The accuracy of the CNN model for predictions with a lead time of 6 days is only reduced by around 4% to 40.9% (Table A.1). But also here, the $RPSS_{spread}$ is with 0.012 close to zero and therefore the ML model is only as skilful as the non-ML spread reference model.

One major difference between the ML models with and without the use of the ensemble spread is not visualised in any of the figures. As explained in Section 3.2.4, 10 models with random initialisation parameters are trained and tested on the training data. The final model, which we analyse here, consists out of all ML models that achieve an accuracy of at least 35% on the training data. For most lead times and using the ML models with the ensemble spread as a predictor, the number of models that learn sufficiently on the training data is close to the maximum, namely 10 models. Without using the ensemble spread as a predictor, the variation in the number of models that learn sufficiently is large. At some lead times, only one or two models are sufficiently learning on the training data. A greatly reduced set of sufficiently skilled ML models is not advantageous, as it increases the probability that the final model will achieve a good result by chance. The results are less robust the fewer models are combined.

In summary, the ML models without the ensemble spread as a predictor perform as well as the non-ML spread reference model, but they perform significantly worse than the ML models that include the ensemble spread as a predictor. Therefore, in the further evaluation we continue with the CNN model that uses the ensemble spread as a predictor field.

4.6 Context to existing literature

With the work of Whitaker and Louche (1998), we can explain the difficulties of the ML models to predict the normal forecast skill category as seen in the confusion matrices (Figure 4.1). As mentioned in Section 2.2.2, the ensemble spread is likely to be most useful as a predictor of skill when it is either very large or very small compared to its climatology. The ML models are relying on the ensemble spread and therefore correctly predicting the good and bad forecast skill category is easier as the ensemble spread anomalies for those categories are generally larger than for the normal forecast skill category.

Mayer and Barnes (2020a) also observe a performance increase with an increase in the confidence and use the confidence of their ML model to determine the forecasts of opportunity, periods of atmospheric conditions that lead to an enhanced predictability. Their research question differs from ours, they predict the sign of the 500 hPa geopotential height anomaly over the North-Atlantic and not the forecast skill. In their study, they analyse the day-22 forecasts and they have only two categories to predict, positive or negative Z500 anomaly. Therefore, their baseline for the accuracy is at 50%, which equals randomly drawing one of the two categories. In our scenario, the baseline is at 33% due to selecting a category out of three categories. Our ML models and their ML model achieve a slightly better accuracy than the baseline if all predictions are considered (around 46% for our models (Table 4.1) and 58% for their model (Figure S3 in Mayer and Barnes (2021)). For all ML models the accuracy is increasing with an increase in confidence. Using the 10% most confident predictions, the accuracy is around 65% for our LSTM and CNN model (Table 4.2) and at 74% for their model (Figure S3 in Mayer and Barnes (2021)). The increase of accuracy is comparable between the two different works.

The reason why the ensemble spread is less valuable for shorter and longer lead times is given by Whitaker and Louche (1998). They analyse the spread-error correlation for the National Centers for Environmental Prediction (NCEP) ensemble forecast. Even though the forecasts are based on a different forecast system (NCEP instead of ECMWF) and are 24 years old, the explanation is likely transferable. The spread-error correlation for the 250 hPa geopotential height field north of 25 °N peaks at a forecast lead time of 5 days. For longer lead times, the ensemble distribution approaches the climatological distribution, which does not change from day to day. At these lead times, the forecast error is simply a random draw from a fixed distribution and the spread-error correlation has to decrease to zero. At shorter lead times, the model is not able to produce ensemble spread variability and therefore the spread-error correlation is low. These two factors lead to the spread-error correlation peaking in the medium-range, at around 5 days which is also seen in the feature importance of our ML models (we observe the peak at 5–8 days, Figure 4.6). Ferranti et al. (2015) and Leutbecher and Palmer (2007) reinforce the findings of Whitaker and Louche (1998) as they state that the spread-error relationship for the ECMWF forecast system is strongest for lead times from 5 to 10 days.

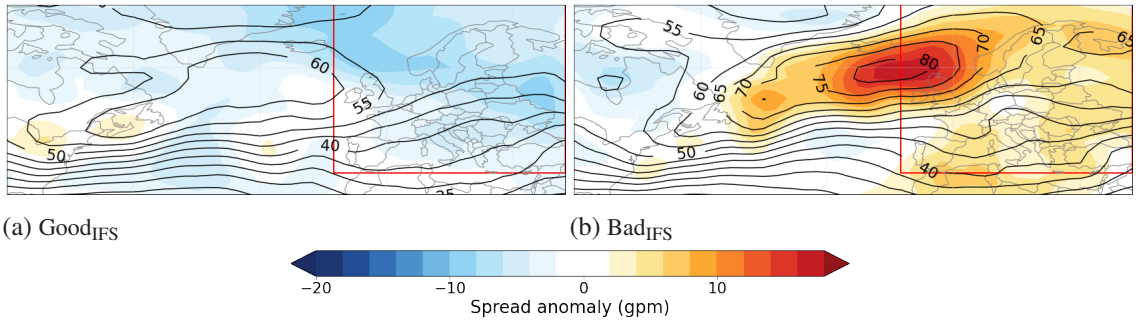
The approach to not include the ensemble spread as a predictor for the ML models is similar to the approach of Scher and Messori (2018). Their ML models are not as good as the ensemble

spread forecast in predicting the forecast error, but they outperform two other reference models. This reinforces our approach of using the ensemble spread as a predictor for our ML models, as they outperform the reference models.

5 Analysing categorised forecast skill

In the next two chapters, we focus again on the forecast skill of day-6 forecasts. We have learned from the feature importance (Figure 4.6) that the ensemble spread of the Z500 reforecast is the most important predictor field for the convolutional neural network (CNN). Therefore, we analyse in the next chapters the Z500 ensemble spread and the Z500 field in more detail. Prior to analysing the correct and incorrect predictions of the CNN in Chapter 6, we need to look at the actual (IFS) forecast skill categories. We limit the analysis to the good_{IFS} and bad_{IFS} forecast skill categories. For the Z500 field, we are not only interested in the field at initialisation time which serves as the predictor for the CNN, but also at lead time, as the ensemble spread is computed for the Z500 field at lead time. We also compare the Z500 fields at initialisation and lead time to better understand if the temporal change of the atmospheric flow pattern could be an indicator of the forecast skill category.

5.1 Ensemble spread



(a) Good_{IFS} (b) Bad_{IFS}

Figure 5.1: Composites of the Z500 ensemble spread for reforecasts with a lead time of 6 days for the good_{IFS} forecast skill category (a) and the bad_{IFS} category (b). The contour indicates the absolute Z500 ensemble spread in gpm and the anomaly of the Z500 ensemble spread to the NDJFM climatology is indicated with the shading. The Euro-Atlantic region is shown and the European region is indicated by the red rectangle. The European region is used for generating the categorised RMSE (the forecast skill categories).

The spread-error relationship explained in Section 2.2.2 suggests that the intrinsic predictability is low (bad_{IFS} forecast skill category) when the ensemble spread is large and high (good_{IFS} forecast skill category) when the ensemble spread is small. The Z500 ensemble spread composites of the good_{IFS} (Figure 5.1a) and bad_{IFS} (Figure 5.1b) forecast skill categories indeed represent the spread-error relationship.

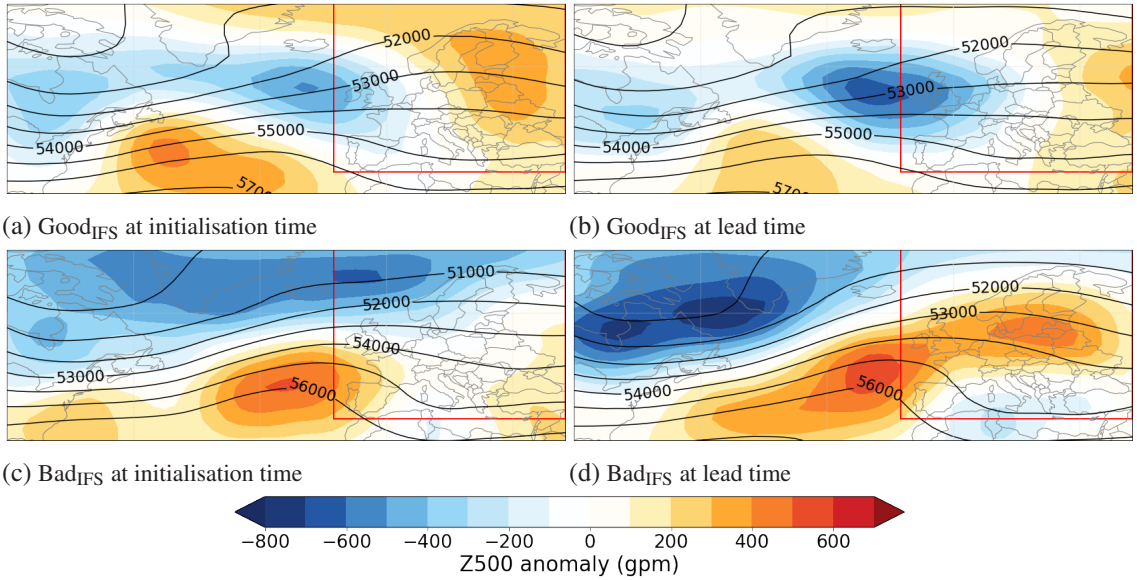
The composite of the good_{IFS} forecast skill (Figure 5.1a) shows a generally low absolute ensemble spread (up to 60 gpm, shown as contour) in the Euro-Atlantic region with a significant negative

anomaly (up to –10 gpm, shown as shading) in the European region, the region for which the RMSE is calculated.

Contrary, the Z500 ensemble spread is anomalously high (up to 80 gpm) during forecasts in the bad_{IFS} forecast skill category (Figure 5.1b) with a positive anomaly up to +16 gpm in the northern Atlantic, south of Greenland. The area with strong positive anomalies and therefore large absolute ensemble spread values extends from the Labrador Sea to northern Norway.

A large Z500 ensemble spread is likely associated with unstable atmospheric conditions on a synoptic scale and strong gradients in the 500 hPa geopotential height field, indicating a strong jet. On the other hand, low Z500 ensemble spread values are probably due to stationary large-scale patterns without strong gradients.

5.2 500 hPa geopotential height



(a) $Good_{IFS}$ at initialisation time

(b) $Good_{IFS}$ at lead time

(c) Bad_{IFS} at initialisation time

(d) Bad_{IFS} at lead time

Figure 5.2: Composites of the Z500 field for a lead time of 6 days for the $good_{IFS}$ forecast skill at initialisation (a) and lead time (b), bad_{IFS} forecast skill at initialisation (c) and lead time (d). As in Figure 5.1, the Euro-Atlantic region is shown and the European region is indicated with the red rectangle. The contour indicates the Z500 field and the Z500 anomaly to the NDJFM climatology is expressed with the shading.

The composites of the Z500 field at the time of initialisation for the $good_{IFS}$ (Figure 5.2a) and bad_{IFS} (Figure 5.2c) forecast skill category reflect the assumptions made in the previous section. In the $good_{IFS}$ composite, the isohypses are further apart in the northern North Atlantic, south of Iceland, indicating a weaker jet stream and more stability than in the bad_{IFS} composite, where the Z500 gradient is stronger. The region with a strong Z500 gradient is identical to the region where large Z500 ensemble spread occurs.

The development of the large-scale circulation (Z500) from initialisation to lead time differs significantly between the $good_{IFS}$ and bad_{IFS} forecast skill categories. For the $good_{IFS}$ forecast skill, a more blocked state in the European region at initialisation time (Figure 5.2a) changes to a more zonal state at lead time (Figure 5.2b). Contrary, a more zonal state in the European region for the

bad_{IFS} forecast skill at initialisation time (Figure 5.2c) changes to a blocking state at lead time (Figure 5.2d).

The large-scale circulation pattern for the different composites represent a mix of different weather regimes (we use the 7 weather regime definition). The Z500 field for the $good_{IFS}$ forecast skill at initialisation time (Figure 5.2a) is ridging in the northern European area. A negative Z500 anomaly in the North Atlantic south of Iceland is accompanied by two positive Z500 anomalies in the northeast European and western North Atlantic regions. This pattern can be a combination of the Scandinavian blocking (ScBL), Zonal regime (ZO) and the Atlantic trough (AT). At lead time (Figure 5.2b), the negative anomaly south of Iceland intensifies and the positive anomalies are weakening. The ridge over Scandinavia is also weakening. The zonal flow in the European region could indicate an increasing contribution of cyclonic weather regimes, especially the ZO and AT.
The characteristic Z500 anomalies at initialisation time for the bad_{IFS} forecast skill category (Figure 5.2c) are a strong positive anomaly over the Azores and a negative anomaly extending from Greenland to northern Norway. This flow pattern with a strong gradient in the Z500 field in the North Atlantic can be a combination of the cyclonic regimes ZO and Scandinavian trough (ScTr). At lead time (Figure 5.2d), the ridging and positive anomaly over the Azores intensifies and extends into northern Europe. The negative anomaly intensifies in the region of the Labrador Sea. This is a strong shift from the cyclonic regimes at initialisation time to blocking regimes such as the Atlantic ridge (AR) or European blocking (EuBL) at lead time.

