universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

"A deep convolutional RNN model for spatio-temporal prediction of wind speed extremes in the short-to-medium range for wind energy applications"

verfasst von / submitted by

Daan Scheepens, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2021 / Vienna, 2021

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 910
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Computational Science
Betreut von / Supervisor:	Univ.-Prof. Dipl.-Inform.Univ. Dr. Claudia Plant

Zusammenfassung

Die Menge der Windfarms und der Windstromproduktion in Europa, sowohl on- als auch off-shore, hat in den letzten Jahren stark zugenommen. Um Netzstabilität sicherzustellen, die Gebühren im Energiehandel zu senken und die rechtzeitige Planung von Wartungsarbeiten zu gewährleisten, sind präzise Vorhersagen von Windgeschwindigkeit und Windleistung erforderlich. Es ist besonders wichtig geworden, die Vorhersagen der Windgeschwindigkeit im kurzen Bereich von ein bis sechs Stunden zu verbessern, da die Windgeschwindigkeitsvariabilität in diesem Bereich die größte betriebliche Herausforderung darstellt. Darüber hinaus sind genaue Vorhersagen extremer Windereignisse für Windfarmbetreiber von großer Bedeutung, da diese durch rechtzeitige Kenntnis Schäden vorbeugen und wirtschaftliche Vorsorge betreiben können. Wir stellen in dieser Arbeit ein auf Deep Convolutional Recurrent Neural Network (RNN) basierendes Regressionsmodell zur raumzeitlichen Vorhersage von extremen Windgeschwindigkeitsereignissen im kurzen bis mittleren Bereich vor. Das wird durch die Kombination eines mehrschichtigen Convolutional Long Short-Term Memory (ConvLSTM)-Netzwerks mit Imbalanced Regression Loss erreicht, das Windgeschwindigkeitsvorhersagen in Europa mit einer Vorlaufzeit von 12 Stunden in 1-Stunden-Intervallen liefert. Wir untersuchen die Prognoseleistung bei hochschwelligen Extremereignissen von drei verschiedenen Imbalanced Regression Loss Functionen und vergleichen sie mit den häufig verwendeten Mean Squared Error (MSE) und dem Mean Absolute Error (MAE) Loss und untersuchen Netzwerkstrukturen mit einer unterschiedlichen Anzahl an Schichten. Zusätzlich zu dem Beitrag, den diese Arbeit zu Deep Learning für Windenergieanwendungen und Imbalanced Spatio-temporal Regression bietet, tragen wir zum Bereich der Vorhersageverifizierung bei, indem wir die Leistung der Vorhersage mit einer umfassenden raum-zeitlichen Verifizierungsstrategie bewerten. Die Ergebnisse indizieren, dass der invers gewichtete mittlere absolute Fehler (W-MAE) für die Vorhersage der in dieser Arbeit untersuchten extremen Windgeschwindigkeitsereignisse am besten geeignet ist und sowohl den standard MSE als auch den MAE Loss deutlich überholt. Zudem bieten wir mit einer umfassenden Verifikation des W-MAE-trainierten Modells Einblicke in die maximale Vorlaufzeit sowie die minimale räumliche Auflösung, die eine ausreichende Vorhersagefähigkeit dieses Modells für Extremereignisse unterschiedlicher Schwellenwerte garantiert. Diese Masterarbeit wurde im Rahmen des MEDEA-Projektes verfasst, welches vom Österreichischen Klimaforschungsprogramm zur weiteren Erforschung erneuerbarer Energien und meteorologisch induzierter Extremereignisse finanziert wird.

Abstract

The amount of wind farms and wind power production in Europe, both on- and off-shore, increased rapidly in the past years. To ensure grid stability, on-time (re)scheduling of maintenance tasks and mitigate fees in energy trading, accurate predictions of wind speed and wind power are needed. It has become particularly important to improve wind speed predictions in the short range of one to six hours as wind speed variability in this range has been found to pose the largest operational challenges. Furthermore, accurate predictions of extreme wind events are of high importance to wind farm operators as timely knowledge of these can both prevent damages and offer economic preparedness. We propose in this work a deep convolutional recurrent neural network (RNN) based regression model for the spatio-temporal prediction of extreme wind speed events in the short-to-medium range. This is achieved by combining a multi-layered convolutional long short-term memory (ConvLSTM) network with imbalanced regression loss, providing wind speed forecasts over Europe with a 12 hour lead-time in 1 hour intervals. We investigate forecast performance on high-threshold extreme events of three different imbalanced loss functions and compare them to the commonly used mean squared error (MSE) and mean absolute error (MAE) loss, as well as investigating network structures with various numbers of layers. In addition to the contributions that this work offers to deep learning for wind energy applications and imbalanced spatio-temporal regression, we contribute to the area of forecast verification by assessing forecast performance with an extensive spatio-temporal verification strategy. The results indicate that the inversely weighted mean absolute error (W-MAE) is best suited for the forecasting of the extreme wind speed events investigated in this work and significantly outperforms both the standard MSE and MAE loss. Furthermore, with an extensive verification of the W-MAE-trained model we offer insight into the maximum lead-time, as well as minimum spatial scale, that warrant sufficient forecasting skill of this model for extreme events of varying thresholds. This Master's thesis was written as a part of the MEDEA project, which is funded by the Austrian Climate Research Program to further research on renewable energy and meteorologically induced extreme events.

Acknowledgements

I want to thank my supervisor Claudia Plant and co-supervisors Katerina Schindlerova and Irene Schicker for helping me develop and materialise the ideas that gave rise to this thesis: Thank you for giving me the opportunity to explore this exciting application of deep learning to meteorologically induced extreme events and renewable energy research. A particular shout-out goes out to Katerina for our countless hours of brain-storming together, for pushing me to develop on my own ideas and for encouraging my optimism. Thank you, Irene, for helping me make sense of all things meteorological, and, Claudia, for ensuring that my data science was always sound. Many thanks, moreover, to everyone from the MEDEA project: I enjoyed contributing to the project with my thesis and I am grateful for being given the opportunity to present my work at the 2021 Annual Meeting of the European Meteorological Society. Lastly, I want to thank my family, friends and my partner Alina for supporting me through this journey.

Contents

Chapter 1

Introduction

Global warming through increasing concentrations of greenhouse gases (GHG) in the earth's atmosphere demands ever more urgently that electricity generation is shifted away from fossil fuels and towards renewable energy sources. The latest report of the Intergovernmental Panel on Climate Change (IPCC) warns that "global warming of 1.5°C and 2°C [compared to pre-industrial levels] will be exceeded during the 21st century unless deep reductions in carbon dioxide and other greenhouse gas emissions occur in the coming decades" (IPCC, 2021). The consequences of global warming by as little as 1.5°C are likely to be disastrous for many regions of the world, exposing millions of people to high risks of displacement due to rising sea-levels (IPCC, 2012; Nicholls et al., 2011), increased frequencies of extreme temperatures, precipitation, droughts, stronger tropical cyclones and increased biodiversity loss and ecosystem collapse (IPCC, 2018). We furthermore risk crossing irreversible 'tipping points', relating to the Greenland and West Antarctic ice sheets, the Atlantic Meridional Overturning Circulation (AMOC) and the Amazon rainforest. As "initiators of tipping cascades" (Wunderling et al., 2021) the ice sheets hold critical importance to stability of the global climate system and their tipping points could already be triggered at 1.5°C warming (IPCC, 2018). The mass loss of the Greenland and Antarctic ice sheets in combination with continuing deep ocean heat uptake will mean that the sea level will continue to rise for centuries beyond 2100 and remain elevated for thousands of years (Oppenheimer et al., 2019). Unsurprisingly, the economic cost implicated by climate change is significant. Different studies have estimated the delayed economic cost for each year of inaction at around 0.5 trillion dollars globally, if a 2°C target is to be met (see: e.g. Keller et al., 2007; Sanderson and O'Neill, 2020). Such estimates are, however, invariably lower-bound, since, as the IPCC (2012) succinctly states, "many impacts such as loss of human lives, cultural heritage, and ecosystem services, are difficult to value and monetize."

While global demands for oil, coal and natural gas are not yet showing signs of decreasing, more than half of the increase in global electricity supply in 2021 are set to be provided by renewables (IEA, 2021). The share of renewables in global electricity generation has been steadily rising over the past decades and is projected to reach close to 30% in 2021 (IEA, 2021) with wind providing the largest contribution to renewables growth, followed closely by solar photovoltaic (PV) (BP, 2021). Over the decade 2010-2020 wind power capacity grew 4-fold, to 773 GW, and solar power 17-fold, to 714 GW (IRENE, 2021). Interestingly, while oil, coal and natural gas demands all fell to multi-year lows during the first year of the COVID pandemic (2019-2020), renewable energy supply chains displayed remarkable resilience and experienced record growth in deployment (BP, 2021; IEA, 2021; IRENE, 2021). Invariably, the costs of renewables have steadily been declining and are projected to continue to do so (IPCC, 2011). Already renewables are economically competitive with fossil fuels. The IRENE (2021) estimates that globally 32 billion USD/year in annual savings and almost 3000 Mt CO2/year in emission reductions can be achieved by replacing uneconomic existing coal-fired power plants with wind and solar. As for emission reductions, the IRENE (2019) states that "renewables and efficiency measures, boosted by substantial electrification, can provide over 90% of necessary CO2 emission reductions [to keep global temperature rise to within 1.5°C] by 2050."[1] It is evident that the potential of renewables in reducing GHG emissions and thereby mitigating the climate crisis is enormous.

Possessing the largest market share among all of the renewables, wind energy has managed to establish itself as a mature, reliable and efficient technology for electricity production and is expected to maintain rapid growth in the coming years (Fyrippis et al., 2010; Huang et al., 2015). The Wiser et al. (2011) report that wind power capacity could grow from meeting roughly 1.8% of global electricity demand in 2009 to in excess of 20% in 2050, the reasons for this being primarily the continued advancements in on- and offshore wind energy technology and the associated continued reduction in the cost of wind energy. The decade following this report, however, has seen such rapid growth in research and development in off-shore wind technologies that 20% of global energy demand may already be provided as soon as 2030 by utilising the massive potential of floating off-shore wind farms (Darwish and Al-Dabbagh, 2020). While electricity costs[2] of off-shore wind are currently still in the upper fossil fuel cost range, they are projected to decline rapidly during the coming decade and compete in the lower range by 2030 (IRENE, 2019). On-shore wind power, on the other hand, is currently the most cost-competitive renewable resource with costs already competing in the lower end of the fossil fuel cost range (IRENE, 2019). By 2030 costs are expected to fall well below the lower fossil fuel cost range and thereby become "fully competitive" (IRENE, 2019).

One of the main challenges to the deployment of wind energy, however, seems to be concerns about its inherent variability and the accordingly lower levels of predictability than are common for other types of power plants (Wiser et al., 2011). The atmosphere is essentially a chaotic system (Visconti, 2001), the evolution of which may transpire highly non-linearly and unpredictably, making it difficult to model and predict accurately. The literature on wind speed forecasting is abundant in stating the difficulty of modelling and predicting wind speed due to its stochastic nature and the challenge this poses to reliable wind power generation (e.g. Chen and Yu, 2014; Lei et al., 2009; Li et al., 2018). Hybrid electric systems that incorporate a substantial amount of wind power therefore require some degree of flexibility from other generators in the system in order to maintain the right supply/demand balance (Wiser et al., 2011). Failing to manage this variability leads to scheduling errors which impact grid reliability and market-based ancillary service costs[3] (Kavasseri and Seetharaman, 2009), while potentially causing energy transportation issues in the distribution network (Salcedo-Sanz et al., 2009) and increased risks of power cuts (Li et al., 2018). This is where wind speed forecasting can play a significant role. Incorporating high-quality wind speed forecasts,

[1] Here "energy efficiency" refers to efficiency measures deployed in end-use applications in the industry, buildings and transport sectors (e.g., improving insulation of buildings or installing more-efficient appliances and equipment) and "electrification" denotes electrification of heat and transport applications, such as deploying heat pumps and electric vehicles (IRENE, 2019).

[2] In terms of global levelized cost of electricity (LCOE).

[3] Maintenance costs relating to controlling voltage and frequency in the grid.

and in return, wind power forecasts, into electric system operations gives the system more time to prepare for large fluctuations and can thereby help mitigate these issues (Wiser et al., 2011). The variability in the short-range, particularly over the time scale of *one to six hours* is found to pose the most significant operational challenges (Li et al., 2018; Wiser et al., 2011). The development of accurate wind speed forecasts in the short-range has therefore become increasingly important. Despite the challenge of wind power variability, however, many countries have already clearly demonstrated that hybrid systems with large contributions of wind energy can operate reliably. As early as 2010, Denmark, Portugal, Spain and Ireland managed to supply between 10 and 20% of annual electricity demand with wind energy (Wiser et al., 2011); and the number have only risen since. Denmark is a clear world-leader in this respect, boasting wind penetration[4] of 58% in 2020 (REN21, 2021). The Wiser et al. (2011) conclude that managing the variability of wind power generation poses "no insurmountable technical barriers" and reiterate that "increased utilisation of wind energy offers the potential for significant near-term GHG emission reductions".

Short-term wind speed prediction is not just a key element in the successful management of hybrid electric power systems, it is also vital for improving the life-span of wind turbines and planning for necessary shut-downs in the face of extreme weather (Chen and Yu, 2014). Most existing turbines stop producing energy when either instantaneous gust speeds or averaged wind speeds exceed a threshold called the 'cut-out speed' which is typically set around 25 m/s, after which, either through changing the shape of the blades (stall control) or the angle of the blades (pitch control), or a combination of both, the rotation of the blades is brought to a halt and the turbine is essentially turned off (Burton et al., 2001). This is called 'sharp cut-out'. This is done to avoid excessive loads on the turbine components that could lead to wind turbine damage or failure (Jelavić et al., 2013). The turbine starts operating again once the wind speed has decreased by several m/s, typically to 20 m/s (Castellani et al., 2019). This process is called hysteresis and is used to prevent frequent restarts and shutdowns (Cutululis et al., 2012). Nevertheless, hysteresis is known to be a highly sub-optimal strategy, leading to significant loss of energy production (Horváth et al., 2007) and poor predictability of power output (Bossanyi and King, 2012). Most importantly, however, the sharp cut-out strategy risks sudden shutdown of the entire wind farm, threatening the security and stability of the entire grid unless energy reserves are on standby to take over under very short notice (Cutululis et al., 2012). Using simulations of off-shore wind power in Denmark, Cutululis et al. (2012) found that loss of wind power production during critical weather conditions can reach up to 70% of installed capacity within an hour. Accurate forecasts of extreme wind events can therefore provide vital foresight to help prepare the electrical grid for such shutdowns as well as the duration of their downtime (Petrović and Bottasso, 2014). Over the course of the last decade, however, other strategies have been put forward to mitigate these issues by allowing wind turbines to continue operating during storms. These are called 'soft cut-out' strategies. The standard soft cut-out strategy, as described by Markou and Larsen (2009) and Bossanyi and King (2012), entails bringing the wind turbine to a gradual halt at a higher cut-out speed, by carefully reducing the power output and rotor speed according to an optimised, predefined ramp shape. While this circumvents hysteresis and thereby improves predictability and annual energy output, it does subject the turbine to increased loads (Bossanyi and King, 2012). Improvements to this strategy are offered, for example, by Jelavić et al. (2013), Petrović and Bottasso (2014), and Petrović and Bottasso (2017) where wind

[4] The percentage of gross annual electricity demand covered by wind power generation.

speed and wind turbines states are monitored in real-time and power output is only reduced when there is a real risk of excessive loads. Both of these methods utilise expected wind speed variations in their optimisation procedures and can thus profit from accurate wind speed forecasts. Thus, improving forecasting capability of extreme winds speeds can help increase grid stability, optimise wind power output, as well as mitigate loads on turbine components and increase the overall wind turbine life-span.

The prediction of extreme winds, however, poses a considerable challenge to computer science research. The distribution of wind speed is typically modelled according to a Weibull distribution, which is a heavy-tailed, type III extreme value distribution (Wilks, 2019, p. 115). While the Weibull distribution can be used within the framework of extreme value theory (EVT) for the statistical study of wind speed maxima or minima (see: e.g. Ghil et al., 2011), heavy-tailed distributions pose a serious problem to data-driven prediction of extreme events e.g. through deep learning methods, where the severe data imbalance of extremes versus non-extremes can significantly impede model-learning. While classification tasks can benefit from a wide range of resampling strategies (e.g. Oliveira et al., 2021) to combat data imbalance, imbalanced regression tasks, as investigated in this thesis, must seek other remedies to this problem, such as the usage of particular loss functions. The subject of imbalanced regression is discussed in further detail in Section 2.1.3 in the next chapter.

Chapter 2

State of the Art

This chapter provides an extensive literature review of the state of the art of forecasting methods and forecast verification relevant to the problem of spatio-temporal forecasting of wind speed extremes. Section 2.1 reviews the literature on wind speed forecasting methods, which can broadly be divided into *physical* and *statistical* methods, where particular attention is granted to the deep learning methods for spatio-temporal forecasting of extreme events. Section 2.2 provides a review of the subject of forecast verification of spatial fields and extreme events.