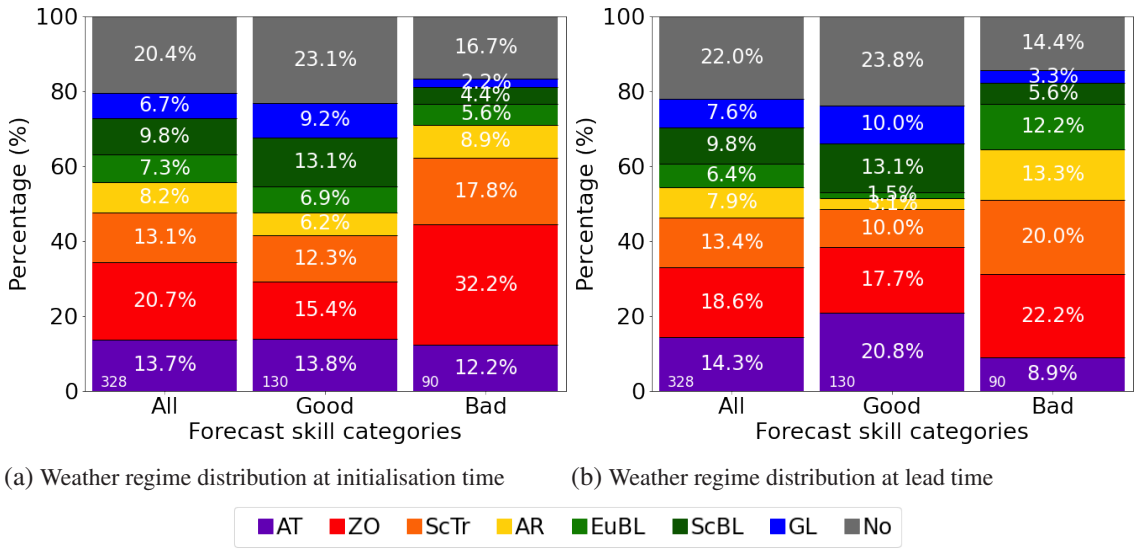
In summary, it is striking that both forecast skill categories seem related to a change in the large-scale circulation. The Z500 field in Europe changes from a ridging to a more zonal flow in the $good_{IFS}$ forecast skill category and vice versa from a zonal flow to a blocking situation in the bad_{IFS} category.

5.3 Weather regimes

We quantify our hypothesis from the previous section that for the $good_{IFS}$ and bad_{IFS} forecast skill categories a transition in the large-scale circulation pattern in the European region is observed. First, we analyse the distribution of the seven weather regimes at initialisation and at lead time for the different forecast skill categories. Afterwards, we compare the frequency of regime transitions for both forecast skill categories. The three cyclonic regimes consist of the Atlantic trough (AT), Zonal regime (ZO) and Scandinavian trough (ScTr) and the blocking regimes are the Atlantic ridge (AR), European blocking (EuBL), Scandinavian blocking (ScBL) and the Greenland blocking (GL).

5.3.1 Distribution of weather regimes at initialisation and lead time

Validating the composites of the large-scale circulation pattern analysed in the previous section by examining the distribution of weather regimes shows that each composite is not dominated by an individual weather regime, but that a tendency in the transition of weather regimes from initialisation (Figure 5.3a) to lead time (Figure 5.3) is apparent.



(a) Weather regime distribution at initialisation time (b) Weather regime distribution at lead time

Figure 5.3: Weather regime frequency distribution at initialisation time (a) and at a lead time of 6 days (b) for the IFS forecast skill categories. The seven weather regime definition, introduced in Chapter 2, is used and each regime is indicated with its colour as shown in the legend. AT, ZO and ScTr are considered as cyclonic regimes. Blocking regimes are: AR, EuBL, ScBL and GL. Analysing all reforecasts (left column), there are more cyclonic (around 47%) than blocking regimes (around 32%) active which is characteristic for the winter time period analysed here. The values at the bottom left of each column indicate the number of reforecasts that are analysed for the specific forecast skill category.

For the good_{IFS} forecast skill category, the contribution of cyclonic regimes to the composite increases from 42% at initialisation (Figure 5.3a) to 49% at lead time (Figure 5.3b). As assumed in the previous section, the AT and ZO are mainly responsible for the increase of cyclonic regimes. The reduction of blocking regimes (from 35% to 28%) is due to the AR and EuBL. The relatively large contribution of the ScBL remains unchanged from initialisation to lead time.

For the bad_{IFS} forecast skill category, a large contribution of the ZO (32%) at the time of initialisation is noticeable. The contribution of the cyclonic regimes decreases from 62% at initialisation to 51% at lead time with a strong decrease of ZO and also AT. The blocking regimes contribute to 21% at initialisation and 34% at lead time to the composite, with a strong increase in EuBL and AR, as suggested in the previous section.

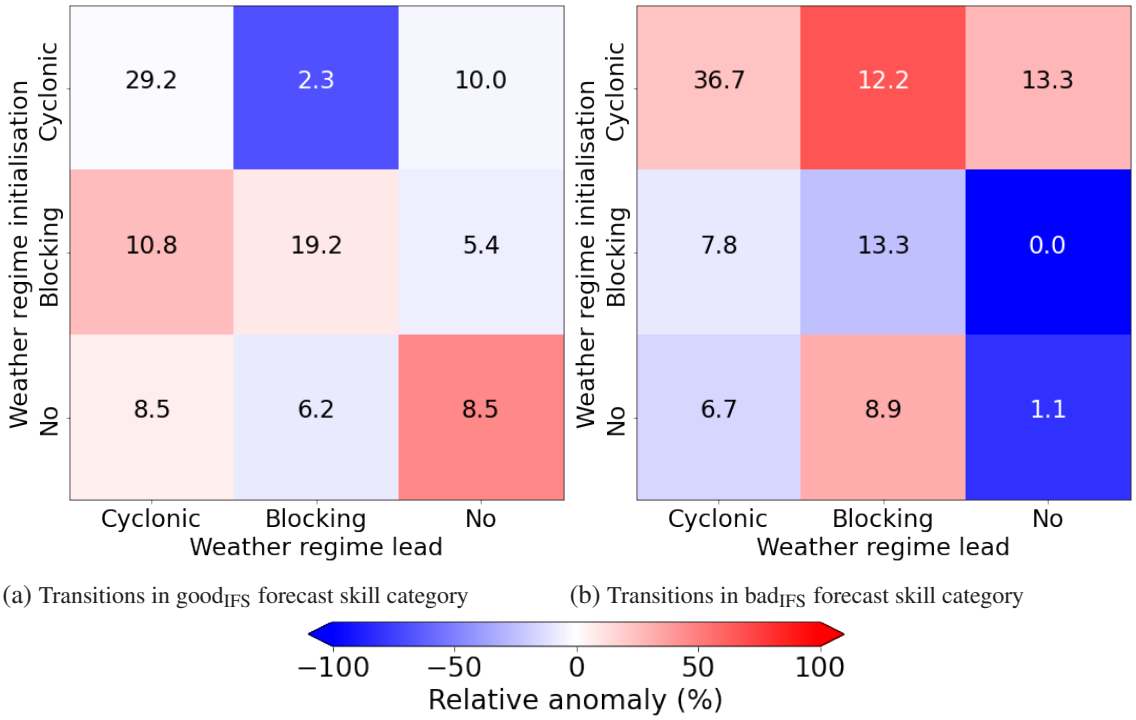
Even though the cyclonic regimes are dominant in all categories, which is characteristic for the winter time period analysed here, a clear tendency in the distribution of weather regimes from initialisation to lead time is noticeable. The analysis of the weather regime distribution confirms the hypothesis from the previous section: Good_{IFS} forecast skill is associated with a transition from ridge dominated regimes in the European region at initialisation to a zonal flow at lead time.

Bad_{IFS} forecast skill is associated with a transition from a zonal flow at initialisation to a ridging at lead time.

5.3.2 Weather regime transitions from initialisation to lead time

In the previous sections, we analysed the Z500 composites and the distribution of weather regimes at initialisation and lead time. The results can indicate change in the large-scale circulation pattern and transitions of the weather regimes. Now, we analyse the transitions of weather regimes directly to confirm the findings from the previous analysis. For a simpler representation, we combine the weather regimes to three categories: cyclonic regimes (AT, ZO, ScTr), blocking regimes (AR, EuBL, ScBL, GL) and no regime.

The mean duration of weather regimes in the analysed period is about 10 days. Analysing 6 day forecasts, weather regimes are not changing at many instances due to their life time. Furthermore, a transition from a cyclonic regime into another cyclonic regime does not count as a transition in our simplified representation, therefore we expect a large contribution of persistent regimes.



(a) Transitions in $good_{IFS}$ forecast skill category (b) Transitions in bad_{IFS} forecast skill category

Figure 5.4: Transition frequencies of weather regimes from initialisation time (y-axis) to lead time (x-axis) of 6 day reforecasts for the $good_{IFS}$ (a) and bad_{IFS} (b) forecast skill categories. The weather regimes from the seven weather regime definition are summarised into cyclonic (AT, ZO, ScTr), blocking (AR, EuBL, ScBL, GL) and no regimes. The 9 values sum up to 100% and therefore show the absolute distribution of the transition frequencies. The shading indicates the relative increase (red) or decrease (blue) of the transition frequency to all reforecasts during the four year testing period.

The anomalies to the climatological regime transition frequencies have an opposite sign for the $good_{IFS}$ and bad_{IFS} forecast skill categories (Figure 5.4).

Transitions from a blocking to a cyclonic regime are anomalously high (red shading) for the

$good_{IFS}$ forecast skill category (Figure 5.4a). The transition from cyclonic to blocking regimes is anomalously low (blue shading). The persistence of cyclonic regimes is in absolute values higher than the persistence of blocking regimes, this is due to two reasons: the climatological higher contribution of cyclonic regimes during winter time, but also transitions from a cyclonic regime to another cyclonic regime, such as from the ZO to AT or ScTr, supporting our hypothesis that the ridge in the European region is weakening for the $good_{IFS}$ forecast skill category.

For the bad_{IFS} forecast skill category (Figure 5.4b), the opposite signs for the transition frequency anomalies are observed. An increase in transition frequencies from cyclonic to blocking regimes (red shading) and a decrease from blocking to cyclonic regimes (blue shading), again confirming the hypothesis of a development towards a blocking situation in the European region for the bad_{IFS} forecast skill category.

The results of all three figures (Figures 5.2-5.4) corroborate each other in the assumption that the atmospheric flow pattern in the European region is changing from a ridging to a zonal flow for the $good_{IFS}$ forecast skill category and from a zonal flow to a blocking situation for the bad_{IFS} forecast skill category.

5.4 Context to existing literature

The Z500 ensemble spread composite (Figure 5.1) for the $good_{IFS}$ (bad_{IFS}) forecast skill category shows low (high) values in the European region. This nicely aligns with the spread-error relationship given by Leutbecher and Palmer (2007) or shown by Ferranti et al. (2015), indicating high (low) forecast skill for forecasts with a small (large) ensemble spread. Ferranti et al. (2015) states that the spread is a good indicator of the expected forecast error and therefore provides information about the forecast uncertainty.

The change of the atmospheric flow pattern for the different IFS categories is consistent with results from Ferranti et al. (2015), investigating the flow-dependent skill of the ECMWF ensemble predictions, using the four Euro-Atlantic climatological regimes. They conclude that poor forecast skill (large RMSE values) is observed in persisting zonal flows (NAO+) and the transition to a blocking pattern (BL). These two transition/persistence frequencies are also increased for the bad_{IFS} category (Figure 5.4b). Davini et al. (2017) and Quinting and Vitart (2019) also state that numerical weather prediction models have poor skill in reproducing atmospheric blocking patterns in the Euro-Atlantic sector. Büeler et al. (2021) mention a decreased forecast skill for the European and Scandinavian Blocking compared to the other regimes. The weather regime distribution for the bad_{IFS} forecast skill category (Figure 5.3) indicates an increase of these two regimes at lead time, which aligns with the findings of Büeler et al. (2021).

6 Meteorological interpretation of convolutional neural networks

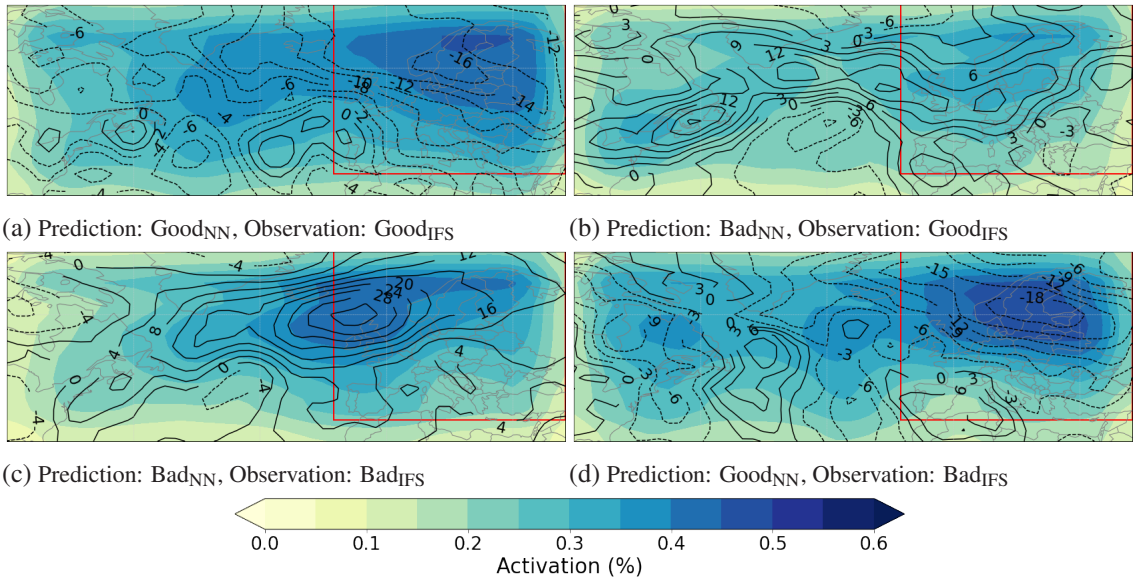
In the previous chapter, the ensemble spread and Z500 field have been analysed for the different forecast skill categories of the IFS. In this chapter, we do a similar analysis, but we include the forecast skill given by the convolutional neural network (CNN). This analysis allows an insight in the CNN. We answer the question from which features and at which area the CNN learns and why the CNN has difficulties in predicting the correct forecast skill category in some scenarios. Afterwards, we discuss the development of the Z500 field from initialisation to lead time and how more a priori knowledge about the Z500 field could benefit the CNN and therefore improve the accuracy of the forecasts.
As in previous chapters, the focus is on the good_{IFS} and bad_{IFS} forecast skill of the IFS model. In both categories, we generate subsets based on the prediction of a forecast skill category by the CNN. We end up with four combinations to analyse: $\text{good}_{\text{NN}}\text{good}_{\text{IFS}}$, $\text{bad}_{\text{NN}}\text{good}_{\text{IFS}}$, $\text{bad}_{\text{NN}}\text{bad}_{\text{IFS}}$, $\text{good}_{\text{NN}}\text{bad}_{\text{IFS}}$. In this notation, the former category always refers to the prediction of the neural network (NN) and the latter to the forecast skill given by the Integrated Forecasting System (IFS). For example, the combination $\text{good}_{\text{NN}}\text{bad}_{\text{IFS}}$ includes all reforecasts where the IFS model has the category bad_{IFS} forecast skill but the neural network predicts good_{NN} forecast skill. Based on the insight from the relative feature importance in Section 4.4, the ensemble spread and Z500 fields are analysed in more detail. Once again, we investigate day-6 forecasts.

6.1 Class activation mapping

In Section 4.4 it was shown that the Z500 ensemble spread is distinctively the most important feature for the CNN. The spread-error relationship (explained in Section 2.2.2) suggests good forecast skill when the ensemble spread is low and bad forecast skill when the ensemble spread is high. Therefore, our hypothesis is that the CNN detects large and small ensemble spread anomalies in the European region and predicts the bad_{NN} and good_{NN} forecast skill categories accordingly.

The CNN seems to have learned the spread-error relationship to predict the forecast skill category. The correct CNN predictions, $\text{good}_{\text{NN}}\text{good}_{\text{IFS}}$ (Figure 6.1a) and $\text{bad}_{\text{NN}}\text{bad}_{\text{IFS}}$ (Figure 6.1c), show the maximum amplitude of the ensemble spread anomalies (contour) in the European region. The highest activation values, and thus the area from which the CNN has learned most, are co-located with the area of maximum amplitudes of the ensemble spread anomalies.

The ensemble spread composites for the wrong predictions by the CNN demonstrate that the CNN did indeed learn the relationship between the ensemble spread and forecast error/forecast skill and that the CNN relies on this relationship. Rodwell et al. (2013) state that the ensemble spread does



(a) Prediction: Good_{NN}, Observation: Good_{IFS}

(b) Prediction: Bad_{NN}, Observation: Good_{IFS}

(c) Prediction: Bad_{NN}, Observation: Bad_{IFS}

(d) Prediction: Good_{NN}, Observation: Bad_{IFS}

Figure 6.1: Class activation mapping for forecasts with a lead time of 6 days for the combinations: $\text{good}_{\text{NN}}\text{good}_{\text{IFS}}$ (a), $\text{bad}_{\text{NN}}\text{good}_{\text{IFS}}$ (b), $\text{bad}_{\text{NN}}\text{bad}_{\text{IFS}}$ (c), $\text{good}_{\text{NN}}\text{bad}_{\text{IFS}}$ (d). The anomalies of the Z500 ensemble spread in relation to the 30 day running climatology (contour) serve as a predictor for the CNN. The activation of the CNN (shading) indicates regions that are more relevant for the CNN (darker colours). The selected region is the Euro-Atlantic region and the European region is indicated with a red rectangle.

not correctly predict the forecast error for every forecast, but averaged over many forecast start dates, the mean ensemble spread should match the mean ensemble error. The CNN relies on the ensemble spread and therefore at some reforecasts it is mislead by an ensemble spread which does not represent the correct forecast skill.

This effect is clearly visible for the $\text{good}_{\text{NN}}\text{bad}_{\text{IFS}}$ combination (Figure 6.1d). The CNN learns the most from the European region where a strong negative ensemble spread anomaly is located, thus indicating the good_{NN} forecast skill category whereas bad_{IFS} forecast skill is observed.

The effect is less clear in the composite of the ensemble spread for the $\text{bad}_{\text{NN}}\text{good}_{\text{IFS}}$ combination (Figure 6.1b). In this composite, the Z500 ensemble spread anomalies are less pronounced but mostly positive in the European region. The small amplitude of the anomalies causes the CNN to be less confident about its prediction. The CNN is not activated in a specific region as much as for the three other composites with larger amplitudes of the anomalies.

In summary, the CNN's decision making is understandable and similar to the approach a „human forecaster" would take if the forecaster was provided with the anomaly of the ensemble spread for the Euro-Atlantic region as a decision support: If the ensemble forecast has large (small) ensemble spread anomalies in the European region, the forecast skill is bad (good). This approach leads in many scenarios to the correct prediction of the forecast skill category, but sometimes it is misleading.

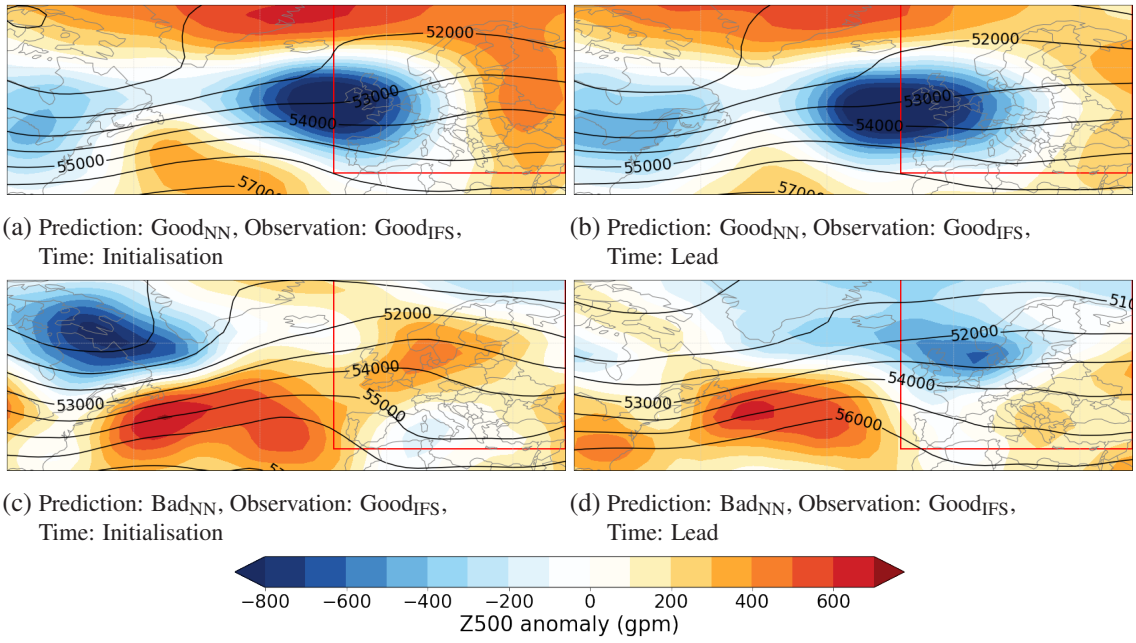
Even though the decision making seems to be similar to the approach of a human learning from the spread-error relationship, the CNN is performing better than using a non-ML approach (see $\text{RPSS}_{\text{spread}}$ in Chapter 4). When comparing the accuracy of the CNN with the accuracy a human

achieves, looking only at the ensemble spread, the CNN might outperform the human. To test this hypothesis I performed a self experiment: I have visualised the Z500 ensemble spread anomaly in the Euro-Atlantic region for each reforecast in the testing period (similar to the composites in Figure 5.1) and assigned one of the three forecast skill categories to each reforecast subjectively. With this procedure, I have achieved an accuracy of 40%, which is worse than the accuracy the CNN has achieved (44.5%).
The success of the CNN over a human can be due to the other features which the CNN analyses in combination with the ensemble spread or the deficit of the human to remember and learn from all scenarios of the training set.

6.2 Development of the 500 hPa geopotential height for different CNN prediction scenarios

The spread-error relationship, which the CNN learned by itself, is often a good indicator of the forecast skill category, but also misleads the CNN in some scenarios. With the knowledge of the change of the Z500 field for different forecast skill categories (Chapter 5), we formulate the hypothesis that the development of the Z500 field is a valuable indicator whether we can trust the prediction by the CNN. We discuss the validity of our hypothesis by analysing the Z500 field at the initialisation and the lead time for the good_{IFS} and bad_{IFS} forecast skill categories.

6.2.1 Good_{IFS} forecast skill category



(a) Prediction: Good_{NN}, Observation: Good_{IFS}, Time: Initialisation

(b) Prediction: Good_{NN}, Observation: Good_{IFS}, Time: Lead

(c) Prediction: Bad_{NN}, Observation: Good_{IFS}, Time: Initialisation

(d) Prediction: Bad_{NN}, Observation: Good_{IFS}, Time: Lead

Figure 6.2: Composites of the Z500 field at initialisation and at lead time of 6 days for all forecasts in the good_{IFS} forecast skill category. The Z500 field for the Euro-Atlantic region is shown in the contour. The Z500 anomaly is shown in the shading. The anomaly fields at initialisation time ((a), (c)) serve the CNN as predictor field. The fields at lead time ((b), (d)) are unknown at the time of initialisation but the ECMWF ensemble forecast of this field is used to generate the ensemble spread which serves as a predictor. In the first (latter) two composites, (a) and (b) ((c) and (d)), the CNN predicts the good_{NN} (bad_{NN}) forecast skill category.

Both subsets of good_{IFS} forecast skill, $\text{good}_{\text{NN}}\text{good}_{\text{IFS}}$ and $\text{bad}_{\text{NN}}\text{good}_{\text{IFS}}$, indicate a transition in the large-scale atmospheric flow (Z500 field) from a ridging pattern in the European region at initialisation time to a zonal flow at lead time.

The change in the atmospheric flow for the correct ML model predictions ($\text{good}_{\text{NN}}\text{good}_{\text{IFS}}$) is less pronounced than for the incorrect predictions ($\text{bad}_{\text{NN}}\text{good}_{\text{IFS}}$). For the correct predictions, the weak riding pattern in Central Europe at initialisation time (Figure 6.2a) changes to a zonal flow at lead time (Figure 6.2b). The weather regimes that contribute most to the transition are the EuBL transitioning into the no regime, but also the no regime transitioning into the cyclonic regimes AT and ZO (Figure A.6a).

For the incorrect ML model prediction, a pronounced blocking pattern at initialisation time (Figure 6.2c) changes to a zonal flow in Europe at lead time (Figure 6.2d). The blocking pattern is a combination of the AR and EuBL regime. The AR transitions into the AT and ScTr and the EuBL into the ZO, all transitions from blocking to cyclonic regimes (Figure A.6b). Next to the transition into cyclonic regimes, the EuBL transitions to an extensive amount into the no regime. Transitions into the no regime could explain to a certain extend why the ML model is not able to predict the forecast skill correct as the atmospheric flow in the no regime is less stable and therefore in general more difficult to predict.