2.1 Forecasting Methods

Due to the growing utilisation of wind power as a renewable energy resource, a large amount of research has flown into the development of new and improved methods for reliable forecasting of wind speed and wind power. Many approaches have been proposed over the past two decades, spanning a wide range of models, temporal horizons as well as pre- and post-processing techniques. Following the dichotomy made by e.g. Costa et al. (2008), Lei et al. (2009) and Jung and Broadwater (2014), which provide excellent overviews of the subject, the approaches for wind speed prediction can broadly be divided into two categories based on whether or not physical information about the terrain is used as input. These categories are termed the *physical model based methods* and the *statistical modelling methods*, respectively. While this review focuses on forecasting approaches of wind speed, approaches targeting other atmospheric variables or objects will be mentioned in certain cases where applicable.

2.1.1 Physical Methods

The physical model based methods used for wind speed and power prediction utilise a detailed physical description at the target forecast location (e.g. the wind farm) (Lange and Focken, 2006). This typically includes a description of the terrain through its orography, roughness, obstacles, etc. but may also include a description of the wind farm such as its layout and its site-specific wind turbine power curve (Jung and Broadwater, 2014). This description is then used in combination with large-scale numerical weather prediction (NWP) data of the atmosphere to solve a system of physical conservation equations at the target site in order to obtain the predicted wind speeds. As the name suggests, a NWP model predicts the future states of the atmosphere by numerically solving large systems of differential equations from thermodynamics and geophysical fluid dynamics (Cheng et al., 2017). Typically a NWP system consists of an ensemble of such models, where each model computes the solution based on slightly perturbed initial conditions in order to obtain a degree of uncertainty in the predictions.

A number of methods have been proposed in this field. Alessandrini et al. (2013) and Deppe et al. (2013) used physical models with ensemble NWP data for improved long-range wind speed forecasts out to 72 hours and 48 hours, respectively. Kikuchi et al., 2017 proposed an algorithm for assimilating NWP data with flight data using sequential importance sampling (SIS) as a wind speed nowcasting model, while Cheng et al., 2017 reported improved wind speed forecasts by assimilating NWP data with wind turbine anemometer observations. The Risoe National Laboratory in Denmark developed a physical model called *Prediktor* (see: e.g. Landberg, 2001) which takes local conditions into account by using high resolution NWP forecast data. A similar physical approach is taken by the *Previento* (see: e.g. Focken et al., 2001) model of the University of Oldenburg in Germany which uses NWP forecasts from the German Weather Service.

The advantage of the physical approach is that it does not require any training of models based on historical data. Rather, the approaches are based on the solving of equations based on NWP forecasts and physical data at the target site. The acquisition of this site-specific physical data is, however, one of the main drawbacks of the approach (Jung and Broadwater, 2014). Furthermore, the predictive capability of NWP models degrades significantly for particularly stochastic variables like wind (Chen and Yu, 2014). Although physical models are nevertheless essential for wind forecasts in the short- and very-short-term, where the influence of atmospheric dynamics becomes important, they are typically outperformed by statistical methods in the mid- to long-term (Ferreira et al., 2019; Jung and Broadwater, 2014). In practice, however, physical approaches are often combined with statistical post-processing methods into so-called hybrid physical-statistical methods in order to utilise the advantages of both methods while mitigating the restrictions of NWP models (Chen and Yu, 2014). For examples of physical-statistical hybrids see e.g. Scheuerer and Hamill (2015), Dabernig et al. (2017) or Cheng et al. (2017) and references therein.

2.1.2 Statistical Methods

As opposed to the physical model based methods, statistical forecasting methods are based solely on historical weather data with no knowledge of the physical mechanisms underlying the wind speed patterns (Chen and Yu, 2014). As such, statistical methods may also be described as 'data-driven methods'. This class of methods encompasses a large variety of approaches. Most of the literature distinguishes here the conventional time-series modelling approaches from the artificial neural network (ANN) based and deep learning approaches. The latter two, along with a number of other machine learning approaches have also been termed 'soft computing methods' (Cheng et al., 2018).

Time-Series Modelling Methods

Among the time-series modelling methods, autoregressive moving average (ARMA) models have been utilised excessively for wind speed and power forecasting purposes (Chen and Yu, 2014). A variety of

ARMA models have been developed, such as the simplified autoregressive (AR), the autoregressive integrated moving average (ARIMA) or the fractional-ARIMA with a fractional differencing parameter (see: e.g. Erdem and Shi, 2011; Kavasseri and Seetharaman, 2009; Riahy and Abedi, 2008). AR and ARIMA can also be used in combination with explanatory variables (exogenous input), called ARX and ARIMAX. For example, Zhao et al. (2016a) investigated a hybrid approach of a ARIMAX model optimised with chaotic particle swarm optimisation (CPSO) for wind speed prediction over Northern China. While these models have received a substantial amount of attention for wind speed forecasting, the models are essentially linear and may therefore not be well suited for modelling the highly stochastic, non-linear dynamics of wind speed (Chen and Yu, 2014).

Artificial Neural Network Methods

Alternatively, there has been a strong trend towards ANN-based approaches. In fact, the ANN is one of the most widely used statistical models for wind speed and power forecasts (Jung and Broadwater, 2014), and renewable energy forecasting in general (Leva et al., 2017). An ANN in its most general sense is a network of interconnected nodes, typically arranged as an input layer, a number of 'hidden layers' and an output layer. The connections between nodes are given weights that determine how strongly one node's activation is received by another node. In *feedforward* ANNs information flows entirely in one direction from input through the hidden layers to the output, while *recurrent* ANNs allow for cyclical connections (Zell, 1994, p. 73). With each new training sample, these weights are adjusted through some form of gradient descent method using a particular learning algorithm such as backpropagation so as to minimise the model's 'loss function', which is a function of the errors between the model output and the target. The weights of the network thus span the learnable parameters of the model and over the course of a training phase the model 'learns' to provide the correct output by adjusting its weights so as to minimise the given loss function. The power of ANNs lies in its ability to model highly complex and non-linear relationships between input and output while the risk of overfitting can be controlled fairly well. In fact, feedforward ANNs with as few as a single hidden layer are so-called *universal approximators*, capable of approximating any continuous function on a compact interval to arbitrary precision, provided sufficiently many hidden nodes are available (Hornik et al., 1989). A more contemporary discussion on this subject is provided by Lu et al. (2017). Furthermore, ANNs require no prior assumption of any mathematical relationship between input and output, as opposed to the previously reviewed physical and statistical approaches (Jung and Broadwater, 2014).

Many different ANN architectures have been proposed for the forecasting of atmospheric variables. We note here that most of the methods mentioned below are multi-layered ANN models and can thus also be considered deep learning (DL) methods, where DL simply refers to ANN models with one or more hidden layers. Li and Shi (2010) compared three types of ANNs, based on adaptive linear elements, backpropagation (BP-NN) and radial basis functions (RBF-NN) respectively, for 1-hour-ahead wind speed forecasting. A similar comparison was carried out by Noorollahi et al. (2016), where the forecasting abilities of a BP-NN, RBF-NN and an adaptive neural fuzzy inference system (ANFIS) were compared for 1-hour-ahead wind speed forecasts over Iran. Liu et al. (2015) and Cheng et al. (2018) provide further applications of ANFIS for wind speed prediction. The method proposed by Cheng et al. (2018) is a hybrid ensemble model utilising ANFIS, long short-term memory (LSTM) and gated recurrent units (GRU) models. The latter two are types of recurrent neural networks

(RNN) known for their effectiveness in modelling time series. Further examples of LSTM applications for wind speed forecasting include the work of Liu et al. (2018a), Liu et al. (2018b), Marndi et al. (2020) and Xie et al. (2021). Extreme Learning Machines (ELM) have also been applied; see e.g. Wan et al. (2014), Zhao et al. (2016b) or Li et al. (2018). ELM models can learn thousands of times faster than conventional ANNs trained with backpropagation while retaining good generalisation abilities (Huang et al., 2006). Blonbou (2011) proposed an ANN in combination with adaptive Bayesian learning and Gaussian process approximation to allow probabilistic wind power prediction in the very short-term. Pourmousavi Kani and Ardehali (2011) investigated potential improvement in the very short-term with an ANN and Markov chain hybrid model. Lee and Baldick (2014) reported high precision of short-term wind power forecasts using an ensemble of 52 ANNs and 5 Gaussian process models. Scher and Messori (2021), furthermore, proposed four methods to generate probabilistic ensemble forecasts from a feedforward ANN, namely random initial perturbations, retraining of the neural network, random dropout in the network and the creation of perturbations with singular vector decomposition. ANNs can also be used to model the probability density function of wind speed rather than the wind speed itself, as in Celik and Kolhe (2013), or used for binary classification of damaging wind events, as in Marzban and Stumpf (1998). Lastly, a novel family of ANNs has recently been proposed by Chen et al. (2018) for time-series modelling, supervised learning and density estimation termed neural ordinary differential equation solvers; to our knowledge this method has not yet been applied to wind speed prediction.

Other statistical methods have been investigated for wind speed prediction that are not ANN-based but do fit into the 'soft computing' category as general machine learning approaches. These include methods utilising support vector machines (SVM) (e.g. Chen and Yu, 2014; Jiang et al., 2017; Yang et al., 2015), Markov chain models (e.g. Carpinone et al., 2015) or regression trees (e.g. McGovern et al., 2017). Wang et al. (2017a) provide a review of multi-step ahead soft-computing models for wind speed forecasting; the authors argue that the multi-step forecasting approach is to be preferred over single-step forecasting. Hybrids of time-series and soft-computing methods have also been proposed. For example, an ARMA-ELM hybrid for wind speed forecasting was proposed by Mi et al. (2017), while do Nascimento Camelo et al. (2018) found that their ARIMA-ANN hybrid performed better than ARIMA or ANN by itself, for wind speed predictions over Northeastern Brazil.

2.1.3 Deep Learning Methods for Spatio-Temporal Forecasting

As a special case of the soft-computing statistical methods, deep learning (DL) methods for spatio-temporal forecasting are of particular relevance to the thesis at hand. The term *deep learning* simply refers to the area of machine learning concerned with multi-layered i.e. 'deep' artificial neural networks. Beyond being universal approximators (Lu et al., 2017), deep ANNs are capable of automatically and effectively learning hierarchical feature representations from the raw input data, with different layers essentially learning to detect different features in the data. This is different from other physical and statistical approaches, where features are first hand-crafted from the data and then fed into the model (Wang et al., 2020). These two qualities have made deep learning models very attractive to the general area of spatio-temporal sequence forecasting (STSF), where complex spatial and temporal correlations are typically present in the data (Wang et al., 2020). With the utilisation of multi-layered structures of both convolutional neural networks (CNN) and recurrent neural network (RNN) such correlations can be automatically and effectively learned directly from the data (Wang et al., 2020). Good review papers of deep learning applications to STSF can be found in Shi and Yeung (2018), Amato et al. (2020) and Wang et al. (2020). Shi and Yeung (2018) posits to divide deep learning for STSF into two major categories: *deep temporal generative models* on the one hand and *feedforward neural network* and *recurrent neural network* based models on the other. The former category includes methods such as the temporal restricted Boltzmann machine and its variants (e.g. Mittelman et al., 2014; Sutskever and Hinton, 2007; Sutskever et al., 2009) and the temporal sigmoid belief network (Gan et al., 2015). We will focus on the latter category here. Zhu et al. (2018) proposed a DL model for spatio-temporal wind speed prediction where spatial correlations in the data are captured with a CNN and temporal correlations are captured with a multi-layered perceptron (MLP). Instead of using a CNN to capture spatial correlations only, 3D CNN models may be used to perform convolution over both spatial and temporal domains using spatio-temporal filters. Vondrick et al. (2016) applied this approach to video frame prediction and Shi et al. (2017) demonstrated its superior performance over a 2D CNN model for precipitation nowcasting. Another way to capture temporal correlations in the data is by using RNNs. Srivastava et al. (2015) proposed the usage of a multi-layered, fully connected LSTM (FC-LSTM) for video frame prediction by flattening the input images directly into arrays to be used by the LSTM network. Oh et al. (2015) instead used 2D CNNs to encode the input frames before feeding them into the LSTM network. Shi et al. (2015) improved upon these methods with the proposed convolutional LSTM (ConvLSTM) network, which the authors applied to precipitation nowcasting. A similar extension to the GRU, the ConvGRU, has been made by Shi et al. (2017). Such ConvRNN networks combine the advantage of CNN and RNN and have been used extensively in the literature as building blocks for DL models for STSF tasks (see: e.g. Shi and Yeung, 2018 and references therein). Improvements to the ConvRNN network structure have been put forward, however. Shi et al. (2017) introduced the Trajectory GRU (TrajGRU) as an improvement to the ConvGRU where the recurrent connection structure is actively learned, while Wang et al. (2017b) proposed the Predictive RNN (PredRNN) as an improvement to the ConvLSTM network by maintaining a global memory state rather than constraining memory states to each ConvLSTM module individually. PredRNN++ was later proposed by Wang et al. (2018), where more nonlinearities were added to the updating process of the global memory state. The authors demonstrate PredRNN++ to be superior to TrajGRU and ConvLSTM for video frame prediction. A different approach was taken by Rao et al. (2020), where two novel spatio-temporal DL methods are proposed based on Functional Neural Networks (FNN) as improvements to the ConvRNN approaches. Another alternative to spatio-temporal modelling using DL is with the help of Generative Adversarial Networks (GAN). A thorough review of the usage of GANs for spatio-temporal data is provided by Gao et al. (2020).

Extreme Events

Deep learning for spatio-temporal forecasting has also been used in the context of extreme weather forecasting. Liu et al. (2016) developed a multi-channel CNN model (taking in multiple atmospheric variables) to classify images of extreme weather events such as tropical cyclones, atmospheric rivers and weather fronts. Racah et al. (2017) followed a similar approach but utilised a multi-channel 3D (spatio-temporal) CNN architecture to classify extreme weather events spatially as well as temporally. The authors used a single network to classify all events, as opposed to Liu et al. (2016) where separate networks were used to classify each of the individual events. Racah et al. (2017) furthermore presented the ExtremeWeather dataset, to encourage further machine learning research in this area and to help mitigate the effects of climate change. Feng and Fox (2021) proposed the TSEQPredictor model for earthquake prediction over Southern California, which combines a CNN autoencoder with a temporal convolutional network (TCN) to classify the occurrences extreme earthquake events. The authors were able to improve their model by employing skip connections and local temporal attention into the network. Rather than classifying the extreme events themselves, Ahmed et al. (2021) used a CNN model to classify the 'dependence structure' of extreme temperature and precipitation events over Iraq. Yu et al. (2017), on the other hand, proposed modelling spatial extreme events by bridging a gap between traditional statistical methods and graph methods via decision trees, while Thomas et al. (2021) employed an unsupervised k-means clustering approach to investigate weather patterns responsible for extreme wind speed events throughout Mexico.

The interpretation of the term 'extreme event' in the spatio-temporal context can vary substantially, however. As we can see from the aforementioned studies, extreme events often refer to hazardous weather *patterns*, present over some spatial or spatio-temporal domain. While these events are certainly rare and extreme within the underlying climatology of the study, this is not usually the case in terms of the actual data distribution. For example, most classification studies of extreme weather patterns ensure that the model is fed an equal number of negative and positive samples, so as to avoid any severe model biases due to class imbalances. Even when class imbalances are tolerated, there are other remedies available such as resampling strategies (see: e.g. Oliveira et al., 2021 for a spatio-temporal approach) as well as one-class classification approaches (e.g. Deng et al., 2018; Goyal et al., 2020) or deep anomaly detection (e.g. Hendrycks et al., 2019). Extreme events in regression problems, on the other hand, typically refer to the tail of the data distribution i.e. those values above or below a certain threshold that are highly underrepresented in the data and thus rarely encountered during model training. As any real-world distribution is rarely uniform, such data imbalances are the rule rather than the exception. Regression problems on non-uniform i.e. imbalanced data distributions are termed *imbalanced regression* problems. Ding et al. (2019) provide a formal analysis on why deep learning regression models suffer from overfitting and underfitting problems when data is imbalanced and propose a novel loss function called the Extreme Value Loss (EVL), based on extreme value theory, which is demonstrated to improve predictions on extreme events in time-series forecasting. The authors furthermore propose a memory network based neural network architecture to memorise past extreme events for better prediction in the future. Ribeiro and Moniz (2020) addressed the problem of imbalanced regression by proposing the squared error-relevance area (SERA) loss function, based on the idea of *relevance functions*. Yang et al. (2021), on the other hand, proposed *distribution smoothing* for both the label (in the continuous sense) and feature distributions to take into account the dependencies between nearby labels, address underrepresented or even missing labels in the label distribution and reduce unexpected similarities within the feature distribution that arise due to the label imbalance. The smoothed label distribution better reflects the effective label density and can be used rather easily for re-weighting methods, where the loss function can be re-weighted by multiplying it with the inverse of the smoothed label distribution for each target. Such re-weighting of the loss function is a cost-sensitive remedy to data imbalance and has been used in the context of spatio-temporal weather forecasting, for example by Shi et al. (2017) for precipitation nowcasting.