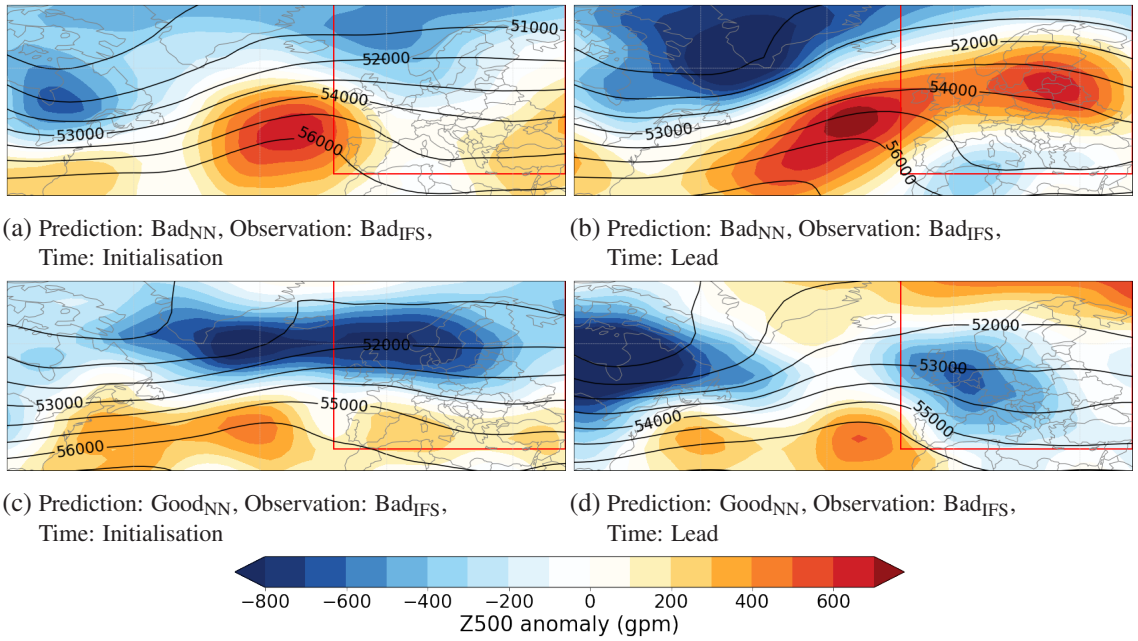
It is remarkable that especially for the incorrect predictions of the CNN the development of the large-scale atmospheric flow indicates a transition from blocking into zonal flow in the European region and therefore indicating the good forecast skill category. For the correct predictions, the transition is at a late state where the ridge at initialisation time is already weak.

6.2.2 Bad_{IFS} forecast skill category

The Z500 development for the bad_{IFS} forecast skill subsets, $\text{bad}_{\text{NN}}\text{bad}_{\text{IFS}}$ and $\text{good}_{\text{NN}}\text{bad}_{\text{IFS}}$ contrasts with the good_{IFS} forecast skill. The atmospheric flow in the European region changes from a zonal flow at initialisation time to a ridging pattern at lead time.

The formation of a ridge in the European region is more pronounced for the correct predictions ($\text{bad}_{\text{NN}}\text{bad}_{\text{IFS}}$) than for the incorrect predictions ($\text{good}_{\text{NN}}\text{bad}_{\text{IFS}}$). A ridge over the Azores at initialisation time (Figure 6.3a), best described by the cyclonic regimes ZO and ScTr (Figure A.7a), evolves into a ridge reaching over northern Europe at lead time (Figure 6.3b), best described by the blocked regimes AR and ScBL.

For the incorrect predictions, $\text{good}_{\text{NN}}\text{bad}_{\text{IFS}}$, the atmospheric flow in the European region is strongly zonal with only a small ridging over the Azores at initialisation time (Figure 6.3c) which should serve as an indicator for the bad_{IFS} forecast skill category. The atmospheric pattern at initialisation time is a combination of mainly the ZO and ScTr regime (Figure A.7b). The ZO transitions into a variety of blocked regimes (AR, EuBL, ScBL) but also into the no regime. The ScTr mainly transitions into the AR. The AR as a blocked regime together with the ScTr and ZO as



(a) Prediction: Bad_{NN}, Observation: Bad_{IFS}, Time: Initialisation

(b) Prediction: Bad_{NN}, Observation: Bad_{IFS}, Time: Lead

(c) Prediction: Good_{NN}, Observation: Bad_{IFS}, Time: Initialisation

(d) Prediction: Good_{NN}, Observation: Bad_{IFS}, Time: Lead

Figure 6.3: Composites of the Z500 field at initialisation and lead time of 6 days for all forecasts in the bad_{IFS} forecast skill category. The composites are identical to the composites in Figure 6.2 except that the IFS forecast skill category is bad_{IFS} instead of good_{IFS}. The anomaly fields at initialisation time ((a), (c)) serve the CNN as predictor field. The fields at lead time ((b), (d)) are unknown at the time of initialisation but the ECMWF ensemble forecast of this field is used to generate the ensemble spread which serves as a predictor. In the first (latter) two composites, (a) and (b) ((c) and (d)), the CNN predicts the bad_{NN} (good_{NN}) forecast skill category.

cyclonic regimes resemble the atmospheric flow at lead time (Figure 6.3d) best, as a ridge evolves in the eastern North Atlantic to northern European region.

In summary, the Z500 field changes from a zonal flow to ridging in the European region for both bad_{IFS} subsets. The ridge formation for the correct predictions is at an advanced state and therefore develops a significant ridge in the northern European region at lead time. For incorrect predictions, the ridge development is at an early state but the tendency of a ridge evolving in the European region is noticeable from initialisation to lead time.
Our hypothesis that the development of the Z500 field is a valuable indicator whether we can trust the CNN's prediction is justified, and we should take the development of the Z500 field into account when predicting the forecast skill.

6.3 Using CNN-based categorical forecast skill and improving the model

The development of the Z500 field from the initialisation to the lead time, which we show in Chapter 5 for the good_{IFS} and bad_{IFS} forecast skill categories, but also in this chapter for the four different subsets, $\text{good}_{\text{NN}}\text{good}_{\text{IFS}}$, $\text{bad}_{\text{NN}}\text{good}_{\text{IFS}}$, $\text{bad}_{\text{NN}}\text{bad}_{\text{IFS}}$ and $\text{good}_{\text{NN}}\text{bad}_{\text{IFS}}$, promises to provide valuable information for the user of our ML model. We propose two different ideas for using the information from the Z500 development: Either the information is directly integrated into the predictors of the ML model, thus improving the skill of the ML model, or the Z500 development is used to evaluate the prediction of the CNN at initialisation time. In order to use the CNN model operationally, we restrict both ideas to a priori knowledge.

The evaluation of the CNN prediction at the initialisation time has the advantage of informing the user of our ML model about the reliability of the prediction by using the Z500 field. If a good_{NN} forecast skill is predicted but a zonal flow is observed in the European region at the initialisation time and the weather forecast forecasts a transition to a ridge at the lead time, the user of our ML model should not trust the CNN prediction. Conversely, the user should not trust the CNN prediction if a bad_{NN} forecast skill is predicted and a ridge is observed in the European region at the time of initialisation, which is forecasted to transition to a zonal flow at lead time.
This method can be used with the current CNN model and does not require any further modification. But the method is probably not the most practical, as one still has to manually check the current and the forecasted Z500 field. Besides, there is already a confidence measure for the CNN model which we have explained in Chapter 4, using the probabilistic prediction of the forecast skill categories.

The more promising method is to include the Z500 fields of the recent past and/or the forecasted Z500 field up to lead time as predictor variables. In this way, the task of understanding the significance of the atmospheric flow change is part of the ML model and no additional human interaction is required. The inclusion of the additional Z500 predictor variables should improve the overall performance of the ML model and thus increase the accuracy and also the confidence of the ML model.
We propose two different approaches to implement the additional Z500 fields in the ML model. Either one stays with the CNN architecture and adds multiple past Z500 fields and forecasted Z500 fields as predictors, or one increases the complexity of the ML model by combining the CNN with an LSTM so that the Z500 predictor has two dimensions in space (longitude and latitude) and one dimension in time (from past Z500 fields to forecasted Z500 fields) instead of multiple Z500 predictors with no temporal dimension. Increasing the temporal dimension is not limited to the Z500 field and could be applied to all other atmospheric field variables used in the CNN. In Section 3.2.7 we described the set of seven predictors with two spacial dimensions as an image with seven colours. The CNN thus performs an image classification. The introduction of a temporal dimension next to the two spacial dimensions corresponds to a video classification. This new ML model has a more complex architecture than the CNN model but we expect it to be more skilful than all of the previous ML models applied in this Master's thesis.

6.4 Context to existing literature

Results from the class activation mapping are well understandable with the knowledge that the ensemble spread is the most important feature for the CNN. The class activation composites (Figure 6.1) indicate that the spread-error relationship as explained in Leutbecher and Palmer (2007) is learned by the CNN. Low (high) Z500 ensemble spread in the European region is an indicator for good (bad) forecast skill. Leutbecher and Palmer (2007) also state that the mean ensemble spread should match the mean ensemble error, therefore we expect that some predictions are mislead by the ensemble spread. These mislead predictions are the wrong predictions (Figures 6.1b and 6.1d). Analysing the ensemble spread field of the correct predictions (Figures 6.1a and 6.1c), we see that the extreme anomalies are more predictable than values close to climatology, as stated by Whitaker and Louche (1998).

Analysing the development of the atmospheric flow and the according transitions of weather regimes shows that for incorrect CNN predictions the contribution of the no regime is increased in comparison to the correct CNN predictions (Figures A.6 and A.7). The CNN has problems in correctly predicting the forecast skill category if the no regime is involved, either at initialisation or lead time. A possible explanation is given by Büeler et al. (2021). They state that it is difficult to predict phases that lack persistence, such as the no regime.

7 Conclusion and discussion

The chaotic nature of the atmosphere limits its intrinsic predictability (Lorenz, 1969). State-of-the-art numerical weather prediction models rarely reach this limit of intrinsic predictability. Their forecast skill horizon, also referred to as the limit of practical predictability, indicates the lead time at which a forecast is no longer skilful. The limit of practical predictability depends on the accuracy of the initial condition of the atmosphere, but also on model errors of the NWP model. The forecast skill horizon can exceed a range of two weeks for the probabilistic prediction of the 500 hPa geopotential height (Z500) field over the Northern Hemisphere, with a spacial resolution of 180 km (Buizza and Leutbecher, 2015). However, the forecast skill generally is flow dependent and sometimes, poor forecasts on shorter lead times can be observed, so called forecast busts (Rodwell et al., 2013). Previous studies such as Vitart and Molteni (2010) or Ferranti et al. (2018) have shown that the ensemble spread of an ensemble forecast, the atmospheric condition at initialisation time and slowly varying modes of the climate system can be indicators of the skill of the forecast. With this a priori knowledge, a statement about the quality of the forecast can be made at the time of initialisation of the forecast.

The aim of the Master's thesis is to develop machine learning (ML) models which predict the forecast skill of ECMWF's subseasonal reforecasts for the European region. 20 years of reforecasts (3980 individual reforecasts with 11 ensemble members) are available and we split them into 16 years (1997-2012) of training data and 4 years (2013-2016) of testing data. We evaluate the skill of the ML models by comparing their performance with the skill of non-ML approaches. The ML models only use a priori knowledge such as the state of different climate modes or atmospheric field variables at and prior to initialisation time or information given by the ensemble forecast such as the ensemble spread of the Z500 field. In the following, we answer the research questions, raised in the introduction (see Chapter 1):

1. **Is it possible to improve the prediction of the forecast skill using ML models?**
   Yes, all three ML models used in this work show an improvement in the prediction of the forecast skill compared to classical approaches using the climatology or the Z500 ensemble spread of the ECMWF reforecast.

2. **Are predictions of the forecast skill improving with an increasingly complex architecture of ML models?**
   We apply three different architectures of ML models, a fully connected neural network (FCNN), a long short-term memory (LSTM) and a convolutional neural network (CNN) (ordered by increasing complexity). The performance of predicting the forecast skill, measured by the accuracy and ranked probability skill score (RPSS), is similar for all ML models if all predictions in the testing period are analysed. The confidence of the predictions is in-

creasing with an increase of complexity. The FCNN is less confident than the LSTM and CNN model. If only confident predictions are analysed, for example the 10% most confident predictions in the test data set, the LSTM and CNN outperform the FCNN. Combining the probabilistic prediction of all three ML models improves the accuracy and RPSS at certain confidence levels and makes the predictions more robust.

3. **On which time-scale are ML models able to skilfully predict the forecast skill and which predictors are driving the decision making process?**
   Our ML models are skilful up to a lead time of 15 days. This does not imply that no ML models can be skilful for longer lead times.
   The predictors used can be split into two categories. First, we use predictors based on atmospheric field variables or parameters derived from the atmospheric field variables such as climate mode indices. Second, we employ predictors which are derived from the actual ensemble forecast, here the ensemble spread of the Z500 field. Both categories are available at initialisation time, though the ensemble spread of course is dependent on the ensemble forecast. In the time range with skilful predictions up to 15 days, the ensemble spread is the most important predictor and the main driver of the decision making process.

4. **When and why are ML models failing to predict the correct forecast skill?**
   The most important predictor for the ML models is the Z500 ensemble spread. Other predictors only play a minor role in the decision making process. Hence, a misleading sign of ensemble spread anomalies in the European region (or in the mean over the European region) can cause the ML model to fail. Accordingly, numerical weather prediction models that exhibit a poorer spread-error relationship than the ECMWF IFS will lead to a decreased skill of the ML models.

Many studies focus on the prediction of the forecast skill. Scher and Messori (2018) take the most similar approach to this Master's thesis. They also aim to predict the weather forecast uncertainty using ML models. Although there are many differences in the setup, it is still worthwhile to compare the results of their work with this Master's thesis. In contrast to this work, Scher and Messori (2018) do not use the ensemble spread as a predictor, as they want to be able to make predictions of forecast skill prior to the model run of the ensemble forecast. Based on the predictions of their ML model, they would be able to adapt for example the number of ensemble members which would eventually reduce the computational expenses of the ensemble forecast. Their models outperform other non-ML methods which are proposed in literature, but are not as skilful as the ensemble spread in predicting the forecast error.
If we do not include the ensemble spread in the predictor field, our ML models perform as good (or slightly better) than the reference models, using the climatology or the ensemble spread. Adding the ensemble spread as a predictor increases the skill and confidence of our ML models significantly.
Scher and Messori (2018) mention that they see their model as a tool for complementary guidance next to the ensemble forecasts. Their approach allows modifications on the ensemble forecast such as fewer ensemble members for highly predictable situations, our approach on the other side leads to more skilful predictions of the forecast skill. In an operational context, one needs to decide

if one wants to use an approach which is less accurate but runs prior to computing the ensemble forecast or an approach which is more accurate but runs after the computation of the ensemble forecast.
Both, their study and this Master's thesis, show that convolutional neural networks are capable to extract information on forecast uncertainty using a priori knowledge about the atmosphere.

Scher and Messori (2018) do not use the confidence of the ML model to make a statement about the reliability of the prediction. Mayer and Barnes (2020a) use the confidence of an interpretable neural network to investigate subseasonal forecasts of opportunity. In their study, they do not predict the forecast skill but the sign of the anomaly of the 500 hPa geopotential height field in the Atlantic region. They use the concept of selecting the most confident predictions and find a strong increase in the accuracy compared to using all predictions, similar to our work. Using all predictions, their ML model accuracy is also only slightly higher than the non-ML model reference accuracy and with an increasing confidence, the difference between the ML model and non-ML reference model accuracy is increasing.
Both, their study and this Master's thesis, conclude that with an increase of the model confidence also the prediction accuracy increases.

The Z500 ensemble spread is correlated with the ensemble forecast error (spread-error relationship. Therefore, the ensemble spread is on average a good indicator of the forecast skill category. Large (small) ensemble spread values in the European region are an indicator for the bad_{IFS} (good_{IFS}) forecast skill category. In the bad_{IFS} forecast skill category, large ensemble spread values from the Labrador Seas to northern Norway are associated with unstable atmospheric conditions on a synoptic scale and strong gradients in the Z500 field which make the atmosphere less predictable.
The CNN model, the most skilful ML model of the three models trained in this Master's thesis, has learned the relationship between the Z500 ensemble spread for the forecasted lead time and the forecast skill of the NWP model. Therefore the CNN imitates a human forecaster using the spread-error relationship to make a statement about the forecast skill of the NWP model at initialisation time. The performance of the CNN predicting the forecast skill is better than the performance of a human forecaster.
Incorrect predictions of the CNN model can be attributed to the fact that the ensemble spread does not always indicate the correct forecast skill. In these scenarios, but also in scenarios where the CNN predicts the correct forecast skill category, the development of the Z500 field from the initialisation time to the lead time is an indicator of the forecast skill category.
Good_{IFS} forecast skill is associated with a ridge in the European region at the initialisation time which weakens towards the lead time to a zonal flow. Conversely, for bad_{IFS} forecast skill, the atmospheric flow in the European region transitions from a zonal flow at the initialisation time to a blocking at the lead time.

If the CNN model from this thesis is used operationally to predict the forecast skill, the state of the Z500 field at initialisation time and the forecasted development of the Z500 field from initialisation to lead time could be used to support or to disagree with the predicted forecast skill.

We propose that implementing the development of the Z500 field as a predictor of the ML model improves the performance and confidence of the ML model. Using a time series of the Z500 field, but also other atmospheric field variables, from past observations to forecasted fields as predictors increases the dimensionality of the predictors. Increasing the complexity of the ML model could significantly improve the performance of the ML model in predicting the forecast skill category. The development of the Z500 field as a predictor for the ML model could eliminate the need for the Z500 ensemble spread as a predictor and thus provide a skilful ML model that predicts the forecast skill category prior to the NWP model run.

A Appendix

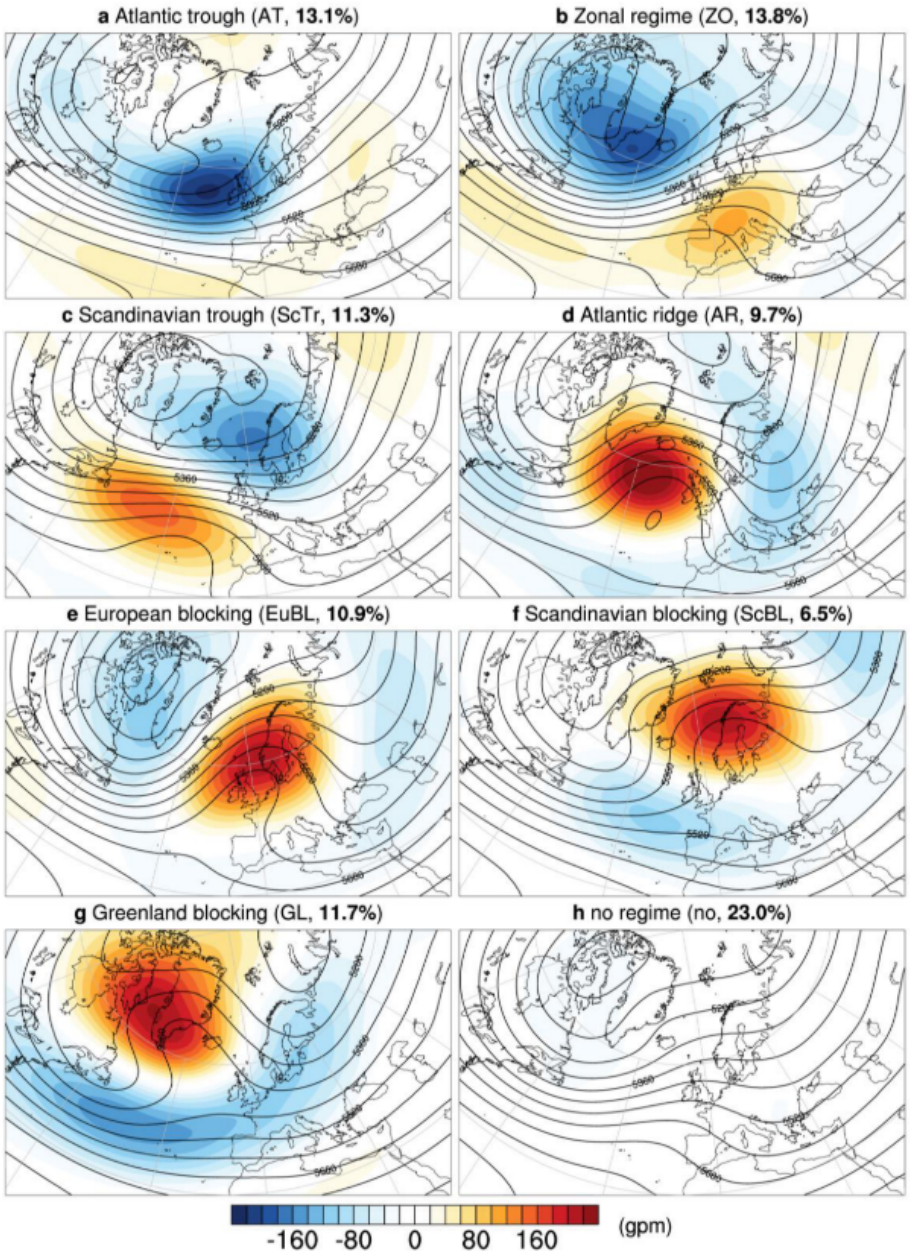
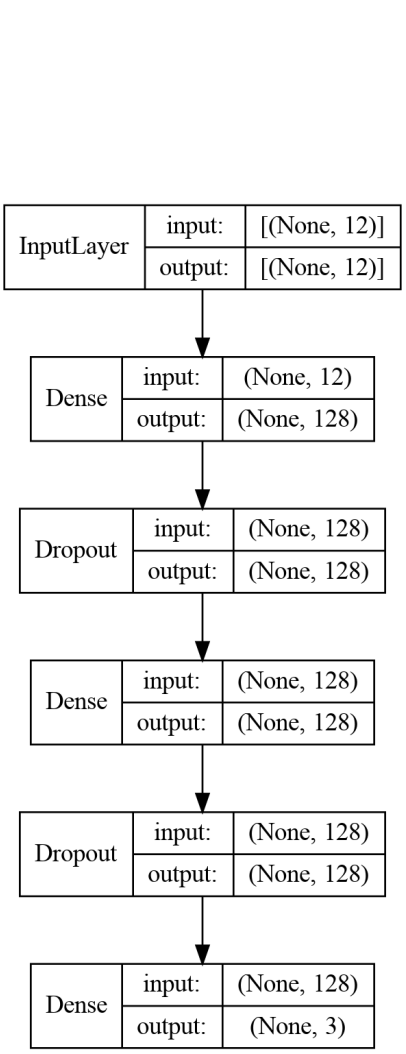
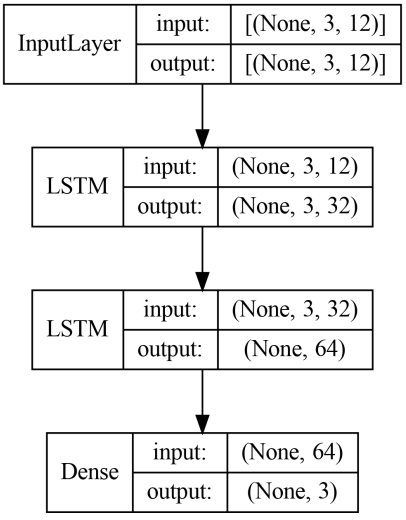


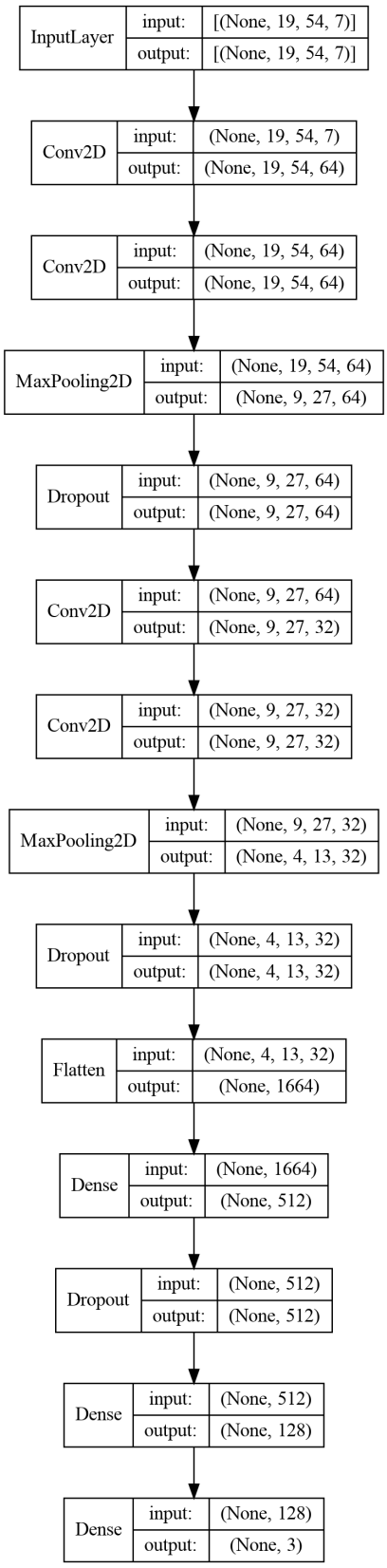
Figure A.1: Atlantic-European weather regimes in winter. Mean low-pass filtered (10 days) 500 hPa geopotential height anomaly (Z500', shading, every 20 geopotential meters), and mean absolute 500 hPa geopotential height (Z500, black contours, every 40 geopotential meters) in winter (DJF) for all days attributed to one of the 7 weather regimes (a-g) and to no regime (h). Although the regime definition is based on normalised data for the entire year, here non-normalised data for DJF are shown. Regime name, abbreviation, and relative frequency (for winter, in percent) are indicated in the sub-figure caption. Adapted from Grams et al. (2017b) supplementary Figure 1.