2.2 Forecast Verification

With a forecasting model in place, a big question arises as to how best to assess the quality of its forecasts. The process through which weather forecasts are assessed to determine their quality is called *forecast verification*. What exactly constitutes 'quality' in this context is a highly non-trivial question as it depends not only on the type of forecast but also on the attributes of the forecast that are of interest to the user. For example, we could judge the merits of a forecast in respect to observation quite simply by comparing their average point-wise error and deem the minimal of the two to be preferable. There is certainly no shortage of error measures that we could compute and compare. Perhaps, however, we are looking at precipitation fields and ask ourselves which forecast predicted the occurrence of rain best: Rather than choosing the forecast with minimal overall error, we could instead look at the fraction of correctly predicted rain events vs. the fraction of falsely predicted rain events. By using a set of thresholds, we can get an idea of the accuracy of the predictions on certain value ranges. We can furthermore ask ourselves, is the forecasting ability on all value ranges equally important, or are we particularly interested in a particular range? Furthermore, is forecast skill on the smallest available scale a necessity or can the forecast still be useful on a coarser spatial (or temporal) scale where its performance is improved? How does the assessment change if the forecast is probabilistic i.e. contains a measure of uncertainty along with its predictions? These examples should give an idea of the many disparate ways that the quality of a forecast can be assessed and what constitutes a good forecast. Accordingly, this active field of research comprises a plethora of different methods and an abundant literature on applications, advantages and disadvantages etc.

Murphy (1993) defines forecast 'quality' simply as "the degree of correspondence between the forecasts and observations", and notes that it is but one type of measure of 'goodness' of a forecast, the other measures being 'consistency' (the correspondence between forecasters' judgments and their forecasts), and 'value' (the incremental economic benefits of the forecasts to users). Following this definition of quality, how do we measure the correspondence between forecasts and observations? Essentially, a forecast verification procedure attempts to capture this correspondence through the computation of a numerical statistic, or *score*. The choice of score depends first and foremost on the type of forecast. Scores that can be used to assess a forecast of a continuous variable are usually not appropriate for forecasts of categorical variables and vice versa. The same holds for forecasts that are probabilistic versus forecasts that are deterministic. In particular, ensemble forecasts, spatial forecasts, forecasts of extreme events and short- and long-range forecasts each require very different scoring frameworks. These different scoring frameworks are discussed in extensive detail in Jolliffe and Stephenson (2012) and Wilks (2019, chap. 9). Furthermore, there are a variety of ways for a forecast to go right and for a forecast to go wrong, which are each captured in different ways by different scores. There is simply no single score that can capture all the strengths and all the failings of a forecast, and collapsing any forecast verification problem into a single number thus necessarily involves a loss of information (Wilks, 2019, p. 383). A forecaster must, therefore, decide which attributes of a forecast are of particular importance, whether to the scientist, the administrator or the end-user, and utilise a set of scores that quantify these attributes (Jolliffe and Stephenson, 2012, p. 6). The general verification framework of Murphy and Winkler (1987) and Murphy (1993) can help get an idea of the many different attributes that may be relevant for a forecast, such as its accuracy, bias, reliability, resolution, discrimination or sharpness.

Rather than presenting a score in absolute terms, it is usually more informative to measure a score against some predefined baseline, or reference (Jolliffe and Stephenson, 2012, p. 6). Such scores are known as *skill scores*. The reference score is obtained from a completely naive, 'unskillful' forecast, typically represented by one of the following choices: a) the *random* forecast i.e. a forecast comprising a random sample from the variable's probability distribution, or b) the *climatology* i.e. the climatological averages of a variable over the particular spatial and temporal domains or c) the *persistence* forecast, which is simply a repetition of the observation at the present time for the duration of the forecast period. Other reference forecasts can be applicable in certain cases, however; for example, when the forecast skill of a newer forecasting system is presented relatively to an older system) (Wilks, 2019, p. 373). Skill scores ensure that we do not get fooled by fantastic scores when in actuality the forecast was entirely naive. This is a common problem when dealing with severely imbalanced categorical problems where a forecast that solely predicts the most dominant category can easily obtain better accuracy scores than the 'informed' forecast. Mathematically, for a particular verification measure A, with respect to the measure A_{ref} of the reference forecast, the skill score is given by:

$$SS_{\text{ref}} = \frac{A - A_{\text{ref}}}{A_{\text{perf}} - A_{\text{ref}}} \times 100\% \tag{2.1}$$

where A_{perf} is the value of the verification measure that would be achieved by a perfect forecast. Clearly, the skill score attains a perfect score of 100% if $A = A_{\text{perf}}$ and 0% if $A = A_{\text{ref}}$, which, as desired, indicates no improvement over the reference forecast.

Excellent overviews of the subject of forecast verification can be found in Jolliffe and Stephenson (2012) and Wilks (2019, chap. 9) and good summaries of current challenges and developments can be found in e.g. Ebert et al. (2013) and Dorninger et al. (2020). We will not attempt to reproduce such an extensive overview here but rather focus on the intersection of two areas of forecast verification that are of particular relevance to the thesis at hand, namely the area of forecast verification of spatial fields i.e. *spatial forecast verification* and the area of forecast verification of extreme events. These areas will be reviewed within a non-probabilistic framework, as the forecasts considered in this work are deterministic.

2.2.1 Spatial Forecast Verification

While forecasts can be limited to singular locations, it is increasingly more common in meteorological practice to see the forecasting of larger spatial arrays of atmospheric variables (Jolliffe and Stephenson, 2012, p. 95). Spatial forecasts are commonly produced by numerical weather prediction (NWP) models, which predict the next state of the atmosphere by solving numerical systems of differential equations on a spatial grid. These forecasts are usually non-probabilistic i.e. do not contain any expression of uncertainty in their predictions (Wilks, 2019, p. 445) (although weather forecasting systems may utilise an *ensemble* of spatial forecasting models where each model forecasts from slightly perturbed initial conditions in order to cast predictions as probabilities). We will introduce here first the traditional approaches used for the verification of spatial forecasts and then review the more modern framework that has since taken its place. For extensive, in-depth overviews of the area of spatial forecast verification we recommend Jolliffe and Stephenson (2012, chap. 6), Wilks (2019, chap. 9.8), Casati et al. (2008) and Gilleland et al. (2009, 2010b).

Traditional Approaches

The traditional approach to spatial forecast verification evaluates the forecasts simply as a collection of point-forecasts: pairs of forecast- and observation points are compared individually using any appropriate verification measure. Here we will briefly review two particular cases, namely (a) the case of *categorical* predictands (forecast variables) and (b) the case of *continuous* predictands.

a. Categorical predictands

Traditional measures for categorical variables (e.g. the occurrence of weather events such as tornadoes and hurricanes or the various categories of heavy wind or rain) usually involve the computation of a variety of scores from a so-called contingency table. For n distinct categories, a contingency table consists of the $n \times n$ frequencies of predicted vs. observed events. An example of a 2×2 contingency table, which represents a binary yes/no problem, is given in Figure 2.1. In this simple case, a denotes the hits, b denotes the false alarms, c denotes the missed hits, d denotes the correct negatives of the forecast and $n = a+b+c+d$ denotes the total. From these numbers a wide variety of scores can be calculated. Common scores include the proportion correct PC $= (a+d)/n$, odds ratio $\Theta = ad/bc$, false alarm rate F $= b/(b+d)$, hit rate H $= a/(a+c)$, false alarm ratio FAR $= b/(a+b)$, threat score TS $= a/(a+b+c)$ and the bias B $= (a+b)/(a+c)$. Each score provides different information about the forecast by quantifying either its accuracy, bias, reliability or discrimination (Wilks, 2019, p. 376). Furthermore, categorical scores are often discussed in terms of the *base rate* $s = (a+c)/n$ and the *forecast rate* $r = (a+b)/n$, which are sample estimates of the probabilities of the event occurring and being forecast, respectively.

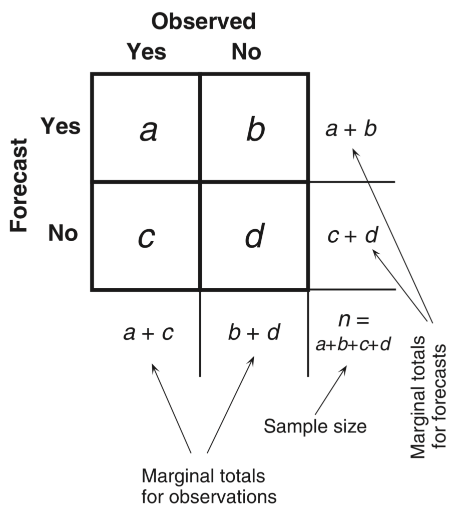

Figure 2.1: The 2×2 contingency table for binary categorical forecasts, from Wilks (2019, p. 375). **© Elsevier Books. Used with permission.**

Some caution must be taken when using the PC or Θ, as these scores can become dangerously misleading when categories are unequally represented in the data. An infamous example of this revolves around the tornado forecasts of Finley (1884). Commonly cited as the starting point of forecast verification, Finley (1884) conducted binary forecasts of tornado appearances over the United States, the results of which are shown in Table 2.1. Finley evaluated his forecasts using the PC score and obtained $PC = (28+2680)/2803 = 0.966$, whereupon he claimed an accuracy of 96.6%. What Finley missed, however, is that the contingency table is dominated by the correct negatives (d), since the actual occurrence of a tornado is exceedingly rare. The critical point is that a higher PC score can actually be obtained by simply forecasting "no tornado" on all occasions, obtaining $PC = (0+2752)/2803 = 0.982$ (as was pointed out shortly after by Gilbert, 1884). This example illustrates the caution that must be taken with categorical scores that involve the 'correct negatives'

d, as they can inflate drastically for imbalanced categories. In such cases it is more appropriate to inspect H, FAR or TS and/or to formulate the results as *skill scores* (see equation 2.1).

Table 2.1: Finley's tornado forecasts.

	Observed	
Forecast	Tornado	No Tornado
Tornado	28	72
No Tornado	23	2680
Total		2803

There are a number of skill scores that are frequently used in the forecast verification of binary contingency tables. Typically these scores are variants of the usual categorical scores, such as the PC or the TS, but measure them against a random forecast. There are a number of desirable properties that these scores should ideally have (see: e.g. Jolliffe and Stephenson, 2012, p. 46), but a central criterion is certainly the notion of *equitability*. The notion of *equitability* was proposed by Gandin and Murphy (1992) to refer to the property of skill scores to yield zero skill for random or constant forecasts and unit skill for perfect forecasts. Equitability furthermore implies that correct forecasts of rare events are weighted more than correct forecasts of common events, which prevents forecasts from attaining high scores simply by exclusively predicting the most common event (Wilks, 2019, p. 383). A distinction is made here between scores that are *asymptotically equitable* i.e. only in the limit of very large n and *properly equitable* scores that are equitable independent of n. An example of the former is the Gilbert skill score (Gilbert, 1884), while properly equitable scores include the Heidke skill score (Heidke, 1926), the Peirce skill score (Peirce, 1884) and the Clayton skill score (Clayton, 1934). The latter three scores are all properly equitable due to the fact that they are both linear and asymptotically equitable - an implication shown by Hogan et al. (2010). Unfortunately, linearity is incompatible with another desirable property, namely the *non-degeneracy* for extreme events (Jolliffe and Stephenson, 2012, p. 47). Skill scores that address non-degeneracy have been put forward, however, and will be discussed at the end of this chapter in Section 2.2.2.

While the focus of this section has been on the binary contingency table, we will briefly mention that for the forecasting of multi-categorical events Gerrity (1992) introduced a family of equitable skill scores termed the *Gandin-Murphy skill scores*. The Gandin-Murphy score utilises a particular weight for each element in the contingency table such that the equitability conditions are satisfied. When the weights are based on the sample climatology, the score is sometimes called the *Gerrity skill score* instead, and for the binary case the Gandin-Murphy skill score recovers the *Peirce skill score* (Wilks, 2019, p. 392). Interestingly, the caveat of non-equitability in the forecasting of rare events is well known in the area of data science within the task of imbalanced classification, where many of the same quantities used in this section appear, albeit under a different nomenclature. For example, the contingency table is otherwise known as the *confusion matrix* while the PC score is simply called the *accuracy*. The remedy that is typically utilised there is to compute a balanced version of the PC, where the PC of each category is weighted according to the inverse prevalence of its observation (see: e.g. Brodersen et al., 2010), which is an approach indeed very similar to the Gerrity skill score.

b. Continuous predictands

If the predictand is continuous (e.g. atmospheric variables such as temperature, pressure or wind speed) common point-wise verification measures include the correlation coefficient (r), mean error (ME), mean absolute error (MAE) and mean squared error (MSE). The correlation coefficient r ranges between -1 and 1, where a score of 1 indicates perfect positive correlation. Although this measure does reflect linear association between two variables, it is sensitive to outliers while being insensitive to biases present in the forecast (Wilks, 2019, p. 396) i.e. a forecast that severely under- or overforecast may nevertheless achieve perfect correlation with the observation. This measure must be therefore used with awareness of its rather strong limitations. The bias of a forecast can be quantified with the ME, by simply computing the difference between the average forecast and the average observation. ME$<$0 indicates forecasts that are, on average, too small while ME$>$0 indicate forecasts that are, on average, too large. Lastly, we come to the MAE and MSE, which are two very popular measures of accuracy for continuous predictands. As a matter of fact, Wilks (2019, p. 394) states that they are the only verification measures for continuous predictands in common usage in the meteorological sciences. Both measures are zero if the forecast is perfect and penalize over- and underforecasts equally. The main difference between the two is that the MSE, by squaring its errors, is more sensitive to large errors and therefore more sensitive to outliers. Interestingly, the MSE skill score against climatology can be expressed as a sum of the squared correlation coefficient and penalties relating to the calibration and the bias of the forecast (Wilks, 2019, p. 398). Thus, for the same skill score, the correlation of a badly calibrated and biased forecast must overestimate skill, which is why the correlation can be regarded as a measure of (potential) skill (Murphy and Epstein, 1989).

There are two other verification measures that must be mentioned in the domain of non-probabilistic verification measures for continuous variables: the S1 score (Teweles and Wobus, 1954) and the anomaly correlation coefficient (ACC). Jolliffe and Stephenson (2012, p. 98) explain that these two scores were specifically devised for early numerical weather prediction (NWP) forecasts of spatial fields and are still in common usage by major national weather services to track forecasting improvements over time, although they are not frequently used in research-related verification activities. The S1 score is different from most other scores used in forecast verification by the fact that it considers not the actual forecast values but rather its' gradients across the grid. The reason for its limited usage is grounded mostly in the following shortcomings: The score can be optimised by forecasting slightly larger forecast gradients than expected (Thompson and Carter, 1972), is sensitive to seasonal variations in gradients, and possesses an undesirable dependency on the domain-size where results may become unstable over smaller regions (Jolliffe and Stephenson, 2012, p. 99). The ACC measures how well the forecast captures the magnitude of anomalies from a reference forecast and can thus be interpreted as a kind of skill score. Just like the correlation coefficient, however, it is insensitive to biases and must be considered as a measure of *potential* skill, rather than actual skill (Murphy and Epstein, 1989). There are, in fact, two forms of the ACC: the *centered* form (Namias, 1952) and the *uncentered* form (Miyakoda et al., 1972). While DelSole and Shukla (2006) recommends the uncentered form for smaller spatial areas (as opposed to e.g. hemispheric domains), it appears that generally the centered form is in somewhat more common use (Jolliffe and Stephenson, 2012, p. 99). The European Centre for Medium-Range Weather Forecasts (ECMWF) employs the centered form.

A more visual verification result can be obtained by graphically presenting certain aspects of the joint distribution for a set of forecasts and observations with the *conditional quantile plot*. The starting point of the conditional quantile plot is the factorisation of the joint distribution $p(y_i, o_j)$

into the conditional distribution $p(o_j|y_i)$ and the marginal (unconditional) distribution $p(y_i)$ i.e. $p(y_i, o_j) = p(o_j|y_i) \cdot p(y_i)$. This is called the calibration-refinement factorisation (Murphy and Winkler, 1987). The conditional distribution specifies how often each possible weather observation o_j occurred given the forecast y_i i.e. how well the forecast was *calibrated*, while the marginal distribution specifies the relative frequencies of each of the forecast values y_i i.e. how *refined* the forecast was (Wilks, 2019, p. 371). The plot presents the conditional distribution in terms of selected quantiles around the 1:1 diagonal (representing perfect forecasts) and presents the marginal distribution as a histogram on the horizontal axis. Compared to basic scalar measures such as the MSE, the conditional quantile plot is a diagnostic verification technique (Wilks, 2019, p. 399) i.e. it allows diagnosis of particular strengths and weaknesses of the forecast rather than simply indicating which forecast scores better. For example, the conditional quantile plot allows for simultaneous assessment of forecast bias, sharpness, calibration and refinement. The accuracy of the forecast can furthermore be inferred from its ability to yield sharpness without sacrificing calibration (Brier, 1950; Murphy and Winkler, 1987).

Modern Framework

In the previous section we reviewed the point-to-point verification methods traditionally used to assess spatial forecasts. Immediate problems and caveats of certain methods were addressed and where appropriate the notions of *skill* and (for categorical predictands) *equitability* were provided as possible remedies. These methods are relatively easy to implement, which, as Wilks (2019, p. 455) rightly states, may seem like a welcome relief from the inherent complexity of the spatial verification problem. There is, however, an inherent problem with the point-to-point verification approach, which arises from the fact that a slight offset of an otherwise correctly predicted event incurs a 'double penalty' on the forecast: The forecast is punished once for missing to predict an event where it did occur, and another time for predicting an event where it did not occur. This also means that traditional scores evaluate a 'near-miss' and a much poorer forecast essentially equally. Furthermore, the higher the resolution of the forecast, the more notorious the 'double penalty' problem becomes, resulting in the fact that traditional scores will generally evaluate coarser forecasts better than finer forecasts, by sheer virtue of resolution (Mass et al., 2002; Rossa et al., 2008). This inherent shortcoming has motivated a great deal of research and development of diagnostic methods for spatial forecast verification (Jolliffe and Stephenson, 2012, p. 101). A large contribution to the ordering and categorising of this progress was made by the initiation of the Spatial Verification Method Intercomparison Project (ICP, see: `https://ral.ucar.edu/projects/icp/`), overviews of which are provided by Gilleland et al. (2010a) and Gilleland et al. (2009, 2010b) and Ahijevych et al. (2009).