(a) Fully connected neural network



(b) Long short-term memory

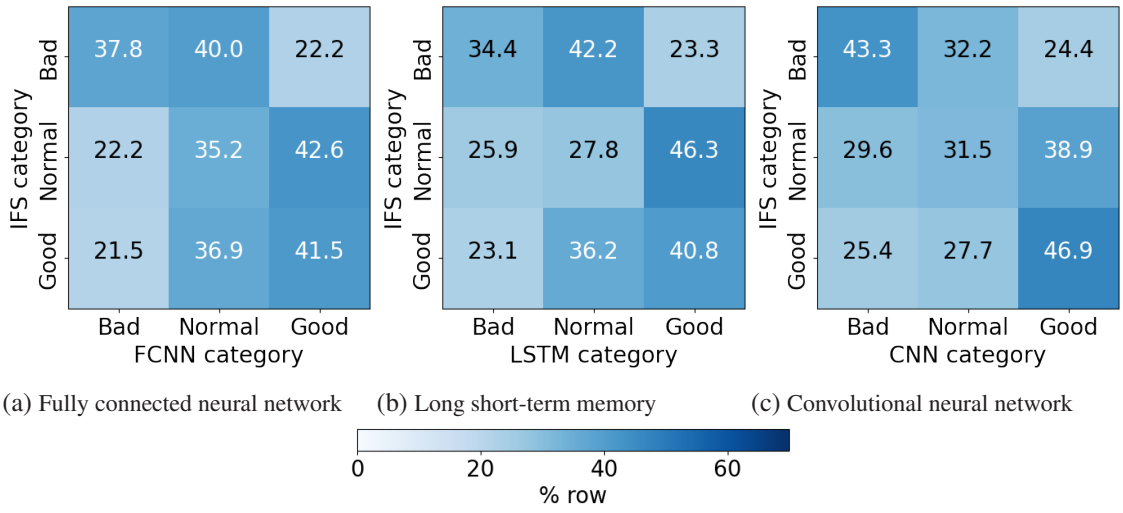


(c) Convolutional neural network

Figure A.2: Visualisation of the ML model architecture and parameters for the fully connected neural network (a), long short-term memory (b) and convolutional neural network (c) for day-6 forecasts.

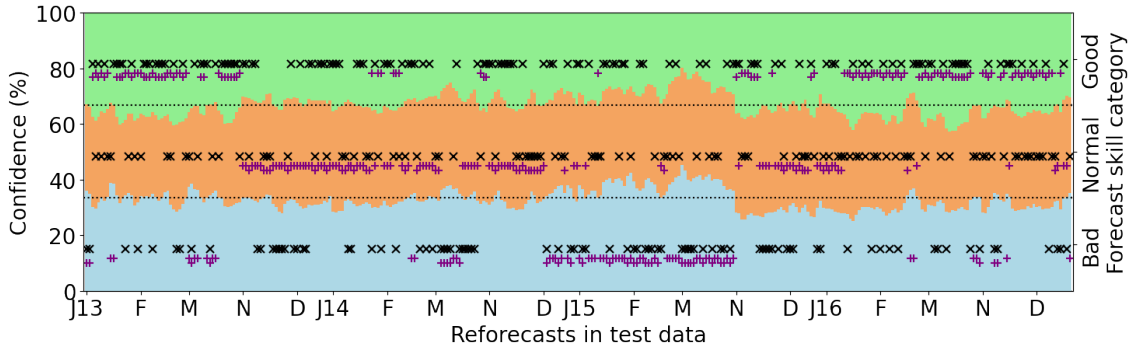
Table A.1: Summary of the performance measures for all ML models (FCNN, LSTM, CNN) at a lead time of 6 days. Here, the ML models do not use the ensemble spread as a predictor which is the difference to Table 4.1.

Model	Accuracy	RPS	RPSSclim	RPSSspread
FCNN	38.4	0.219	0.056	0.006
LSTM	34.8	0.221	0.048	−0.003
CNN	40.9	0.217	0.062	0.012

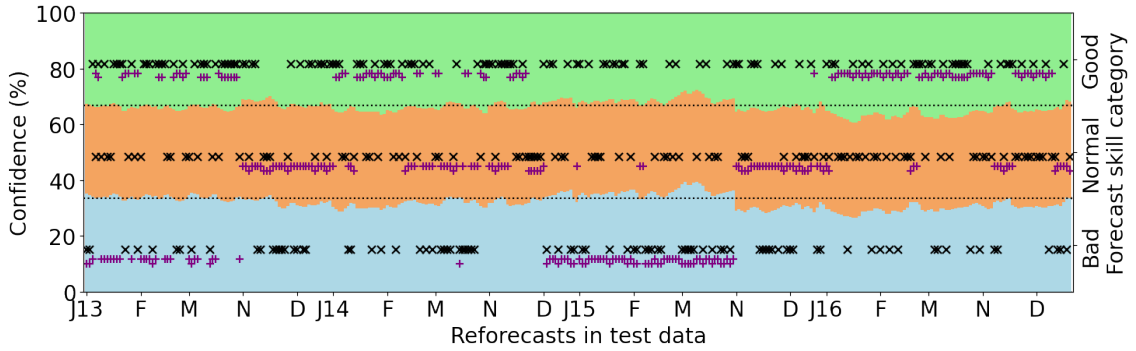


(a) Fully connected neural network (b) Long short-term memory (c) Convolutional neural network

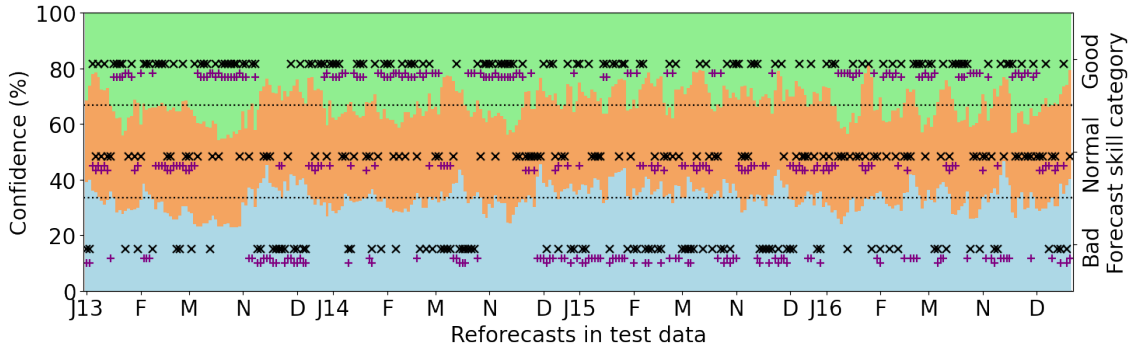
Figure A.3: Confusion Matrices for the three different ML models and a lead time of 6 days. Here, the ML models do not use the ensemble spread as a predictor which is the difference to Figure 4.1. The matrices indicate for each IFS category (y-axis) the distribution of the predictions made by the ML models (x-axis).



(a) Fully connected neural network

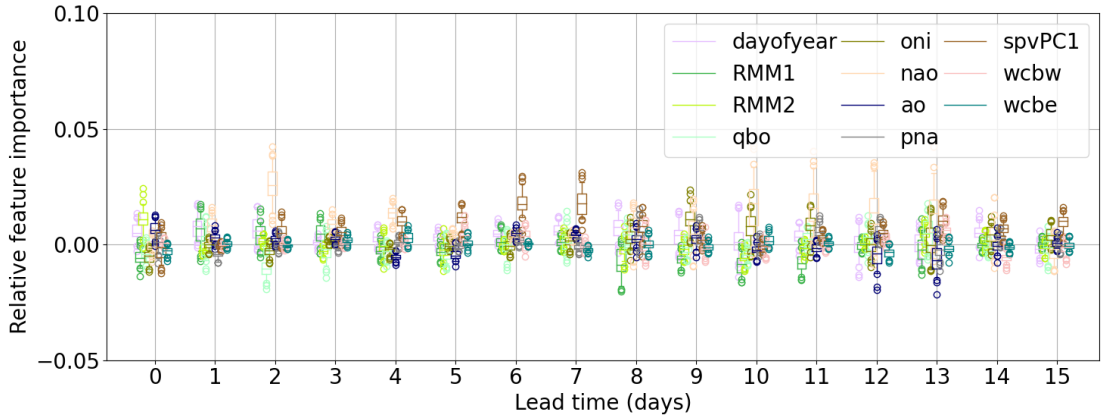


(b) Long Short-term memory

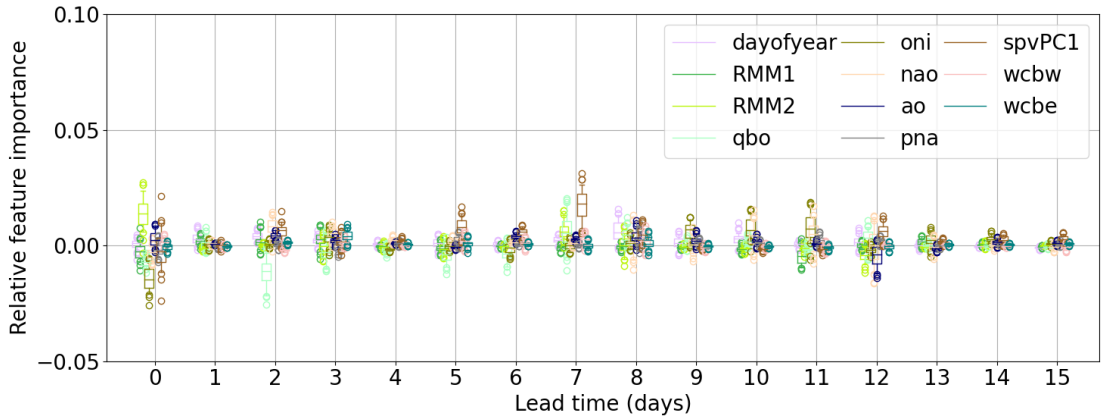


(c) Convolutional neural network

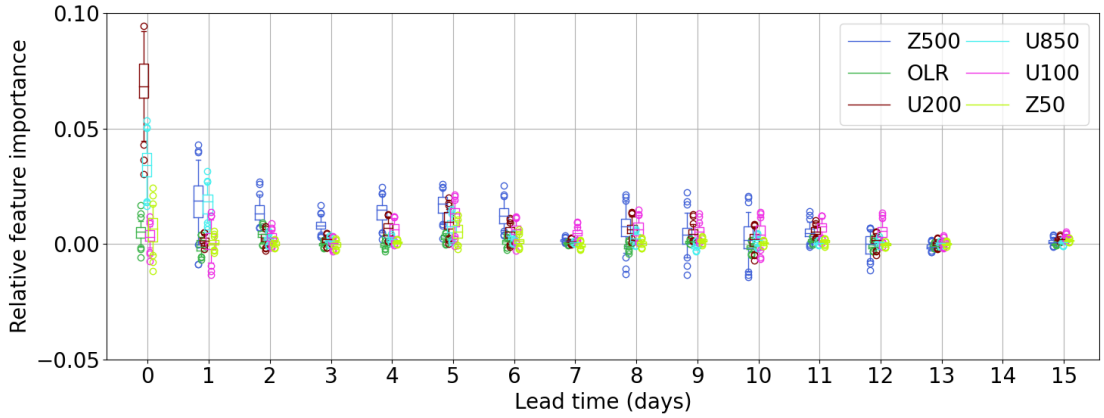
Figure A.4: Deterministic and probabilistic predictions of the three ML models for a lead time of 6 days. Here, the ML models do not use the ensemble spread as a predictor which is the difference to Figure 4.2. The black crosses represent the IFS forecast skill category, visible on the right y-axis, for all reforecast dates in the testing period, shown on the x-axis. The purple plus signs indicate the predicted forecast skill category by the ML model. To identify the correct predictions more easily, the purple plus signs are lowered for correct predictions. The bar plots in the background indicate for each prediction the confidence (left y-axis) of the three different categories. The green, orange and blue bars represent the good, normal and bad forecast skill category respectively.