The ICP established that spatial forecasting verification methods can broadly be categorised into four categories: (a) neighbourhood (or "fuzzy") methods, (b) scale separation (or scale decomposition) methods, (c) feature based (or object based) methods, and (d) field deformation methods. A schematic illustration of these categories is provided in Figure 2.2. The first two categories can be further generalised as "filtering" methods as both share a common strategy of applying spatial filters to either, or both, the forecast and observation fields and then compute the verification measures on the filtered fields. These methods allow us to determine the quality of the forecast at various spatial scales and determine at which scale the forecast attains a desired level of skill. The latter

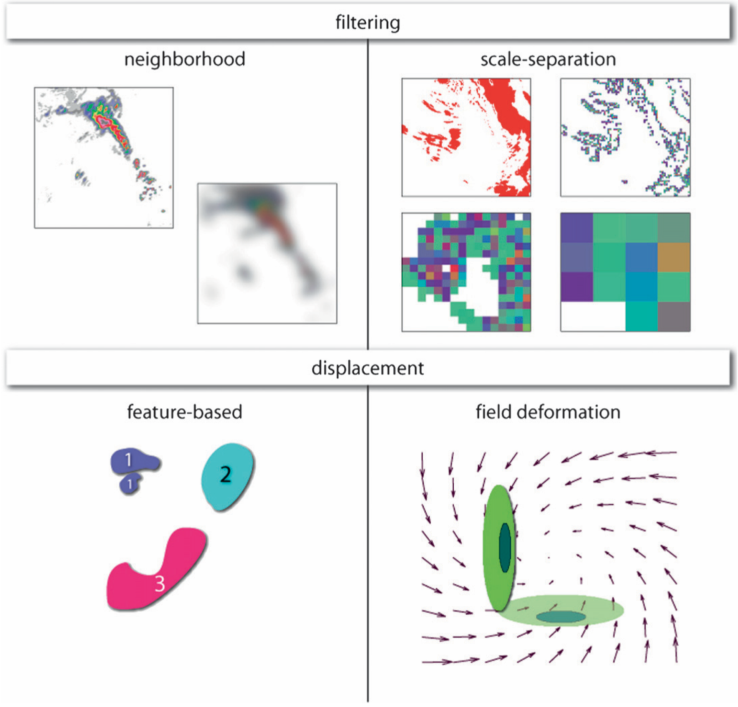

Figure 2.2: Schematic representation of the four categories of spatial verification methods, from Gilleland et al. (2010a). (top) The neighbourhood and scale-separation "filtering" methods and (bottom) the feature-based and field deformation "displacement" methods. **© American Meteorological Society. Used with permission.**

two categories share a common goal of attempting to describe the amount of spatial manipulation (displacement, rotation, scaling, etc.) necessary to transform one field into the other and can thus be generally thought of as "displacement" methods. Of course, it is possible for methods to fall into more than one category; in fact, Jolliffe and Stephenson (2012, p. 134) notes that the combining of filtering and displacement methods into hybrid schemes would be a logical next step for spatial forecast verification, and indeed such approaches have been proposed (e.g. Lack et al., 2010).

a. neighbourhood methods

neighbourhood, or "fuzzy", verification methods, allow a degree of displacement between forecast and observation based on a local *neighbourhood*. This neighbourhood is typically a square of scale s around the forecast point. Some strategies compare the forecast neighbourhood to a single observation while others compare neighbourhood to neighbourhood. Ebert (2008, 2009) provides a thorough review of the general fuzzy framework and the wide variety of methods that have been devised within this framework. The basic idea is the following: Select a set of scales and a set of intensity thresholds to distinguish events from non-events, and for each combination of scale and threshold compare some property of the neighbourhoods (e.g. the mean, maximum, number of points exceeding the threshold) to determine whether to label the neighbourhood as an event or a non-event. Due to the division between events and non-events, based on a particular intensity threshold, the verification is therefore typically done through a binary contingency table and, as such, all the traditional scores for categorical predictands can be used here. The results are typically presented in a table for each intensity threshold and for each spatial scale.

The simplest, and earliest, method is the *upscaling* method (e.g. Yates et al., 2006; Zepeda-Arce et al., 2000) where forecast and observations are consecutively averaged to coarser and coarser scales; the averages are compared to the threshold to determine events and non-events, which can then be assessed using traditional categorical scores. The main disadvantage of this method is that small-scale, high-impact weather events such as extreme wind are lost in the averaging process. Rather than averaging, the *minimum coverage* method (Damrath, 2004) classifies the neighbourhood as an event if a certain minimum fraction of the neighbourhood is covered by events. A second approach by Damrath (2004) termed *fuzzy logic* defines the fraction of events in a neighbourhood as the probability for the occurrence of the event in the neighbourhood and he probabilities themselves can be used to construct a binary contingency table (see: e.g. Ebert, 2002). The *Multi-event contingency table* (Atger, 2001) extends the contingency table to multiple intensity thresholds and allows for 'closeness' to be temporal rather than solely spatial (Ebert, 2008). The *fractions skill score* (FSS) of Roberts and Lean (2008) and Roberts (2008) compares the fractional coverage of forecast and observation events within the respective neighbourhoods directly and from these computes a variation of the fractions Brier score (FBS). An analysis of properties of the FSS is provided by Skok and Roberts (2016) while Zhao and Zhang (2018) provides a recent example of the score to precipitation forecasts. As opposed to using the traditional categorical scores for verification, the *CSRR* (Germann and Zawadzki, 2004) and the *Area-related RMSE* (Rezacova et al., 2007) provide an alternative scoring strategy where intensity is incorporated directly into the calculation of the scores (Ebert, 2008).

The neighbourhood methods are generally straightforward and they provide useful information about forecast skill at various scales and intensities. The disadvantage of the framework is that there is a limited possibility for diagnostic assessment due to the fact that the methods do not inform about specific error types and spatial structure errors (Gilleland et al., 2010a).

b. Scale-separation methods

Similarly to the neighbourhood methods, the scale-separation verification methods inform about forecast skill at various spatial scales and intensities. Scale-separation methods, however, obtain different scales by the application of a spatial scale-decomposition, e.g. by using the Fourier or wavelet transform. The neighbourhood methods essentially smooth the forecast over a range of increasing scales, which means that while small-scale features are lost in the smoothing process, the broad structure of the field is basically retained (Gilleland et al., 2009). In contrast, the scale-separation approach assesses each scale independently; individually they may not resemble the original field, but they may be combined to reproduce it again perfectly (Gilleland et al., 2009).

The first scale-separation approach, introduced by Briggs and Levine (1997), utilised 2D wavelet filters and assessed the different scales using traditional verification measures for continuous predictands such as the correlation and RMSE. The *intensity-scale* (IS) method of Casati et al. (2004) and Casati (2010) follows an approach similar to that of the neighbourhood methods by assessing the forecast at various combinations of scale and intensity-threshold. The particular threshold is applied to the forecast and observation fields to turn them into binary fields of events and non-events, from which the difference is taken to obtain the 'binary error field'. The difference field is then decomposed into different scales using 2D Haar wavelets. The different scale components are assessed using an MSE skill score. Marzban and Sandgathe (2009) investigates a somewhat different approach using *variogram* plots. The approach investigates distributional properties of the two fields rather than providing assessment on different spatial scales. The method does not fit into any of the four categories very well but since the variogram of the field is tied to the covariance of the coefficients of the wavelet decomposition, it is judged to be most like a type of scale-separation method (Gilleland et al., 2009).

Like the neighbourhood methods, the scale-separation methods can provide a useful assessment of the forecast at various scales and intensities. The methods may, however, be able to identify more meaningful features, as the chosen scales are often representative of physical features, such as large-scale frontal systems or small-scale convective showers, in the case of precipitation (Gilleland et al., 2009).

c. Feature-based methods

The feature-based verification methods focus verification on the forecast's ability to capture the overall structures of meteorological features (Gilleland et al., 2009). Similarly to the previous methods, features are typically identified by applying a threshold to the field, leaving behind images akin to that in the bottom-left of Figure 2.2. The basic idea of the feature-based approach is to evaluate attributes such as size, shape and average density over each object in the forecast and observation and compare how well these correspond to one another (Gilleland et al., 2009). These methods seem to have found a lot of application in the verification of precipitation fields and some methods have been developed precisely for that aim (see: e.g. Ebert and McBride, 2000, Grams et al., 2006 or M. E. Baldwin, 2003) but in principle feature-based methods can be applied to any desired field.

The *Method for Object-based diagnostic Evaluation* (MODE), developed by Davis et al. (2006), first applies a convolution procedure to the field before the field is thresholded and features are identified. This ensures that scattered events on a small-scale are correctly identified as belonging to the same 'object'. The algorithm then matches objects between forecast and observation based on various attributes such as the centroid position, total area, area overlap, intensity distribution etc., although other measures of similarity have also been proposed (see: e.g. Gilleland et al., 2008, Venugopal et al., 2005). An application of the approach to wind events is provided by Nachamkin (2004). One of the limitations of the method is that although it identifies individual features, it does not provide information about them (Gilleland et al., 2009). Marzban and Sandgathe (2006) instead identify features as the clusters obtained from a hierarchical cluster analysis. Traditional categorical scores can be used for verification after defining hits, misses and false alarms based on the proximity of clusters between two fields (Gilleland et al., 2009). Another approach is proposed by Micheas et al. (2007), termed the *Procrustus* method. Features in the Procrustus method are called 'cells' and are defined as any cluster of non-zero pixels of a user-defined size. It is therefore important that continuous fields have been thresholded in some manner before hand. Forecast and observation cells are matched based on either their centroid distances or a shape criterion. Procrustus provides diagnostic information about forecast quality in terms of rotation, dilation, translation and intensity-based errors (Gilleland et al., 2009). Lack et al. (2010) make a large departure from the typical feature-based methods and modify the Procrustus scheme by employing a scale-separation technique for identifying features, as well as utilising cell-by-cell verification metrics to provide summary statistics for individual forecast and observation objects (Gilleland et al., 2009).

Feature-based methods can provide useful diagnostics about the forecast by providing information about spatial and structural errors of individual features. The main disadvantage of the methods is the matching and merging of features, which can be "tricky" (Gilleland et al., 2010a).

d. Field deformation methods

The field deformation methods were among the first methods to be used for spatial forecast verification (e.g. Hoffman et al., 1995 or Alexander et al., 1999) and essentially attempt to find an optimal 'morphing' of the forecast field into the observation field (Gilleland et al., 2009). The methods produce a field of distortion vectors, indicating the directions, and degree of spatial manipulation necessary for the morphing, which can then be assessed (see: Figure 2.2, bottom-right). As Gilleland et al. (2009) mentions, these methods could be applied to individual features and as such be used within a feature-based framework. Field deformation methods are therefore distinguished from feature-based methods by the fact that they deal with the entire field at once, rather than performing localised, feature-to-feature comparisons.

Pioneering the usage of image warping for spatial verification purposes, Dickinson and Brown (1996) and Alexander et al. (1999) applied image warping techniques using polynomial warp functions to morph the forecast field into the observation field. A difficulty regarding these methods is that image warping requires a set of control points, the selection of which is crucial for ensuring physically meaningful warps but also far from trivial (Gilleland et al., 2009). The methods of Keil and Craig (2007, 2009) employ an image warping approach akin to *optical flow*, which is method whereby the pixel motion of adjacent frames is represented by a vector field (Zhai et al., 2021). Forecast performance is then summarised by a metric which captures the amount of movement and amplitude change (Gilleland et al., 2009). In Keil and Craig (2007) this metric is referred to as the forecast quality measure (FQM), while in Keil and Craig (2009) an improved version termed the displacement and amplitude score (DAS) is used. Marzban and Sandgathe (2010) and Marzban et al. (2009) provide reviews of the usage of optical flow in forecast verification. Venugopal et al. (2005) introduce a metric called the forecast quality index (FQI) to assess closeness of forecast and observation fields. It computes a quotient of the normalised partial Hausdorff distance (Huttenlocher et al., 1996) between the two fields and a measure of the intensity error. Although the method does not strictly perform any field deformation, it does measure spatial displacement across the entire field (Gilleland et al., 2009).

Vector fields can be a great diagnostic tool, and like the feature-based methods, the field deformation methods can provide useful and physically meaningful information about spatial and structural errors between the forecast and observation fields (Gilleland et al., 2010a). The obvious drawback is that the methods do not inform about individual features but can only assess the entire field at once.

Looking Ahead

As a "second phase" of the ICP, Dorninger et al. (2018) review the MesoVICT project, which was established in 2014 and addresses the particular challenge of deterministic and ensemble forecasts of atmospheric variables over complex terrain at near convection-resolving resolution. The project proposes adding another category of emerging spatial verification techniques to the ICP framework, termed *distance metrics*, which can be considered a hybrid between the feature-based and the field-deformation methods. A novel set of geometric verification measures for this new category is investigated by Gilleland et al. (2020).

2.2.2 Forecast Verification of Extreme Events

Extreme weather events pose a significant challenge to the forecast verification methodology. An extreme event in the atmospheric sciences loosely refers to any meteorological event that is both *rare* and *severe* Murphy (1991). While the forecasting of extreme weather events is one of the most important goals of weather prediction (Casati et al., 2008), and indeed poses its own challenges, there is a "critical gap" (Jolliffe and Stephenson, 2012, p. 187) in verification research in relation to such events. As Jolliffe and Stephenson (2012) goes on to state, without appropriate verification methods to *evaluate* forecasts in regard to extreme events, it will be correspondingly difficult to *improve* forecasts of extreme events, the demand of which is growing rapidly as certain types of high-impact hazards appear to be increasing in frequency and intensity due to climate change. The need for further research in the area of forecast verification of extreme events has been clearly recognised by weather services worldwide, such as the ECMWF, which stated in 2009 (Jolliffe and Stephenson, 2012, p. 187):

> "In view of the high priority given by the National Meteorological Services to the accurate prediction of severe weather events, the Subgroup recognised the lack of fundamental research into related verification within the Member States' meteorological services and universities. Without an enhanced research effort, it will continue to be difficult to provide those National Meteorological Services with robust performance measures."

and has made the forecasting of extreme weather events one of its four main goals in the ECMWF strategy of 2011-2020 (Magnusson et al., 2014). To that end, the renowned weather service has developed the Extreme Forecast Index (EFI) to evaluate the extreme weather predictions made by its probabilistic forecast ensembles. The World Meteorological Organization's THORPEX programme provides another example of dedication to the improvement of extreme weather prediction.

Binary Extreme Events

This section will briefly touch on the verification of deterministic forecasts of binary events, as most published research in this area has addressed this case (Jolliffe and Stephenson, 2012, p. 194) and the approaches can be combined with ease to the spatial verification methods described in the previous section, typically by choosing a threshold to distinguish extreme from non-extreme events. For good overviews of the subject we recommend Jolliffe and Stephenson (2012, chap. 10), Casati et al. (2008) and Magnusson et al. (2014).

We saw from the example of Finley's tornado forecasts (see: Table 2.1) that the forecast verification of rare events from a binary contingency table can produce severely misleading results when evaluated with the proportion correct PC, or, indeed, any categorical score involving the *correct negatives*, d. It can be shown, in fact, that the PC always tends to 100% as the base rate

s of the event decrease to zero, with the exception of when the event is so strongly overforecast that the false alarm rate F does not decrease to zero (Jolliffe and Stephenson, 2012, p. 186). As a measure of accuracy independent of the correct negatives, the threat score (TS, also known as the critical success index) has been widely used for rare events instead (Mason, 1989), while the problem can also be mitigated by measuring forecast skill relative to a naive reference forecast with an equitable skill score such as the Peirce skill score (PSS). Stephenson et al. (2008) have shown, however, that practically all categorical scores, including the Peirce skill score, *degenerate* to trivial values such as 0, 1 or infinity as the base rate of the event decreases to zero. This degeneracy is not limited to the binary contingency table but pertains to any multi-category forecast where some of the categories are rare (Jolliffe and Stephenson, 2012, p. 186). While the degeneracy problem makes it difficult to provide a fair comparison of scores between rare events with different base rates, this is not to say, however, that the traditional categorical scores have no place in the forecast verification of extreme events. These scores remain valuable indicators of particular attributes of the forecast and can nevertheless be used to track performance over time or compare a set of forecasts against the same observations (and hence the same base rate).

As a remedy to the degeneracy problem, Stephenson et al. (2008) proposed the extreme dependency score (EDS), which provided the first non-degenerate assessment of skill for extreme events (Jolliffe and Stephenson, 2012, p. 49). Since then, Ferro and Stephenson (2011) have proposed an improved version of the score called the extremal dependence index (EDI) and its symmetric counterpart, the SEDI. Beyond non-degeneracy, the EDI and SEDI have a number of other desirable properties (see: e.g. Jolliffe and Stephenson, 2012) which make them preferable over the EDS. In fact, the SEDI score combines more desirable properties into one skill score than any other categorical skill score (Jolliffe and Stephenson, 2012, p. 54). One important property that the EDS, EDI and SEDI lack, however, is true equitability. To address this Hogan et al. (2010) proposed the equitably transformed SEDI, $\mathrm{SEDI}_{\mathrm{ET}}$, combining the desired properties of true equitability, non-linearity and non-degeneracy into a single measure.

We mentioned in the previous discussion that vanishing base rates cause categorical scores to degenerate. A base rate of an event can, however, vanish in two ways: a) the number of observations of the event is very small *relative* to the total number of observations of all events, or b) there are simply very few observations of the event in an *absolute* sense. In the latter case, we encounter another common problem in the verification of extreme events, which is *sampling uncertainty*. Essentially, small samples in the contingency table can inflate the variation of verification measures and thereby increase uncertainty about the quality of the forecasts. On the one hand, the variance of most performance measures is inversely proportional to the sample size, and so small samples can inflate uncertainty, while on the other hand, certain measures may depend particularly critically on small numbers of hits, misses and false alarms, such as the hit rate or bias (Jolliffe and Stephenson, 2012, p. 186). One very common method to reduce sampling uncertainty and to obtain more precise performance estimates is to fit a statistical model to the joint forecast-observation distribution (see: e.g. Bradley et al., 2003) and to derive hits, misses etc. and the resulting scores from the model rather than computing them directly from the samples (Jolliffe and Stephenson, 2012, p. 191).

Economic Value

Lastly, we provide some consideration to the third type of measure of 'goodness' as originally proposed by Murphy (1993): the economic value of a forecast. While such considerations are rarely made in the literature of deep learning weather forecasting research, they bear some particular relevance to the forecasting of extreme events. One way to consider the economic value of a forecast of extreme events is to use a simple cost-loss decision model, where preventative action taken in preparation of an extreme event is given by a cost C and some loss L will be suffered if an extreme event occurs and no action is taken. Such a model can be summarised by an expense matrix, as shown in Table 2.2. Jolliffe and Stephenson (2012, chapter 9) derive that the value V of the forecast, defined as "the reduction in mean expense relative to the reduction that would be obtained by having access to perfect forecasts" can be expressed as:

Table 2.2: The expense matrix of a simple cost-loss decision model, showing cost C and losses L for different outcomes.

	Event occurs	
Action taken	Yes	No
Yes	C	C
No	L	0

$$V = \frac{\min(\alpha, s) - F(1-s)\alpha + Hs(1-\alpha) - s}{\min(\alpha, s) - s\alpha} \tag{2.2}$$

where $\alpha = C/L$ is the cost-loss ratio. This equation shows that economic value is dependent not just on the quality of the model (H and F) but also on the base-rate of the event s as well as the user's cost-loss ratio α. Shown in Figure 2.3, the curve of V can be plotted over a range of $0 < \alpha < 1$ for different combinations of H and F, arising, for example, from evaluating the model forecasts using different thresholds to distinguish extremes from non-extremes. The curves need only be considered on the range $0 < \alpha < 1$ since there would be no sense in taking action if the cost C is larger than the potential loss L. The maximum of the value curve can be shown to always lay at the point $\alpha = s$, where the value becomes $V_{\max} = H - F$, which is identical to the PSS, mentioned earlier. Hence, the PSS can be interpreted as a measure of potential forecast value as well as forecast quality (Jolliffe and Stephenson, 2012, p. 171). Likewise, the difference in the upper and lower limits of the cost-loss range over which the value curve is positive, and is thus of any benefit to the user, can be shown to be equal to the aforementioned Clayton skill score (CSS). As such the PSS and CSS quantify the two most important features of the value curve.

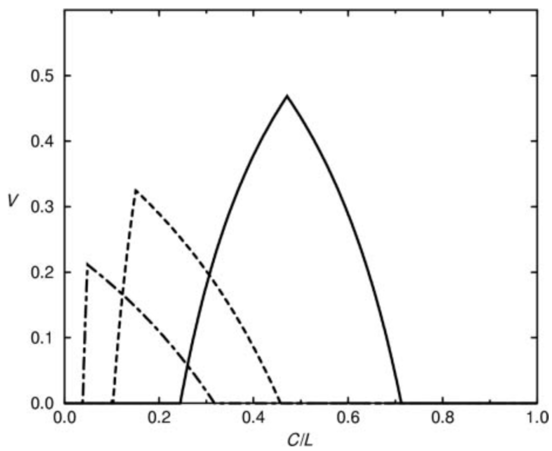

Figure 2.3: Three different value curves over the range $0 < C/L = \alpha < 1$, corresponding to different extreme precipitation events, defined as rainfall exceeding 1mm (solid line), 5mm (dashed line) and 10mm (chain-dashed line), from Jolliffe and Stephenson (2012, p. 171). **© Wiley Books. Used with permission.**

Chapter 3

Methodology

In this chapter the methodology of the thesis is addressed. The data collection and preprocessing steps are detailed in Section 3.1. By standardising the data using Z-normalisation locally at each grid-point we take a relative approach to wind speed forecasting and define extreme events as the 99th percentile of the locally standardised wind speed distribution. This approach focuses the definition of an extreme event on its *rarity* rather than its *severity* and mitigates a severe spatial imbalance of extreme events in the data between land and sea regions. Section 3.2 provides a description of the spatio-temporal forecasting model used in this thesis, which is an adaptation of the convolutional long short-term memory (ConvLSTM) model as proposed by Shi et al. (2015). By combining the ConvLSTM model with two types of loss functions borrowed from the literature on imbalanced regression, the weighted loss and the squared error-relevance area loss (SERA), we propose to tackle the severe data imbalance of the extreme wind speed forecasting problem. Finally, Section 3.3 details the forecast verification approach used to assess the quality of the forecast results. While much of the literature on spatio-temporal forecasting follows a very basic approach to forecast verification, we propose the application of a thorough spatio-temporal forecast verification based on the state-of-the-art of forecast verification methods for spatial fields and extreme events.

3.1 Data

3.1.1 Collection

The wind speed data was downloaded from the Copernicus Climate Change Service (C3S) Climate Data Store (CDS) of the ECMWF (see: Hersbach et al., 2018). The reanalysis data of the U and V components of the horizontal wind velocity (in m/s) were taken at 1000 hPa (ca. mean sea-level) from the *ERA5 hourly data on pressure levels from 1979 to present* dataset. By computing the sum of the squares of the two wind velocity components the scalar wind speed was obtained. The data was downloaded with a temporal resolution of 1 hour between 01-01-1979 and 01-01-2019 (40 years) on a spatial grid over Europe. The grid comprises 64×64 coordinates between latitudes 40-56° and

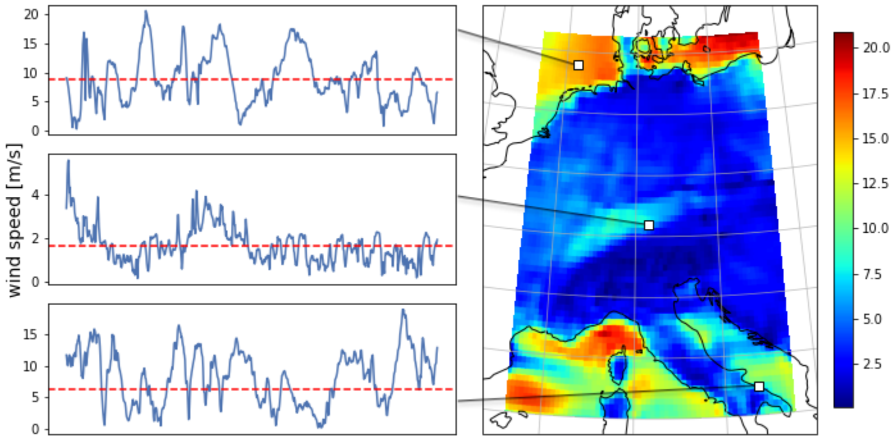

Figure 3.1: A visualisation of the wind speed data (in m/s) used in this thesis. The right figure shows a color-map of an example data frame, overlaid on an cartographic map showing the coastlines of the region. On the left, the wind speed time series of three arbitrary locations within the frame (white squares) are plotted for the duration of one month, as well as the climatological means at these locations (dotted red lines).

longitudes 3-19°, the spatial resolution being 0.25° (approximately 28km). This region was selected for its geographical variation, as it includes both land and sea regions as well as flat and mountainous areas, while the region is furthermore divided into different climatic regions such as the Pannonian climate region in Eastern Austria and the Alpine climate region covering the Austrian Alpine range. Interplay between these features can result in highly complex wind dynamics, which is where we expect the application of deep learning to be particularly promising.

A visualisation of the data is provided in Figure 3.1. The figure shows a color-map of an example data frame as well as the wind speed time series of three arbitrary locations over the duration of one month including the climatological means at these locations. Evidently the climatological means (and by extension the entire wind speed distributions) vary substantially throughout the region, where striking differences in magnitude can be seen between the sea and land regions. To highlight the spatial differences, Figure 3.2 shows the maximum, mean and standard deviation of the wind speed over the grid. The color-maps unveil a sharp division of the statistics with the underlying coastlines of the region. Indeed, extreme winds (e.g. larger than 25 m/s i.e. 10/12 on the Beaufort scale) seem to occur almost exclusively on the sea. If there were, in fact, stronger winds present over this region of mainland Europe between 01-01-1979 and 01-01-2019 then they have not been captured by the ERA5 reanalysis.

Such extremely localised occurrences of extremes are likely to impede model learning in favor of a fairly (spatially) naive model with a simple baseline prediction of extremes on the sea and non-extremes on the mainland. Therefore, rather than defining extreme winds in absolute terms, we proceed to take a relative approach and define extreme winds in terms of the 99th distributional percentile at each coordinate. This definition focuses more on the (relative) *rarity* of an extreme event and less on its *severity* (Murphy, 1991). The corresponding wind speeds in absolute terms may then easily be recovered from the local wind speed distribution.

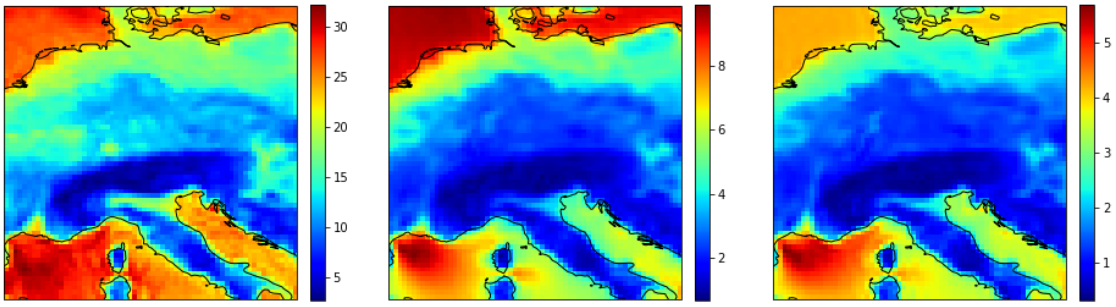

Figure 3.2: Color-maps of the maximum (left), mean (center) and standard deviation (right) of the wind speeds (in m/s) over the spatial grid. The figures display a sharp division of the statistics along the underlying coastlines of the region.

3.1.2 Preprocessing

The raw wind speed data was standardised with a local Z-normalisation (Goldin and Kanellakis, 1995) at each coordinate, which centers the local distributions around zero mean with unit standard deviation according to the following transformation:

$$x' = \frac{x - \mu}{\sigma} \tag{3.1}$$

where x' denotes the transformed variable, x the original variable and μ and σ the mean and standard deviation, respectively. The Z-normalised data thus represents the wind speed in terms of the number of standard deviations from the respective mean at each location. Not only is this type of standardisation typically a prerequisite for the successful training of ANN models (due to limiting gradient sizes), it also ensures that extreme values i.e. the tail of the data distribution is now represented more equally over all locations. In Figure 3.3 the full data distributions of the raw data and the standardised data are provided, along with the locations of the 90th, 95th and 99th percentiles (P90, P95 and P99). The 99th percentile of the standardised wind speed distribution is located at 2.7 σ, which we define here as the threshold between extremes and non-extremes.

Figure 3.3 furthermore highlights the severe imbalance of the data distribution. Wind speed is typically distributed according to a Weibull distribution, which is described by the two parameters, $k > 0$, the shape parameter, and $\lambda > 0$, the scale parameter. This family of distributions is typically long-tailed and includes the exponential distribution ($k = 1$) and the Rayleigh distribution ($k = 2, \lambda = \sqrt{2}\sigma$). Using the Weibull calculator tool of the Swiss wind power data website from 'Suisse Eole'[1] to compute k and λ for the non-standardised wind speed distribution obtains $\lambda = 4.49 m/s$ and $k = 1.18$. While λ is a measure for the characteristic wind speed of the distribution, k specifies the shape of the distribution and takes on values between 1 and 3. Touching on the lower end of this range, a k-value of 1.18 indicates highly variable wind speeds, the Suisse Eole website explains, while constant winds are characterised by large k-values. The low k-value is likely due to the fact that the distribution samples over a large spatial grid, within which wind speeds are indeed bound to vary widely; it underlines, however, the dynamic complexity of the spatio-temporal wind speed forecasting problem.

[1] https://wind-data.ch/tools/weibull.php, visited 14-10-2021.

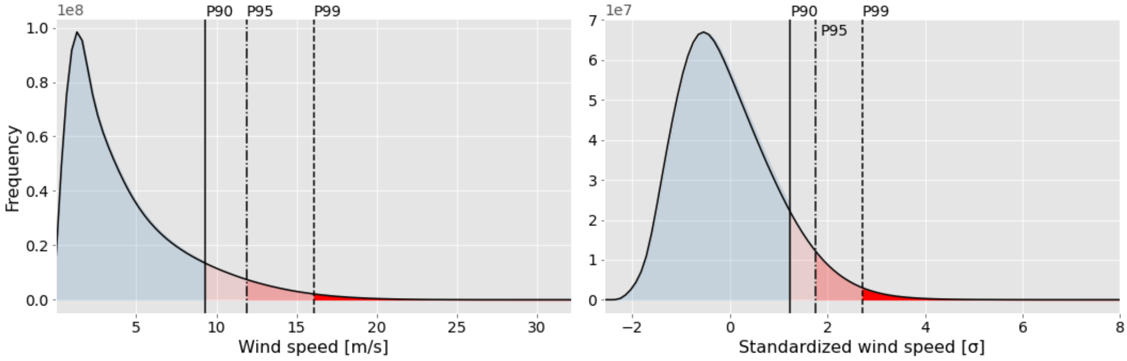

Figure 3.3: The full data distributions of the raw wind speed (left) and the locally standardised wind speed (right). The locations of the 90th, 95th and 99th percentiles (P90, P95 and P99) are shown in both graphs, where P99 defines the threshold of an extreme wind event in this work. The 99th percentiles of either distribution are located at 16.1 m/s and 2.7 σ (standard deviations), respectively. The standardised wind speed distribution was cut at 8 σ due to visual considerations - in reality the tail of the distribution reaches out to 13.7 σ.

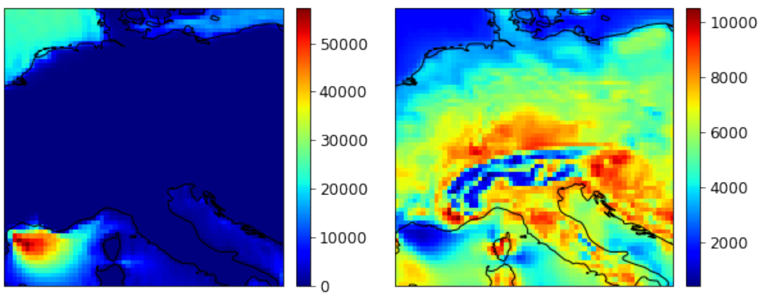

Figure 3.4: Color-maps of the absolute occurrences of wind speeds exceeding the 99th percentiles at each location on the spatial grid, for the raw wind speed data (left) and the standardised wind speed data (right).

The absolute occurrences of wind speeds exceeding the 99th percentile at each location on the grid are shown in Figure 3.4. We can see that locally standardised extreme events are significantly more evenly distributed over the spatial grid than non-standardised extreme winds. A model tasked with forecasting extreme winds based on the non-standardised data distribution is presumably faced with a less challenging spatio-temporal forecasting problem regarding the modelling of spatial correlations in the data, since extreme winds evidently occur primarily around a single location. Defining wind extremes relatively based on the local wind speed distributions creates a much more spatially varied forecasting problem, the investigation of which appears to be of more value to the area of spatio-temporal sequence forecasting than the former case.