(a) Fully connected neural network

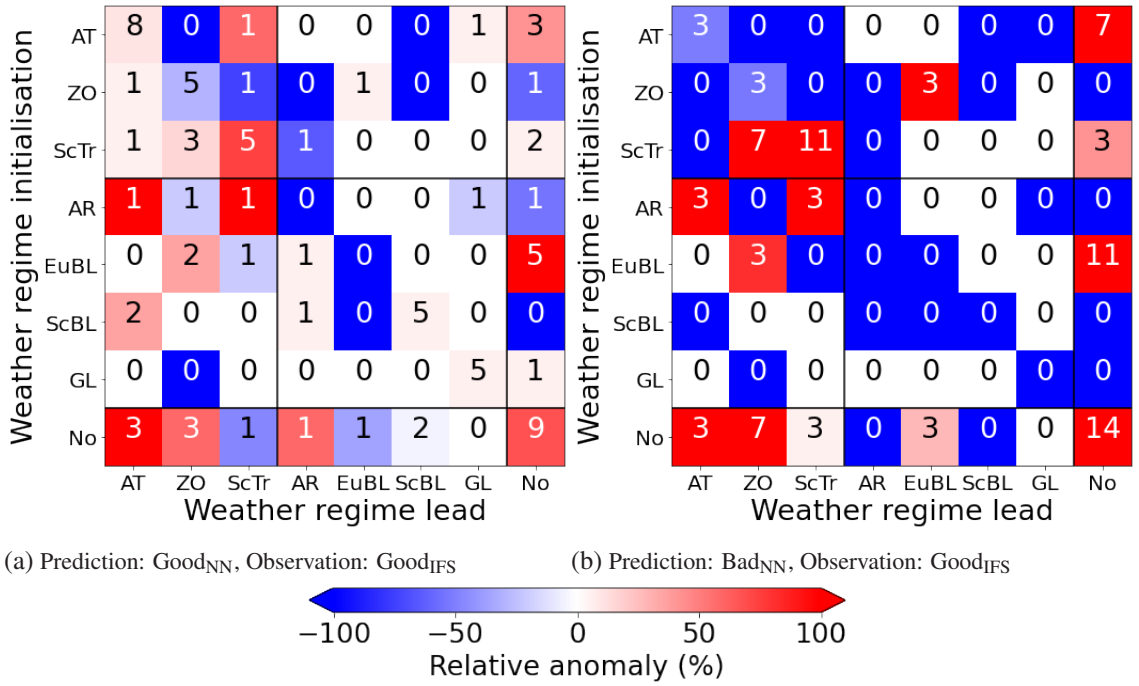


(b) Long short-term memory



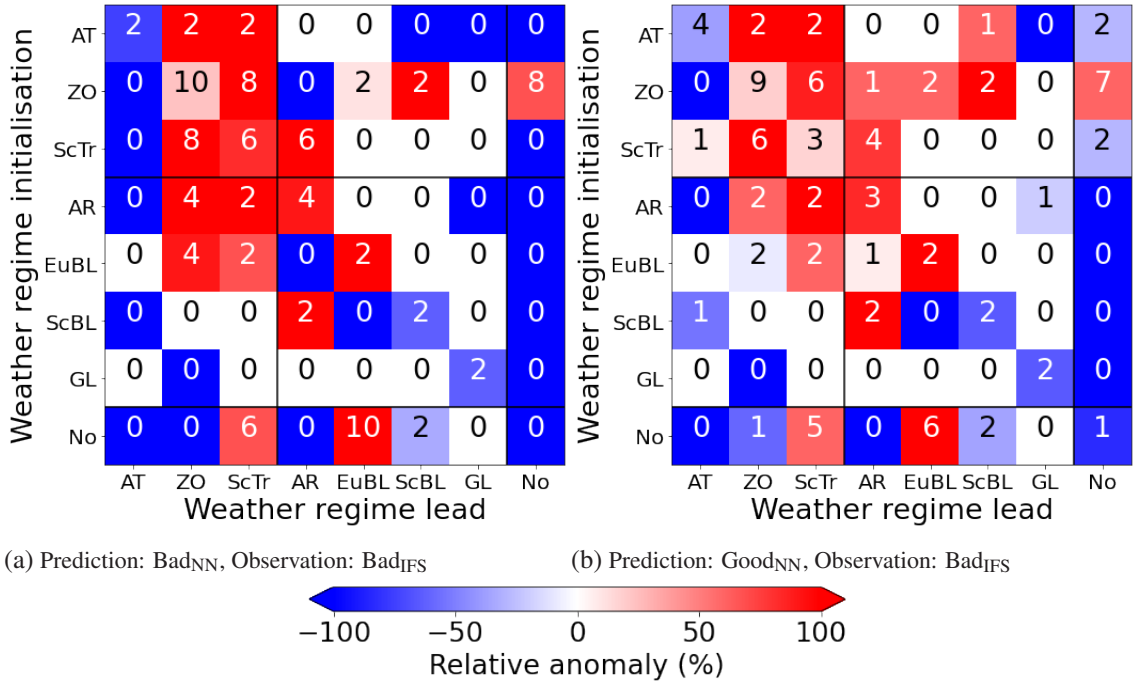
(c) Convolutional neural network

Figure A.5: Relative feature importance for each of the ML model architectures and lead times from 0 to 15 days. Here, the ML models do not use the ensemble spread as a predictor which is the difference to Figure 4.6. Positive values indicate that the ML models perform better with the use of the specific feature and the larger the positive value is, the more important is the particular feature for the model. The box and whisker plots represent the distribution of the feature importance for 50 independent runs, each time randomising the selected feature. The boxes (whiskers) range from the 25–75% (5%–95%) percentile. The solid line inside the boxes represents the median. If at one lead time no box and whisker plot is visible, the ML model was not able to produce skilful predictions for this lead time.



(a) Prediction: Good_{NN}, Observation: Good_{IFS} (b) Prediction: Bad_{NN}, Observation: Good_{IFS}

Figure A.6: Transition frequencies of weather regimes from initialisation time (y-axis) to lead time (x-axis) of day-6 reforecasts for the good_{IFS} forecast skill category in combination with the predicted forecast skill by the CNN, good_{NN} (a) and bad_{NN} (b). The 64 values sum up to 100% and therefore show the absolute distribution of the transition frequencies. The shading indicates the relative increase (red) or decrease (blue) of the transition frequency to all reforecasts during the four year testing period.



(a) Prediction: Bad_{NN}, Observation: Bad_{IFS} (b) Prediction: Good_{NN}, Observation: Bad_{IFS}

Figure A.7: Transition frequencies of weather regimes from initialisation time (y-axis) to lead time (x-axis) of day-6 reforecasts for the bad_{IFS} forecast skill category in combination with the predicted forecast skill by the CNN, bad_{NN} (a) and good_{NN} (b). The 64 values sum up to 100% and therefore show the absolute distribution of the transition frequencies. The shading indicates the relative increase (red) or decrease (blue) of the transition frequency to all reforecasts during the four year testing period.

Bibliography

Albers, J. R. and M. Newman, 2019: A Priori Identification of Skillful Extratropical Subseasonal Forecasts. *Geophysical Research Letters*, **46 (21)**, 12 527–12 536.

Australian Bureau of Meteorology, 2022: MJO indices. Accessed: 2022-03-08, http://www.bom.gov.au/climate/mjo/.

Baggett, C. F., E. A. Barnes, E. D. Maloney, and B. D. Mundhenk, 2017: Advancing atmospheric river forecasts into subseasonal-to-seasonal time scales. *Geophysical Research Letters*, **44 (14)**, 7528–7536.

Baldwin, M. P., L. J. Gray, T. J. Dunkerton, K. Hamilton, P. H. Haynes, J. R. Holton, M. J. Alexander, I. Hirota, T. Horinouchi, D. B. A. Jones, C. Marquardt, K. Sato, and M. Takahashi, 2001: The Quasi-Biennial Oscillation. *Reviews of Geophysics*, **39 (2)**, 179–229.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525 (7567)**, 47–55.

Bjerknes, V., E. Volken, and S. Brönnimann, 1904: The problem of weather prediction, considered from the viewpoints of mechanics and physics. *Meteorologische Zeitschrift*, **18 (6)**, 663–667.

Büeler, D., L. Ferranti, L. Magnusson, J. F. Quinting, and C. M. Grams, 2021: Year-round sub-seasonal forecast skill for Atlantic–European weather regimes. *Quarterly Journal of the Royal Meteorological Society*, **147 (741)**, 4283–4309.

Buizza, R. and M. Leutbecher, 2015: The forecast skill horizon. *Quarterly Journal of the Royal Meteorological Society*, **141 (693)**, 3366–3382.

CAWCR, 2015: Forecast Verification Research. Accessed: 2022-03-08, https://www.cawcr.gov.au/projects/verification/#Introduction.

Chollet, F., 2018: *Deep Learning with Python*. Manning Publications Co., Shelter Island, NY 11964.

Davini, P., S. Corti, F. D'Andrea, G. Rivière, and J. von Hardenberg, 2017: Improved Winter European Atmospheric Blocking Frequencies in High-Resolution Global Climate Simulations. *Journal of Advances in Modeling Earth Systems*, **9 (7)**, 2615–2634.

ECMWF, 2018: Forecast Ensemble (ENS) - Rationale and Construction. Accessed 2022-03-21, https://confluence.ecmwf.int/display/FUG/5+Forecast+Ensemble+%28ENS%29+-+Rationale+and+Construction.

———, 2020a: Data Assimilation Fact Sheet.

———, 2020b: ENS Mean and Spread. Accessed: 2022-03-08, https://confluence.ecmwf.int/display/FUG/ENS+Mean+and+Spread.

———, 2020c: Reanalysis Fact Sheet.

Ferranti, L., S. Corti, and M. Janousek, 2015: Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. *Quarterly Journal of the Royal Meteorological Society*, **141 (688)**, 916–924.

Ferranti, L., L. Magnusson, F. Vitart, and D. S. Richardson, 2018: How far in advance can we predict changes in large-scale flow leading to severe cold conditions over Europe? *Quarterly Journal of the Royal Meteorological Society*, **144 (715)**, 1788–1802.

Grams, C. M., R. Beerli, S. Pfenninger, I. Staffell, and H. Wernli, 2017a: Balancing Europe's wind-power output through spatial deployment informed by weather regimes. *Nature Climate Change*, **7 (8)**, 557–562.

———, 2017b: Balancing Europe's wind-power output through spatial deployment informed by weather regimes. *Nature Climate Change*, **7 (8)**, 557–562.

Grazzini, F. and F. Vitart, 2015: Atmospheric predictability and Rossby wave packets. *Quarterly Journal of the Royal Meteorological Society*, **141 (692)**, 2793–2802.

Grönaas, S., 1985: A pilot study on the prediction of medium range forecast quality. *ECMWF Technical Memorandum*, ECMWF.

Hersbach, H., B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J. N. Thépaut, 2020: The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **146 (730)**, 1999–2049.

Holton, J. R., 2002: Encyclopedia of Atmospheric Sciences.

Hopson, T. M., 2014: Assessing the ensemble spread-error relationship. *Monthly Weather Review*, **142 (3)**, 1125–1142.

Kaggle, 2019: Permutation Importance. Accessed: 2022-03-08, https://www.kaggle.com/dansbecker/permutation-importance.

Kalnay, E. and A. Dalcher, 1987: Forecasting forecast skill. *Monthly Weather Review*, **115 (2)**, 349–356.

Keras, 2022: Dropout Layer. Accessed: 2022-03-20, https://keras.io/api/layers/regularization_layers/dropout/.

König, G., C. Molnar, B. Bischl, and M. Grosse-Wentrup, 2020: Relative feature importance. *Proceedings - International Conference on Pattern Recognition*, 623–630, 2007.08283.

Lee, S. H., 2021: The stratospheric polar vortex and sudden stratospheric warmings. *Weather*, **76 (1)**, 12–13.

Leutbecher, M. and T. N. Palmer, 2007: Ensemble Forecasting. *Journal of Computational Physics*.

Liebmann, B. and C. A. Smith, 1996: Description of a complete (interpolated) outgoing longwave radiation dataset.

Lillo, S. P. and D. B. Parsons, 2017: Investigating the dynamics of error growth in ECMWF medium-range forecast busts. *Quarterly Journal of the Royal Meteorological Society*, **143 (704)**, 1211–1226.

Lim, Y., S. W. Son, A. G. Marshall, H. H. Hendon, and K. H. Seo, 2019: Influence of the QBO on MJO prediction skill in the subseasonal-to-seasonal prediction models. *Climate Dynamics*, **53 (3-4)**, 1681–1695.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21 (3)**, 289–307.

———, 1996: Predictability Problem Partly Solved.

Madden, R. A. and P. R. Julian, 1971: Detection of a 40-50 Day Oscillation in the Zonal Wind in the Tropical Pacific. *Journal of the Atmospheric Sciences*, **28 (5)**, 702–708.