The standardised wind speed data was subsequently divided into a training, validation and test set, where four-fold cross validation was utilised on the first 32 years of data (24 years training, 8 years validation) and the last 8 years were held out as test data. It appears common, however, in the literature on meteorological forecasting with machine learning methods to use a single training-validation split, and thus neglect cross validation of the results (see: e.g. Ahmed et al., 2021; Feng and Fox, 2021; Ferreira et al., 2019; Grönquist et al., 2021; Li et al., 2018; Li and Shi, 2010; Liu et al., 2018b; Liu et al., 2016; Marndi et al., 2020; McGovern et al., 2017; Mi et al., 2017; Racah et al., 2017; Scher and Messori, 2021; Shi et al., 2015, 2017; Wang et al., 2017a; Xie et al., 2021; Zhu et al., 2018; Zucatelli et al., 2020). We expect that this is done with the underlying assumption that for large datasets any arbitrary split samples the underlying distribution sufficiently well and thus statistical variances over splits are expected to be small. Indeed, among this literature, the work of Grönquist et al. (2021) is the only work we are aware of that supports this assumption with results obtained from a trial run. Following Grönquist et al. (2021), we investigated the variances of the results in a trail run with four-fold cross validation and found that the variances in the results for forecasts on long lead-times can, in fact, be substantial, with relative standard deviations ranging up to 11%. We therefore opted to cross validate our models using four-fold cross-validation and present the results as the mean ± one standard deviation. This is is line with the work of e.g. Ding et al. (2019), Marzban and Stumpf (1998), Rao et al. (2020), and Scheuerer et al. (2020). We caution against the presentation of model performance on a single training-validation split without prior investigation of the variances of the results.

3.2 Model Description

The spatio-temporal forecasting model that we propose in this work is an adaptation of the Convolutional Long Short-Term Memory (ConvLSTM) Network proposed by Shi et al. (2015) for precipitation nowcasting. While Shi et al. (2015) trained their ConvLSTM models using cross-entropy loss, we propose utilising two types of loss functions borrowed from the literature on imbalanced regression: weighted loss and the squared error-relevance area (SERA) loss. The ConvLSTM is an example of a ConvRNN model i.e. a combination of a convolutional neural network (CNN) and a recurrent neural network (RNN). By utilising the strengths of the CNN to capture spatial correlations and the RNN to capture temporal correlations in the data, ConvRNN based models have demonstrated promising forecasting ability in the spatio-temporal setting. Before discussing the ConvLSTM model in Section 3.2.3, we provide two brief overviews of the concepts of CNN and LSTM networks in Section 3.2.1 and 3.2.2. Lastly, Section 3.2.4 discusses the loss functions used for model training and Section 3.2.5 details how the ConvLSTM model was implemented.

3.2.1 Convolutional Neural Networks

CNNs are a class of feedforward artificial neural networks, used primarily for data mining tasks involving spatial data and have gained a lot of attention in the area of computer vision and natural language processing (Ghosh et al., 2020). Through a series of alternating convolution and pooling layers, the CNN creates a series of feature maps of increasing abstraction of the original input image. This is achieved with a number of spatial filters called *kernels*, the values of which are learnable weights in the network (typically integer values). The convolutional layer computes the convolution of the input image (or the previous feature layer) with each of the n kernels to create a set of n feature maps comprising the next feature layer. The step-size in which the kernel is convolved over the input image is called the *stride*. In general, the calculation of the convolutional layer can be expressed as (Zhu et al., 2018):

-1	0	2
0	1	0
1	-2	-1

Figure 3.5: Example of a 3×3 kernel, the values of which are learnable weights in the CNN.

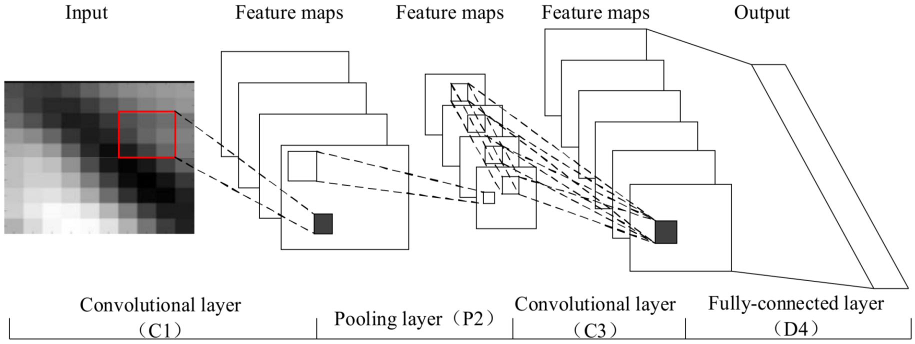

Figure 3.6: Typical network structure of a (two-dimensional) CNN connected to a fully connected output layer, from Zhu et al. (2018).

$$x_q^l = \sigma\big(\sum_p x_p^{l-1} * k_{pq}^l + b_q^l\big) \tag{3.2}$$

where x_q^l is the qth feature map of the lth layer, k_{pq}^l is the trainable convolutional kernel and b_q^l the bias matrix. The summation is performed over the p input maps and the $*$ operator denotes the convolution operation. σ denotes the non-linear activation function of the nodes in the network e.g. sigmoid, tanh or ReLu activation.

The pooling layer is essentially a down-sampling layer which attempts to conserve only the most dominant features, typically by computing the maximum of the values within the kernels (Ghosh et al., 2020). Through the supervised training of the network, the kernels learn to adopt appropriate weights in order to construct feature representations i.e. abstractions of the input image that are useful to the task at hand. The final feature layer may be connected to a fully connected layer to collapse the image into a one-dimensional output, as is done in classification tasks, but it may also, for example, be connected to a reverse engineered decoder network to reconstruct the image from its feature representation e.g. for clustering applications. Figure 3.6, taken from Zhu et al. (2018), provides a visualisation of the former, fully connected case. The basic CNN structure can easily be extended to 3-dimensional, spatio-temporal input by using 3-dimensional kernels (see: e.g. Racah et al., 2017).

3.2.2 Long Short-Term Memory

Arguably the most popular RNN for sequence modelling, the long short-term memory (LSTM) network structure is known for its powerful ability to effectively model long-range temporal dependencies (Shi et al., 2015). The trademark feature of the LSTM is its memory cell c. The state of the memory cell at time t, c_t, is carefully controlled by a series of gates: The activation of the input gate i_t determines whether new input is accumulated into the memory cell; the activation of the forget gate f_t determines whether the previous cell state c_{t-1} is cleared, and, lastly, the activation of the output gate o_t determines whether the latest cell state c_t is passed through to the final 'hidden state' of the network h_t. In the fully connected (FC-LSTM) case, the above quantities are computed as follows (Shi et al., 2015):

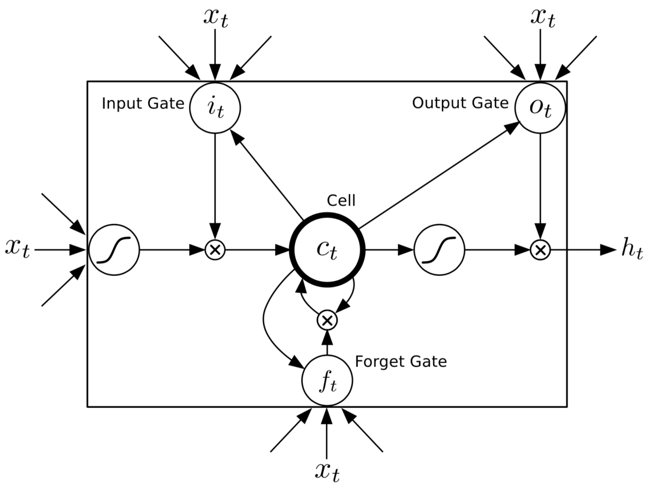

Figure 3.7: The LSTM cell, from Graves (2013). The curves symbols mark the tanh activation functions and crosses ('×') mark the Hadamard product. Compare with equations 3.3.

$$
\begin{aligned}
i_t &= \sigma\big(W_{x,i}x_t + W_{h,i}h_{t-1} + W_{c,i} \circ c_{t-1} + b_i\big) \\
f_t &= \sigma\big(W_{x,f}x_t + W_{h,f}h_{t-1} + W_{c,f} \circ c_{t-1} + b_f\big) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \tanh\big(W_{x,c}x_t + W_{h,c}h_{t-1} + b_c\big) \\
o_t &= \sigma\big(W_{x,o}x_t + W_{h,o}h_{t-1} + W_{c,o} \circ c_{t-1} + b_o\big) \\
h_t &= o_t \circ \tanh(c_t)
\end{aligned}
\tag{3.3}
$$

where $W_{x,y}$ denotes the network weights for layer/cell x and gate/cell y and the '$\circ$' operator denotes the Hadamard product (i.e. element-wise matrix multiplication). For a complete picture of the LSTM network we refer the reader to Graves (2013). A diagram of the LSTM cell, taken from Graves (2013), is provided in Figure 3.7. Single LSTM cells can be stacked and temporally concatenated to form more complex models, which have found usage in many sequence modelling problems (Shi et al., 2015).

3.2.3 Convolutional Long Short-Term Memory

As Shi et al. (2015) notes, the major drawback of the LSTM network in handling spatio-temporal data is that no spatial information is encoded in the input-to-state and state-to-state transitions. The authors propose making all the inputs $\mathcal{X}$, cell outputs $\mathcal{C}$, hidden states $\mathcal{H}$ and gates 3-dimensional spatio-temporal tensors and using convolution in the input-to-state and state-to-state transitions. This is the essence of the ConvLSTM network. Analogously to the LSTM equations (3.3), the ConvLSTM equations are given as (Shi et al., 2015):

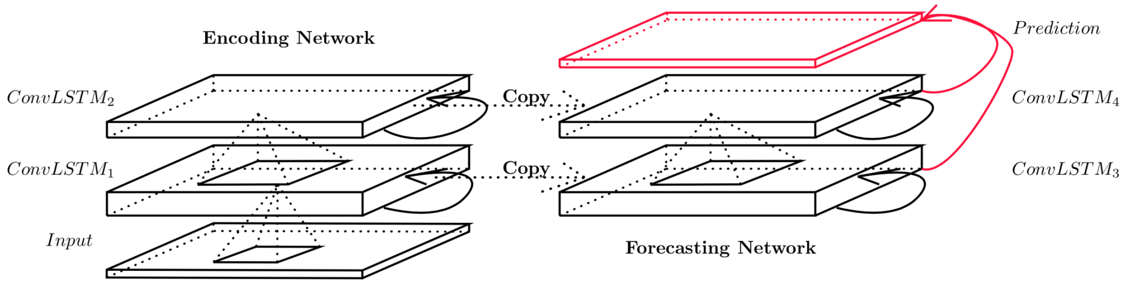

Figure 3.8: The multi-layered encoding-forecasting ConvLSTM network, from Shi et al. (2015). The hidden states and cell outputs of the encoding network are copied to the forecasting network, from which the final prediction is made.

$$
\begin{aligned}
i_t &= \sigma\big(W_{x,i} * \mathcal{X}_t + W_{h,i} * \mathcal{H}_{t-1} + W_{c,i} \circ \mathcal{C}_{t-1} + b_i\big) \\
f_t &= \sigma\big(W_{x,f} * \mathcal{X}_t + W_{h,f} * \mathcal{H}_{t-1} + W_{c,f} \circ \mathcal{C}_{t-1} + b_f\big) \\
\mathcal{C}_t &= f_t \circ \mathcal{C}_{t-1} + i_t \circ \tanh\big(W_{x,c} * \mathcal{X}_t + W_{h,c} * \mathcal{H}_{t-1} + b_c\big) \\
o_t &= \sigma\big(W_{x,o} * \mathcal{X}_t + W_{h,o} * \mathcal{H}_{t-1} + W_{c,o} \circ \mathcal{C}_{t-1} + b_o\big) \\
\mathcal{H}_t &= o_t \circ \tanh(\mathcal{C}_t)
\end{aligned}
\tag{3.4}
$$

As is common for spatio-temporal sequence forecasting, the ConvLSTM model of Shi et al. (2015) adopts an encoding-forecasting structure, where both encoding and forecasting networks consist of several stacked ConvLSTM layers, which we adopt here too. Analogously to the typical encoder-decoder CNN structure, the encoding ConvLSTM network compresses the input into a hidden state tensor and the forecasting ConvLSTM network unfolds this hidden state into the final prediction. The model is, furthermore, a multi-frame forecasting model, with the prediction of the network having the same dimensionality as the (spatio-temporal) input. A diagram of the encoding-forecasting network structure, taken from Shi et al. (2015), is provided in Figure 3.8. As a deep ConvRNN model, the ConvLSTM has the potential to effectively model complex and dynamical spatio-temporal systems like the wind speed forecasting problem investigated in this thesis.

3.2.4 Loss

Weighted Loss

In order to combat the effects of data imbalance on the spatio-temporal extreme wind forecasting problem, we test the ConvLSTM model with two different loss functions that have been proposed for imbalanced regression problems. The first of these is the straightforward *weighted loss*, which consists of assigning a weight $w(y)$ to each value in the input frame according to its target wind speed y. While these weights can be chosen arbitrarily (as done by Shi et al., 2017), we propose to use weights equal to the inverse of the data distribution for each target, as suggested by Yang et al. (2021). For a loss function L of the (spatio-temporal) target y and prediction $\hat{y}$, consisting of N time-frames of $M \times M$ spatial coordinates, the weighted loss L_W is computed as:

Table 3.1: Discrete intervals of the standardised wind speed distribution and their respective proportions.

standardised wind speed [σ]	Proportion [%]
$x < 0$	55.85
$0 \leq x < 1$	28.28
$1 \leq x < 2$	11.87
$2 \leq x < 3$	3.27
$3 \leq x < 4$	0.62
$4 \leq x$	0.12

$$L_W(\hat{y}, y) = \frac{1}{N} \sum_{n=1}^{N} \sum_{i,j=1}^{M} w(y_{n,i,j}) \cdot L(\hat{y}_{n,i,j}, y_{n,i,j}) \tag{3.5}$$

As weighted loss functions we investigate both the weighted mean squared error (W-MSE) loss and the weighted mean absolute error (W-MAE) loss. The weights are computed from the inverse of the standardised wind speed distribution, for discrete wind speed intervals. The proportions of the distribution for each discrete interval are shown in Table 3.1. Due to the rapidly exploding inverses of the distribution at the tails, we assign all target values $y \geq 4$ equivalent weight. Values $y < 0$ are assigned unit weight since we want to focus predictive performance on the right tail of the distribution. Thus, taking the inverses of the proportions from Table 3.1 and dividing all values such that values $y < 0$ get assigned unit weight results in the following weights (simplified to two significant figures):

$$w(y) = \begin{cases} 1.0 & \text{if} \quad y < 0 \\ 1.5 & \text{if} \quad 0 \leq y < 1 \\ 3.5 & \text{if} \quad 1 \leq y < 2 \\ 13 & \text{if} \quad 2 \leq y < 3 \\ 66 & \text{if} \quad 3 \leq y < 4 \\ 420 & \text{if} \quad 4 \leq y \end{cases} \tag{3.6}$$

Squared Error-Relevance Area Loss

As a second approach to combating data imbalance, we implement the squared error-relevance area (SERA) loss, as proposed by Ribeiro and Moniz (2020) for imbalanced regression problems. The SERA loss is based on the concept of a *relevance function* $\phi : \mathcal{Y} \longrightarrow [0, 1]$, which maps the target variable domain $\mathcal{Y}$ onto a $[0, 1]$ scale of relevance. The relevance function ϕ is determined through a cubic Hermite polynomial interpolation of a set of 'control-points'. The set of control-points $S = \{\langle y_k, \phi(y_k), \phi'(y_k)\rangle\}_{k=1}^{s}$ are user-defined points where the relevance may be specified, which are typically local minima or maxima of relevance and thus all have derivative $\phi'(y_k) = 0$ (Ribeiro and Moniz, 2020). We define the location of the 90th percentile (P90) as the point of minimum relevance $\langle y_1 = 1.2,\ \phi(y_1) = 0.0,\ \phi'(y_1) = 0.0\rangle$ and the 99th percentile (P99) of the standardised wind speed distribution as the point of maximum relevance $\langle y_1 = 2.7,\ \phi(y_1) = 1.0,\ \phi'(y_1) = 0.0\rangle$.

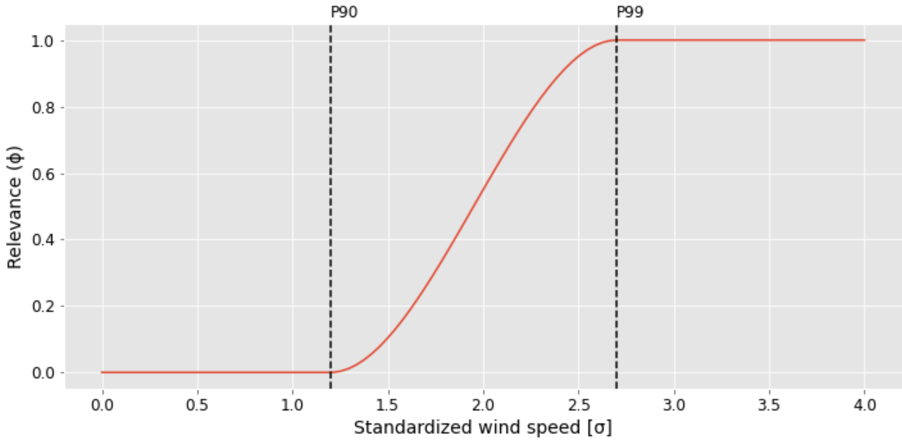

Figure 3.9: The relevance function ϕ obtained by interpolating the 90th percentile (P90 $= 1.2$) as control-point of minimum relevance and the 99th percentile (P99 $= 2.7$) as control-point of maximum relevance, using the *pchip* algorithm of Ribeiro and Moniz (2020).

The interpolation is carried out according to Ribeiro and Moniz (2020) by using the piecewise cubic Hermite interpolating polynomials (*pchip*) algorithm. The obtained relevance function is shown in Figure 3.9. Defining D^t as the subset of data pairs for which the relevance of the target value is greater or equal than a cut-off t, i.e. $D^t = \{\langle x_i, y_i \rangle \in D | \phi(y_i) \geq t\}$, the squared error-relevance SER_t of the model with respect to the cut-off t is computed as follows:

$$SER_t = \sum_{i \in D^t} (\hat{y}_i - y_i)^2 \tag{3.7}$$

where $\hat{y}_i$ and y_i are the i'th prediction and target values, respectively. The curve obtained by plotting SER_t against t is decreasing and monotonic (Ribeiro and Moniz, 2020) and provides an overview of how the magnitudes of the prediction errors change on subsets comprising varying degrees of relevant samples ($t = 0$ representing all samples and $t = 1$ representing only the most relevant samples). Finally, the squared error-relevance area (SERA) is defined as the area under the SER_t curve:

$$SERA = \int_0^1 SER_t \, dt \tag{3.8}$$

The smaller the area under the curve is, the better the model is. We note that assigning uniform relevance values to all data points recovers the MSE loss.

3.2.5 Implementation

The model was implemented in Pytorch, based heavily on an implementation by J. Huang, freely available on Github (see: `https://github.com/jhhuang96/ConvLSTM-PyTorch`). Primarily, changes had to be made to the data loaders to compute the relevance values used for the computation of the SERA loss, to the training function in order to incorporate the novel loss functions, and to the network parameters in order for the model to take 12, 64×64 frames as input and to provide a prediction for the succeeding 12 frames (12 hour lead-time). To this end, the data was divided into

12-frame tensors of size (12, 64, 64) in steps of 12 frames (i.e. non-overlapping). This resulted in 29,198 input-target pairs from the 40 years of wind speed data.

Beyond the different loss functions, we investigate model architectures with different numbers of ConvLSTM layers, ranging from 2-5 layers (in both the encoder and the forecasting networks). The number of parameters of each of the different model architectures are shown in Table 3.2. In line with Shi et al. (2015), all layers utilise 3×3 kernels. The convolution over each successive filter operates such as to successively halve the spatial dimensions of the input, while the number of hidden states (features) are successively doubled (starting from 16 hidden states).

The loss function was optimised through mini-batch gradient descent (batch-size of 16) using the adaptive moment estimation (Adam) optimiser. Adam optimiser is a popular and reliable choice for deep learning neural networks which computes adaptive learning rates for each parameter of the model, based on their update frequency (Ruder, 2017). The initial learning rate of the Adam optimiser was set to 10^{-3}. During training, early-stopping was performed on the validation set to ensure that the model with the lowest validation loss was saved as the best model i.e. to avoid overfitting the model. The early-stopping mechanism was set up to stop training when the validation loss did not decrease for 20 consecutive epochs.

Table 3.2: The number of parameters of the ConvLSTM model with different numbers of layers.

Layers	Parameters
2	2,385,953
3	10,061,025
4	34,201,185
5	62,060,641

3.3 Verification

Since the ConvLSTM model investigated in this thesis is a spatio-temporal forecasting model it appears in order to evaluate the model with a set of verification methods that capture the model's forecasting ability at different temporal horizons as well as at different spatial scales. This section gives an overview of the forecast verification methods used in this thesis to assess the quality of the model forecasts against the observations in the data. It is astonishingly rare to find examples of spatio-temporal verification in the literature on spatio-temporal forecasting, with most papers assessing forecast quality using single point-wise measures such as the MSE, RMSE, the correlation coefficient, or the basic categorical scores such as the hit rate H, false alarm ratio FAR or threat score TS, with no regard for a spatial or temporal assessment (see: e.g. Feng and Fox, 2021, Noorollahi et al., 2016, Rao et al., 2020, Robert et al., 2013, Shi et al., 2015, Shi et al., 2017, Ferreira et al., 2019). We are only aware of one exception, which is the work done by Amato et al. (2020). We put forward to make use of the vast literature on spatial forecast verification in order to provide a spatio-temporal assessment of forecast quality and demonstrate one way of doing so here.

In order to evaluate the model at different spatial scales we utilise three different filtering-based spatial verification methods: The minimum coverage method, the fractions skill score (neighbourhood methods) and the intensity-scale method (scale-separation method). Unlike the displacement-based spatial verification methods, filtering methods work well for verifying "messy" forecasts that do not contain well-defined features (Ebert, 2009), which we expect to be particularly applicable to wind speed data due to its highly stochastic behaviour. Moreover, displacement-based methods typically require the specification of parameters such as the threshold for defining an object and the search-distance used for matching objects (Ebert, 2009), while the filtering-based methods used here are parameter-free. While the methods used here are spatial forecast verification methods, they can be applied in a straightforward manner *per* lead-time for a temporal assessment.

Remark Initially in this work a multi-categorical approach was followed where the continuous wind speed values were divided into discrete bins, where a discrete category i denoted the wind speed interval $[i, i+1)$. The results were then verified on a per-category basis. While some filtering methods can accommodate multiple categories most of them were devised for binary problems where a set of intensity thresholds is chosen to divide the data into events and non-events. Thus, while the methods used here could principally be applied *per* category, this is not their intended use. More importantly, however, cutting up continuous data into discrete bins for comparative purposes creates a clear problem along the bin-edges. For example, forecast-target pairs 0.01 and 0.99 would both be put into category 0 and the forecast would be judged as correct while pairs 0.99 and 1.01 would be put into different categories 0 and 1 and the forecast would be judged as incorrect, although, clearly, the latter forecast was more accurate than the former. This issue is greatly mitigated when using just a single threshold to divide the data into events and non-events and repeating comparisons for different thresholds. When this work was presented during the 2021 Annual Meeting of the European Meteorological Society[2] we pursued a multi-categorical verification approach and the results were presented as such; in this work we instead present the results using a set of intensity thresholds and submit that this approach is to be preferred over the multi-categorical approach.

3.3.1 Minimum Coverage

Proposed by Damrath (2004), the minimum coverage (MC) method essentially states that 'a forecast is useful if the event is predicted over a minimum fraction of the region of interest' (Ebert, 2008). Denoting $\langle P \rangle_s = \frac{1}{n}\sum_n I$ to be the fraction of grid points with events $I \in \{0, 1\}$ within a neighbourhood of scale s, the entire neighbourhood is classified as the event $\langle I \rangle_s$ according to:

$$\langle I \rangle_s = \begin{cases} 0 & \langle P \rangle_s < P_e \\ 1 & \langle P \rangle_s \geq P_e \end{cases} \tag{3.9}$$

where P_e is the minimum fraction of the neighbourhood that must be covered by events in order for the neighbourhood to be classified as an event. A neighbourhood of scale s refers to a squared area of dimension $s \times s$ grid-points (sometimes also presented in kilometers). Due to the scarcity of extreme events in the data, we choose to use a minimum coverage criterion with P_e set to the value $1/n$ i.e. we require only a single event to be present in the neighbourhood for the neighbourhood to be classified as an event. The neighbourhood events can then be evaluated from a typical 2×2 contingency table using any desired categorical score. The categorical scores used here are the symmetric extremal dependence index (SEDI), the Peirce skill score (PSS) and the Clayton skill score (CSS). These scores are defined as:

[2]Scheepens, D., Hlavackova-Schindler, K., Plant, C., and Schicker, I.: A deep CNN model for medium-range spatio-temporal wind speed prediction for wind energy applications, EMS Annual Meeting 2021, online, 6–10 Sep 2021, EMS2021-238, https://doi.org/10.5194/ems2021-238, 2021.

$$\text{SEDI} = \frac{\log F - \log H + \log(1-H) - \log(1-F)}{\log F + \log H + \log(1-H) + \log(1-F)} \qquad \in [-1,1] \tag{3.10}$$

$$\text{PSS} = \frac{ad - bc}{(b+d)(a+c)} = H - F \qquad \in [-1,1] \tag{3.11}$$

$$\text{CSS} = \frac{ad - bc}{(c+d)(a+b)} \qquad \in [-1,1] \tag{3.12}$$

where $H = a/(a+c)$ and $F = b/(b+d)$ are the hit rate and false alarm rate, respectively, and a, b, c and d are the elements from the 2×2 contingency table (see: Figure 2.1). The SEDI was chosen due to its property of non-degeneracy for rare events, while the score also combines more desirable properties into one skill score than any other categorical skill score (see: Section 2.2.2). The usage of the PSS and CSS is motivated by the fact that they are both measures of potential forecast *quality* as well as forecast *value*, while both scores are also truly equitable skill scores. As explained in Section 2.2.2, the PSS can be shown to be equal to the maximum of the economic value curve of a simple cost-loss decision model, while the CSS corresponds to the cost-loss range over which the value curve is positive. We propose to prioritise the PSS over the CSS in order to maximise the potential economic value of the model, granted that the CSS remains greater than 0 and thus a cost-loss range exists over which the value curve is positive.

As for all neighbourhood methods, the scores are computed for a set of neighbourhood scales and a set of intensity thresholds for a diagnostic assessment of forecast quality on spatial scale and intensity. To this end we computed the scores for the set of scales $S = \{1,\ 3,\ 5,\ ...,\ 15\}$ (grid-points) and the set of thresholds $T = \{1.2,\ 1.8,\ 2.7,\ 4.0\}$ (standardised wind speed) corresponding to the 90th, 95th, 99th and 999th percentiles of the standardised wind speed distribution. The 999th percentile was included to investigate forecasting ability beyond the extreme/non-extreme threshold. In order to obtain an aggregated result over all forecast and observation frames, the elements in the contingency table are aggregated over all frames and the scores are computed afterwards from the aggregated contingency table.

Further Analysis

For very rare events, where the number of true negatives d becomes very large in comparison to a, b or c, we offer another interpretation of the PSS and CSS based on the concepts of *recall* and *precision* from the machine learning literature. Recall $= a/(a+c)$ (equivalent to the hit rate H) measures how many of the total observed events $a + c$ were forecast as events, a, and thus focuses performance on the missed hits c, while precision $= a/(a+b)$ measures how many of the forecast events, $a + b$, were actually observed events, a, and thus focuses on forecast quality on the false alarms b. We show here that for very rare events the PSS and CSS approximate the recall and precision, respectively, by using the fact that $b+d \approx d$ and $c+d \approx d$ for d much larger than b and c:

$$\text{PSS} = \frac{ad - bc}{(b+d)(a+c)} \approx \frac{ad - bc}{d(a+c)} = \frac{a}{a+c} - \frac{bc}{d(a+c)} \approx \frac{a}{a+c} \tag{3.13}$$

$$\text{CSS} = \frac{ad - bc}{(c+d)(a+b)} \approx \frac{ad - bc}{d(a+b)} = \frac{a}{a+b} - \frac{bc}{d(a+b)} \approx \frac{a}{a+b} \tag{3.14}$$

where the subtracted fractions in the last step vanish when, again, d is much larger than a, b, and c. When faced with the choice, forecasts of extreme events are desired to have high recall rather than precision since the losses associated with missing an extreme events are typically much larger than the cost of preparing for an extreme event (as is assumed in the basic cost-loss model mentioned before). This observation supports the prioritisation of the PSS over CSS for extreme events.

3.3.2 Fractions Skill Score

The computation of the fractions skill score (FSS), as proposed by Roberts and Lean (2008) and Roberts (2008), follows the same basic approach as the minimum coverage method, however, rather than classifying a whole neighbourhood as an event or non-event based on the fraction of events in the neighbourhood, and comparing these between forecast and observation, the FSS compares the fractions of events themselves. It essentially states that 'a forecast is useful if the frequency of forecast events is similar to the frequency of observed events' (Ebert, 2008). The fractions are used first to compute the fractions Brier score (FBS) and the worst possible fractions Brier score ($\mathrm{FBS}_{\mathrm{worst}}$), defined as:

$$\mathrm{FBS} = \frac{1}{N}\sum_{n}\left(\langle P_y\rangle_{s,n} - \langle P_x\rangle_{s,n}\right)^2 \tag{3.15}$$

$$\mathrm{FBS}_{\mathrm{worst}} = \frac{1}{N}\Big[\sum_{n}(\langle P_y\rangle_{s,n})^2 + (\langle P_x\rangle_{s,n})^2\Big] \tag{3.16}$$

where n sums over all fractions of neighbourhood windows of scale s in the spatial domain, the total of which is given by N. The $\mathrm{FBS}_{\mathrm{worst}}$ is essentially the largest FBS that could be obtained in the worst possible situation where there is no "collocation of non-zero fractions" (Roberts, 2008) between the forecast and observation whatsoever i.e. either the forecast fraction is always zero when the observation is not or vice versa. The FSS is then computed as a skill score of the FBS:

$$\mathrm{FSS} = 1 - \frac{\mathrm{FBS}}{\mathrm{FBS}_{\mathrm{worst}}} \qquad \in [0, 1] \tag{3.17}$$

The FSS takes values between 0 (complete mismatch) and 1 (perfect forecast). Roberts (2008) explains that for a collection of forecasts the FBS and $\mathrm{FBS}_{\mathrm{worst}}$ are computed for each forecast individually and combined in the end to compute the FSS. The FSS is typically compared to the skill of a random forecast with same fractional coverage f_0 of the event as the target observation, which is given as $\mathrm{FSS}_{\mathrm{random}} = f_0$. A random forecast on the 99th percentile ($f_0 = 0.01$) would thus obtain $\mathrm{FSS}_{\mathrm{random}} = 0.01$. Minimal sufficient skill is then defined as the FSS halfway between the random forecast skill and perfect skill, which represents the skill that would be obtained by uniform forecast at the grid scale ($s = 1$) forecasting fraction f_0 at every point: $\mathrm{FSS}_{\mathrm{uniform}} = 0.5 + f_0/2$. Comparison of the forecast skill with $\mathrm{FSS}_{\mathrm{uniform}}$ allows an objective assessment of the spatial scale at which, and/or the lead-time until which, the forecast is sufficiently skillful. The FSS is computed for the same set of neighbourhood scales S and intensity thresholds T as in the minimum coverage method.

3.3.3 Intensity-Scale

Rather than using neighbourhoods of increasing scale to assess spatial forecasting ability, Casati et al. (2004) and Casati (2010) propose a scale-separation approach called the intensity-scale (IS) method. Similarly to the neighbourhood methods, the forecast is evaluated for a set of spatial scales and intensity thresholds, however, the scales in the IS method are obtained by a wavelet decomposition of the spatial field using 2D Haar wavelets. The wavelets have squared support with dimensions equal to 2^L grid points, where L is the scale component, and thus over a field of $2^L \times 2^L$ grid points there are $L+1$ scale components. In our case of 64×64 fields this results in 7 scale components of size 1, 2, 4, 8, 16, 32, 64.

First, the forecast and observation fields are thresholded into binary fields using the same set of thresholds T as used in the previous two methods. After performing the 2D Haar wavelet decompositions of the binary fields, the forecast and observation field are then reconstructed from the wavelet coefficients at each individual scale component, capturing different features at each scale, and the MSE is computed between them. Finally, the intensity-scale score (ISS) is computed as a skill score from the MSE against a random reference forecast. The random MSE at threshold u is estimated from the 2×2 contingency table obtained from the binary (thresholded) observation and forecast fields, using the base rate $s = (a+c)/n$, forecast rate $r = (a+b)/n$ and frequency bias index $B = r/s$:

$$\mathrm{MSE}_{u,\mathrm{rand}} \sim B \cdot s \cdot (1-s) + s \cdot (1 - B \cdot s) \tag{3.18}$$

The ISS is then computed as a skill score of the MSE for each threshold u and each scale l according to:

$$\mathrm{ISS}_{u,l} = 1 - \frac{\mathrm{MSE}_{u,l}}{\mathrm{MSE}_{u,\mathrm{rand}}/(L+1)} \qquad \in [-\infty, 1] \tag{3.19}$$

and takes values between 1 (perfect forecast) and negative infinity, where 0 corresponds to random forecast skill. Casati (2010) explains that for a collection of forecast-observation frames, the aggregated $\mathrm{MSE}_{u,l}$ is obtained simply by averaging $\mathrm{MSE}_{u,l}$ over all frames and $\mathrm{MSE}_{u,\mathrm{rand}}$ is obtained from the aggregated r and s over all frames. These aggregated MSE scores are then used to compute the aggregated skill score $\mathrm{ISS}_{u,l}$. Interestingly, Casati (2010) shows that the average of the scale components of the IS skill score is equal to the Heidke skill score (HSS). The ISS skill score can thus be interpreted as a scale decomposition of the HSS.

Chapter 4

Results

In this section the results of the thesis are presented. Section 4.1 presents the comparison of the different ConvLSTM models on the extreme/non-extreme threshold of 2.7. The results indicate the 4-layered model trained with W-MAE loss to be the preferable model for the forecasting of extreme winds due to its superior SEDI and PSS scores. A thorough spatio-temporal forecast verification for all intensity thresholds was conducted for the 4-layered model with W-MAE loss in Section 4.2. The forecast verification is summarised in Table 4.1, which displays, for each threshold, the maximum lead-time up until which the forecasts can be considered sufficiently skillful and the minimum spatial scale at which the forecast must be displayed at that lead-time to ensure sufficient skill, as determined from the FSS scores. The table furthermore displays each of the remaining scores at the respective lead-time and scale.