———, 1972: Description of Global-Scale Circulation Cells in the Tropics with a 40–50 Day Period. *Journal of the Atmospheric Sciences*, **29 (6)**, 1109–1123.

———, 1994: Observations of the 40–50-Day Tropical Oscillation—A Review. *Journal of the Atmospheric Sciences*, **122 (5)**, 814–837.

Mariotti, A., C. Baggett, E. A. Barnes, E. Becker, A. Butler, D. C. Collins, P. A. Dirmeyer, L. Ferranti, N. C. Johnson, J. Jones, B. P. Kirtman, A. L. Lang, A. Molod, M. Newman, A. W. Robertson, S. Schubert, D. E. Waliser, and J. Albers, 2020: Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bulletin of the American Meteorological Society*, **101 (5)**, E608–E625.

Matsueda, M. and T. N. Palmer, 2018: Estimates of flow-dependent predictability of wintertime Euro-Atlantic weather regimes in medium-range forecasts. *Quarterly Journal of the Royal Meteorological Society*, **144 (713)**, 1012–1027.

Mayer, K. J. and E. A. Barnes, 2020a: Subseasonal Forecasts of Opportunity Identified by an Interpretable Neural Network. 1–11.

———, 2020b: Subseasonal midlatitude prediction skill following Quasi-Biennial Oscillation and Madden–Julian Oscillation activity. *Weather and Climate Dynamics*, **1 (1)**, 247–259.

———, 2021: Supporting Information: Subseasonal Forecasts of Opportunity Identified by an Explainable Neural Network. *Geophysical Research Letters*, **48 (10)**, 1–10.

Melhauser, C. and F. Zhang, 2012: Practical and intrinsic predictability of severe and convective weather at the mesoscales. *Journal of the Atmospheric Sciences*, **69 (11)**, 3350–3371.

NOAA, 2021: Pacific-North American (PNA). Accessed: 2021-05-17, https://www.ncdc.noaa.gov/teleconnections/pna/.

———, 2022a: Arctic Oscillation Index. Accesed: 2022-03-08, https://ftp.cpc.ncep.noaa.gov/cwlinks/norm.daily.ao.index.b500101.current.ascii.

———, 2022b: North Atlantic Oscillation Index. Accessed: 2022-03-19, https://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/nao_index.html.

———, 2022c: Ocean Nino Index. Accessed: 2022-03-08, https://www.cpc.ncep.noaa.gov/data/indices/oni.ascii.txt.

Palmer, T. N. and S. Tibaldi, 1988: On the prediction of forecast skill. *Monthly Weather Review*, **116 (12)**, 2453–2480.

Parsons, D. B., S. P. Lillo, C. P. Rattray, P. Bechtold, M. J. Rodwell, and C. M. Bruce, 2019: The role of continental mesoscale convective systems in forecast busts within global weather prediction systems. *Atmosphere*, **10 (11)**.

Pasquier, J. T., S. Pfahl, and C. M. Grams, 2019: Modulation of Atmospheric River Occurrence and Associated Precipitation Extremes in the North Atlantic Region by European Weather Regimes. *Geophysical Research Letters*, **46 (2)**, 1014–1023.

Quinting, J. F. and C. Grams, 2021: EuLerian Identification of ascending Air Streams (ELIAS 2.0) in Numerical Weather Prediction and Climate Models. Part I: Development of deep learning model. *Geoscientific Model Development Discussions*, 1–29.

Quinting, J. F. and F. Vitart, 2019: Representation of Synoptic-Scale Rossby Wave Packets and Blocking in the S2S Prediction Project Database. *Geophysical Research Letters*, **46 (2)**, 1070–1078.

Rodwell, M. J., L. Magnusson, P. Bauer, P. Bechtold, M. Bonavita, C. Cardinali, M. Diamantakis, P. Earnshaw, A. Garcia-Mendez, L. Isaksen, E. Källén, D. Klocke, P. Lopez, T. McNally, A. Persson, F. Prates, and N. Wedi, 2013: Characteristics of occasional poor medium-range weather forecasts for Europe. *Bulletin of the American Meteorological Society*, **94 (9)**, 1393–1405.

Rodwell, M. J., D. S. Richardson, D. B. Parsons, and H. Wernli, 2018: Flow-dependent reliability: A path to more skillful ensemble forecasts. *Bulletin of the American Meteorological Society*, **99 (5)**, 1015–1026.

Scher, S. and G. Messori, 2018: Predicting weather forecast uncertainty with machine learning. *Quarterly Journal of the Royal Meteorological Society*, **144 (717)**, 2830–2841.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, 2020: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, **128 (2)**, 336–359, 1610.02391.

Tennekes, H., A. Baede, and J. Opsteegh, 1986: Forecasting-Forecast-Skill. ECMWF, Shinfield Park, Reading.

Trenberth, K. and National Center for Atmospheric Research Staff (EDS), 2020: The Climate Data Guide: Nino SST Indices. Access: 2021-06-25, https://climatedataguide.ucar.edu/climate-data/nino-sst-indices-nino-12-3-34-4-oni-and-tni.

Tripathi, O. P., A. Charlton-Perez, M. Sigmond, and F. Vitart, 2015: Enhanced long-range forecast skill in boreal winter following stratospheric strong vortex conditions. *Environmental Research Letters*, **10 (10)**.

Vautard, R., 1990: Multiple Weather Regimes over the North Atlantic: Analysis of Precursors and Successors. *Monthly Weather Review*, **118 (10)**, 2056–2081.

Vitart, F., 2014: Evolution of ECMWF sub-seasonal forecast skill scores. *Quarterly Journal of the Royal Meteorological Society*, **140 (683)**, 1889–1899.

———, 2017: Madden—Julian Oscillation prediction and teleconnections in the S2S database. *Quarterly Journal of the Royal Meteorological Society*, **143 (706)**, 2210–2220.

Vitart, F., C. Ardilouze, A. Bonet, A. Brookshaw, M. Chen, C. Codorean, M. Déqué, L. Ferranti, E. Fucile, M. Fuentes, H. Hendon, J. Hodgson, H. S. Kang, A. Kumar, H. Lin, G. Liu, X. Liu, P. Malguzzi, I. Mallas, M. Manoussakis, D. Mastrangelo, C. MacLachlan, P. McLean, A. Minami, R. Mladek, T. Nakazawa, S. Najm, Y. Nie, M. Rixen, A. W. Robertson, P. Ruti, C. Sun, Y. Takaya, M. Tolstykh, F. Venuti, D. Waliser, S. Woolnough, T. Wu, D. J. Won, H. Xiao, R. Zaripov, and L. Zhang, 2017: The subseasonal to seasonal (S2S) prediction project database. *Bulletin of the American Meteorological Society*, **98 (1)**, 163–173.

Vitart, F. and F. Molteni, 2010: Simulation of the Madden-Julian oscillation and its teleconnections in the ECMWF forecast system. *Quarterly Journal of the Royal Meteorological Society*, **136 (649)**, 842–855.

Wandel, J., J. F. Quinting, and C. M. Grams, 2021: Toward a Systematic Evaluation of Warm Conveyor Belts in Numerical Weather Prediction and Climate Models. Part II: Verification of Operational Reforecasts. *Journal of the Atmospheric Sciences*, **78 (12)**, 3965–3982.

Wanner, H., S. Brönnimann, C. Casty, D. Gyalistras, J. Luterbacher, C. Schmutz, D. B. Stephenson, and E. Xoplaki, 2001: North Atlantic oscillation - Concepts and studies. *Surveys in Geophysics*, **22 (4)**, 321–381.

Wheeler, M. C. and H. H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Monthly Weather Review*, **132 (8)**, 1917–1932.

Whitaker, J. S. and A. F. Louche, 1998: The relationship between ensemble spread and ensemble mean skill. *Monthly Weather Review*, **126 (12)**, 3292–3302.

White, C. J., H. Carlsen, A. W. Robertson, R. J. Klein, J. K. Lazo, A. Kumar, F. Vitart, E. Coughlan de Perez, A. J. Ray, V. Murray, S. Bharwani, D. MacLeod, R. James, L. Fleming, A. P. Morse, B. Eggen, R. Graham, E. Kjellström, E. Becker, K. V. Pegion, N. J. Holbrook, D. McEvoy, M. Depledge, S. Perkins-Kirkpatrick, T. J. Brown, R. Street, L. Jones, T. A. Remenyi, I. Hodgson-Johnston, C. Buontempo, R. Lamb, H. Meinke, B. Arheimer, and S. E. Zebiak, 2017: Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteorological Applications*, **24 (3)**, 315–325.

Wilks, D. S., 2011a: *Forecast Verification*, Vol. 100. 301–394 pp.

———, 2011b: *Statistical Forecasting*, Vol. 100. 215–300 pp.

Wobus, R. and E. Kalnay, 1995: Three Years of Operational Prediction of Forecast Skill at NMC. *Monthly Weather Review*, **123**, 2132–2148.

Yoo, C. and S. W. Son, 2016: Modulation of the boreal wintertime Madden-Julian oscillation by the stratospheric quasi-biennial oscillation. *Geophysical Research Letters*, **43 (3)**, 1392–1398.

Zhang, F., Y. Qiang Sun, L. Magnusson, R. Buizza, S. J. Lin, J. H. Chen, and K. Emanuel, 2019: What is the predictability limit of midlatitude weather? *Journal of the Atmospheric Sciences*, **76 (4)**, 1077–1091.

Glossary

ACC Anomaly Correlation Coefficient.

AO Arctic Oscillation.

AR Atlantic Ridge.

AT Atlantic Trough.

BL European Blocking.

CNN Convolutional Neural Network.

DJF Winter, December–February.

ECMWF European Centre for Medium-Range Weather Forecast.

ENSO El Niño Southern Oscillation.

EOF Empirical Orthogonal Function.

EuBL European Blocking.

FCNN Fully Connected Neural Network.

GL Greenland Blocking.

Grad-CAM Gradient-Weighted Class Activation Mapping.

IFS Integrated Forecast System.

LSTM Long Short-Term Memory.

MJO Madden-Julian Oscillation.

ML Machine learning.

NAO North-Atlantic Oscillation.

NDJFM Extended Winter, November–March.

NN Neural Network.

NWP Numerical Weather Prediction.

ONI Oscillation Niño Index.

PNA Pacific-North American pattern.

QBO Quasi-Biennial Oscillation.

RFI Relative Feature Importance.

RMM Real-time Multivariate MJO.

RMSE Root Mean Squared Error.

RPS Ranked Probability Score.

RPSS Ranked Probability Skill Score.

S2S Subseasonal to Seasonal.

ScBL Scandinavian Blocking.

ScTr Scandinavian Trough.

SPV Stratospheric Polar Vortex.

WCB Warm Conveyor Belt.

Z500 500 hPa Geopotential Height Field.

ZO Zonal Regime.

Acknowledgement

First and foremost, I have to thank my research supervisors, Jun.-Prof. Dr. Christian Grams, Prof. Dr. Peter Knippertz, Dr. Julian Quinting and Dr. Sebastian Lerch. Without your assistance and dedicated involvement in every step throughout the process, this Master's thesis would have never been accomplished.

Julian, I am deeply grateful for your support. You were always available for me and helped me with your impressive meteorological knowledge, but also the knowledge about machine learning. Thank you for your countless suggestions and ideas during our discussions and thank you for proofreading my work.

Christian, I am deeply grateful to you for giving me the chance to work on this exciting topic and for supporting me in expanding my knowledge in your research field. Thank you for letting me be part of your working group „Large-scale Dynamics and Predictability". Thank you also for all the constructive feedback during our discussions and proofreading.

Sebastian, you laid the foundation for my interest and knowledge of machine learning with your lecture „Methods of Data Analysis". I would also like to thank you for the numerous discussions, which gave me excellent thought-provoking impulses.

Peter, your knowledge of tropical meteorology is impressive. With your lecture „Tropical Meteorology" you gave me a great insight into the topic of climate modes and thus made the start of my Master's thesis much easier. Thank you for your constructive feedback in our discussions with approaches that I had not thought about before.

I would also like to thank all colleagues from the young investigator group „Large-scale Dynamics and Predictability" at IMK-TRO. The daily conversations in the online coffee breaks brought a bit of normality to my working day at the home office and were always an event I could look forward to during the day. You integrated me into your work group, for which I am very grateful.

Finally, I would like to express my very profound gratitude to my parents and brother for your unwavering support and encouragement throughout my studies and during the research and writing of this thesis. Mum and Dad, you took a lot of the burden off my shoulders so that I could focus on my Master's thesis. My friends should also not go unmentioned, you have always reminded me that there is also a life besides research. My achievements in my studies and in my Master's thesis would not have been possible without all of you.

Erklärung

Ich versichere wahrheitsgemäß, die Arbeit selbstständig angefertigt, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten anderer unverändert oder mit Abänderungen entnommen wurde.
Karlsruhe, den 01.04.2022

Fabian Mockert