4.1 Model Comparison

4.1.1 Loss

By utilising the spatial verification strategy as described in the methodology we can obtain a spatio-temporal picture of forecast quality. In order to provide an oversee-able comparison of the different ConvLSTM models investigated in this thesis, we compare the models only for the threshold of 2.7 (the 99th percentile of the standardised wind speed distribution), defined as the threshold between extreme and non-extreme winds in this thesis. The results are plotted over spatial scale (in km) for lead-times 3, 6, 9 and 12h and are shown as the mean ± one standard deviation of the four-fold cross validation, computed on the test set. Figure 4.1 compares scores between the ConvLSTM model trained with different loss functions, while Figure 4.2 compares scores between the ConvLSTM model with varying numbers of layers. The compared scores are the symmetric extremal dependence score (SEDI), Peirce skill score (PSS), Clayton skill score (CSS), fractions skill score (FSS) and the intensity-scale score (ISS). We remind the reader that between the PSS and the CSS we prioritise the maximization of the PSS, although both scores are shown for completion, and that high scores on fine scales are preferred over high scores on coarse scales. Lastly, we note that we use 'short lead-times' here to refer to lead-times 3 and 6h and 'long lead-times' to refer to 9 and 12h.

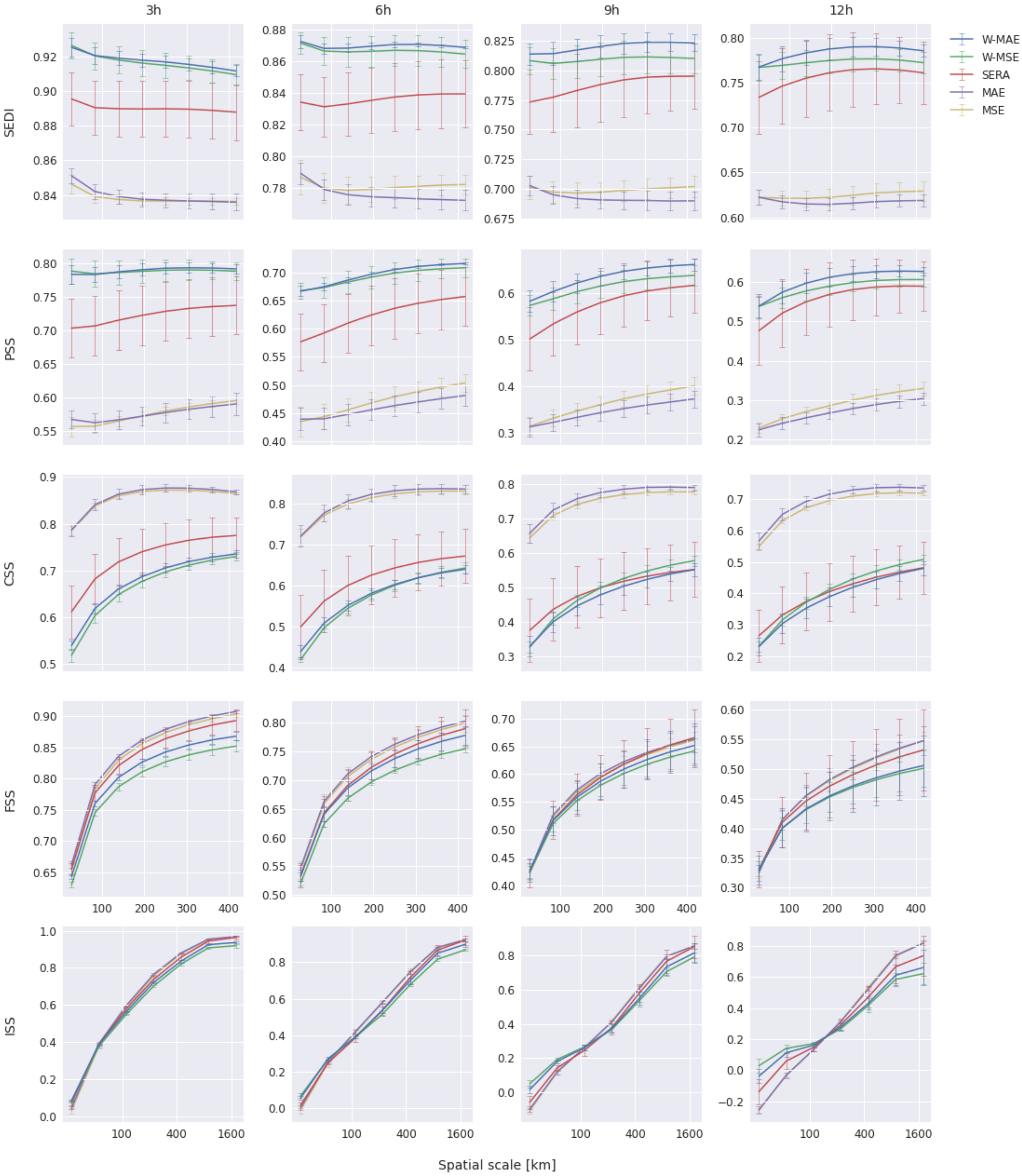

Figure 4.1: The scores obtained for the 3-layered ConvLSTM model trained either with W-MAE (blue), W-MSE (green), SERA (red), standard MAE (purple) or MSE (yellow) loss. Results are shown as the mean ± one standard deviation from the four-fold cross validation and are plotted against spatial scale (in km).

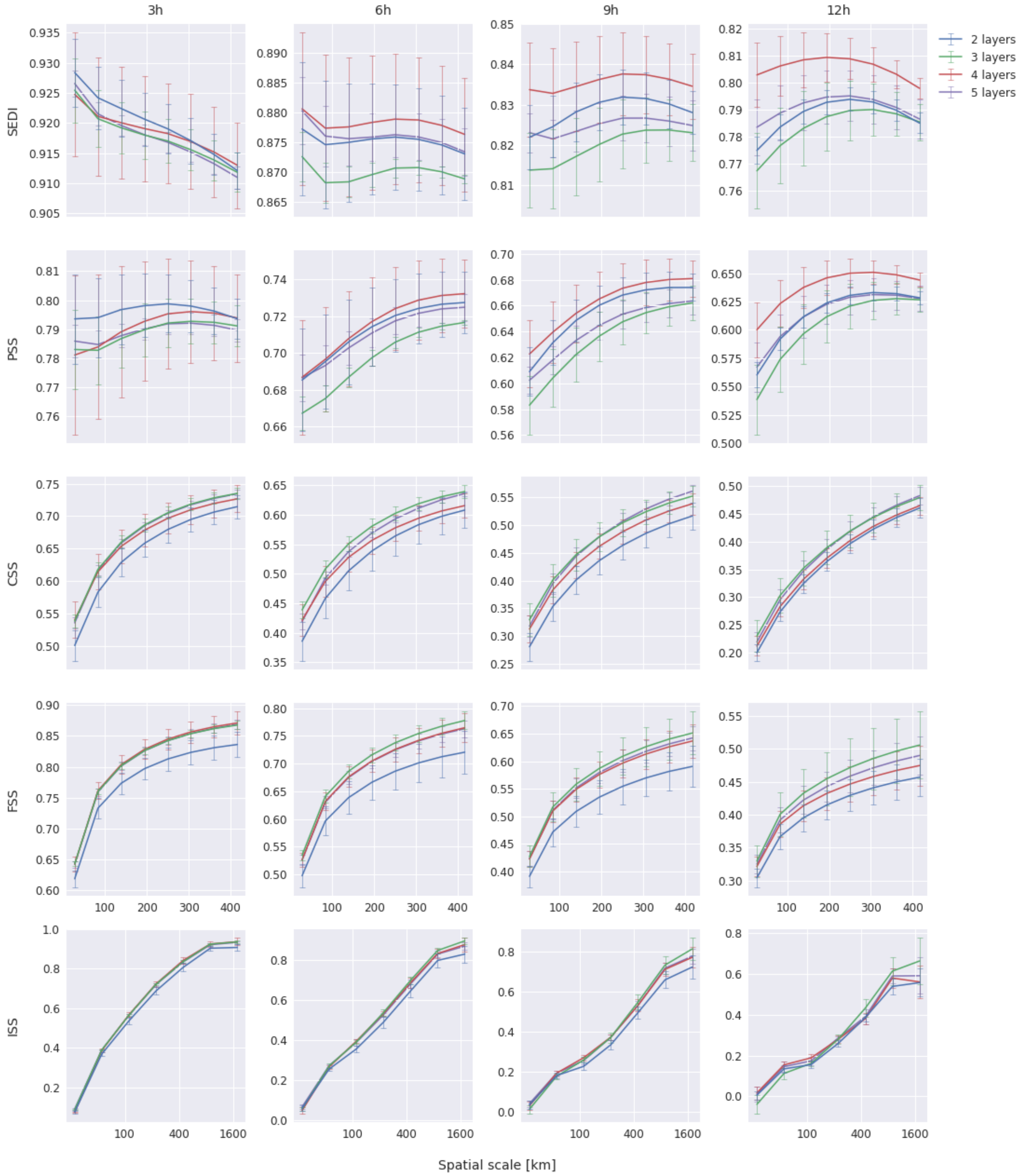

Figure 4.2: The scores obtained for the ConvLSTM model trained with W-MAE loss with varying numbers of layers between 2-5. Results are shown as the mean ± one standard deviation from the four-fold cross validation and are plotted against spatial scale (in km).

Figure 4.1 shows the results for the ConvLSTM model trained either with weighted mean absolute error (W-MAE), weighted mean squared error (W-MSE), squared error-relevance area (SERA), standard MAE or standard MSE loss. The latter two loss functions are provided as a base-line. The ConvLSTM model used for the comparison was chosen to be 3-layered. The results show a general tendency to increase for larger spatial scales, as it becomes easier for coarser predictions to be correct, and to decrease with lead-time. An exception to the former observation is the SEDI on lead-time 3h, where all models scored highest on the lowest spatial scale. While this trend vanishes for most of the models on longer lead-times, it persists particularly long for the standard MAE up until lead-time 12h.

The results on the SEDI and PSS show that the three imbalanced regression losses significantly outperform the standard MAE and MSE, regardless of lead-time or spatial scale. Moreover, between the three imbalanced regression losses, the weighted losses show themselves superior to the SERA loss, particularly on fine scales. The substantially larger uncertainty in the results of the SERA loss furthermore reduces its desirability when compared to the smaller uncertainties of the weighted losses. Between the weighted losses, the W-MAE displays a slighly improved SEDI and PSS on longer lead-times and coarser scales beyond approx. 100 km.

On the other hand, the standard MAE and MSE show significantly higher CSS scores than the imbalanced regression losses while the SERA loss shows some improvement over the weighted losses for short lead-times. The CSS scores of the weighted losses are very similar, although the W-MSE shows slight improvements on long lead-times and coarser spatial scales beyond approx. 100 km. The standard MAE and MSE furthermore tend to obtain higher FSS scores than the imbalanced regression losses and the SERA loss also tends to be superior to the weighted losses here, although differences vanish for long lead-times and fine scales. The weighted losses show very similar FSS scores on long lead-times, while the W-MAE is superior on short lead-times. Lastly, the ISS scores are very similar on short lead-times, where-after a gradual division can be observed on very fine as well as very coarse scales. On long lead-times and fine scales the imbalanced regression losses show superior ISS scores, where the weighted losses are superior to the SERA loss and the W-MSE shows some improvement over the W-MAE, while the reverse can be observed on very coarse scales.

Due to the superior SEDI and PSS scores, superior ISS scores on fine scales and long lead-times, while suffering only minimal reductions in CSS and FSS on fine scales and long lead-times, we judge the weighted losses to be superior to the SERA loss for the extreme wind forecasting problem. Finally, we judge the W-MAE to be preferable over the W-MSE due to its improvements of SEDI and PSS on long lead-times and its improvements of FSS on short lead-times, at the cost of minimal reductions in CSS and ISS. By comparing the imbalanced regression losses with the standard MSE directly (see: Figure A.3 in the Appendix), we obtain for the wind intensity threshold of 2.7 (the 99th percentile) the usage of the W-MAE as opposed to the standard MSE improves the SEDI by 0.09 ± 0.01 points and the PSS by 0.24 ± 0.02 points. With these improvements reductions of CSS of -0.26 ± 0.05 points, of FSS of -0.06 ± 0.03 and ISS of -0.04 ± 0.06 must be taken into account, however. The improvements of the W-MAE over the standard MSE furthermore appear to increase with increasing thresholds. For the threshold of 4.0 (the 999th percentile), the usage of the W-MAE loss improves the SEDI by 0.11 ± 0.02 and the PSS by 0.30 ± 0.05 while reductions in the FSS and ISS are effectively eliminated. Large reductions in the CSS appear unavoidable, however.

4.1.2 Layers

Secondly, Figure 4.2 shows the results for the ConvLSTM model trained with W-MAE loss with varying numbers of layers between 2-5. Analogously to the previous analysis, the figure indicates superior SEDI and PSS on long lead-times for the 4-layered model, while the model shows only moderate reductions in the CSS, FSS and ISS scores on long lead-times and large spatial scales as compared to the 3- and 5-layered models. The 4-layered ConvLSTM model with W-MAE loss was therefore judged the top model for the forecasting of extreme wind speeds from the experiment. A thorough spatio-temporal verification of this model is provided in the next section.

4.2 Top Model Verification

In the previous section the 4-layered ConvLSTM model trained with W-MAE loss was determined to be the preferable model for the forecasting of extreme winds. Before we proceed to present the spatio-temporal forecast verification of this model, we provide three additional figures regarding the model training, the target and forecast distributions and a visualisation of the forecasts. Figure 4.3 presents the loss curves of the model on the training and the validation set for each of the four training-validation splits of the four-fold cross validation. The loss curves show the steady decreasing of the training-loss and the plateau reached by the validation-loss. The dotted vertical lines mark the early-stopping points where the models were saved.

Figure 4.4 provides a comparison of the distributions of the model predictions and the observations i.e. targets. The predicted distribution is shown as the mean $\pm$ one standard deviation from the four-fold cross validation. The figure shows a decent agreement between the two distributions, especially on the hard-to-model right tail. We do observe some undercasting on the left tail (between -3 and -1 σ) and overcasting on the right tail (between 1 and 5 σ) of the distribution, respectively. The model thus exhibits a bias towards the right end of the distribution i.e. high (standardised) wind speeds. This is in accordance with the higher weighting of the right tail of the distribution in the weighted loss function, as opposed to the unit weighting on the left tail.

Figure 4.5 provides a visualisation of the forecasts using five selected input-target-forecast triplets, served to give an impression of the forecasts delivered by the model. For each triplet the first row presents the 12 input frames, the second row presents the 12 target (observation) frames and the third row presents the 12 forecast frames. The frames are plotted as heat-maps to visualise the standardised wind speed from blue (-2.0) to red (4.0). From top-to-bottom, the first two triplets provide examples of cases where the model captured an oncoming intensification well, the third and fourth triples provide examples of cases where the model correctly forecast a dissipation of high wind speeds, while the fifth triplet provides an example of where the model incorrectly forecast a region of intensifying high winds in the south-west of the grid.

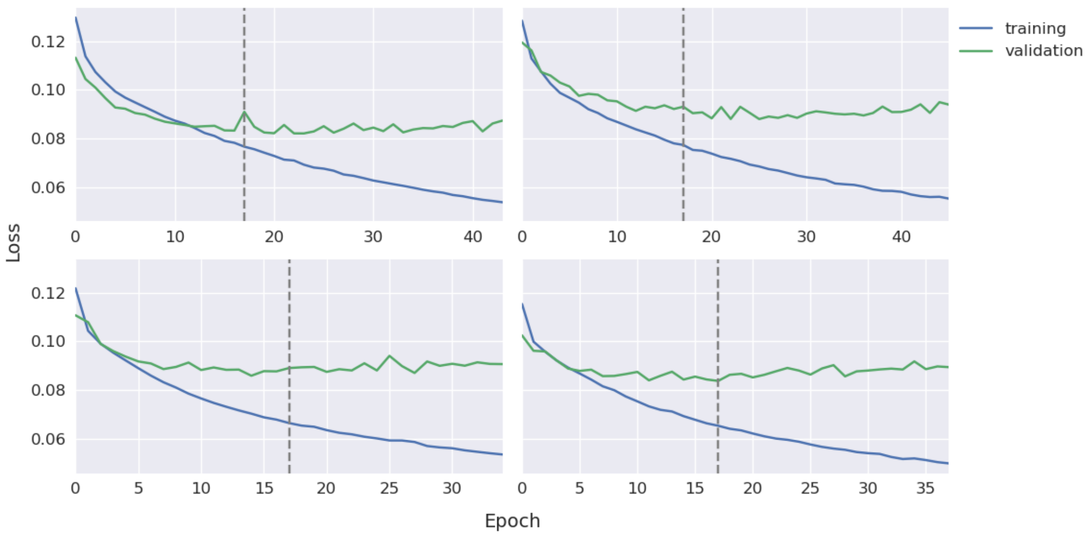

Figure 4.3: Loss curves of the 4-layered ConvLSTM model trained with W-MAE loss on the training and the validation data, for each of the four training-validation splits of the cross validation. The dotted lines designate the points of minimum validation loss and thus the 'early-stopping' points where the models were saved.

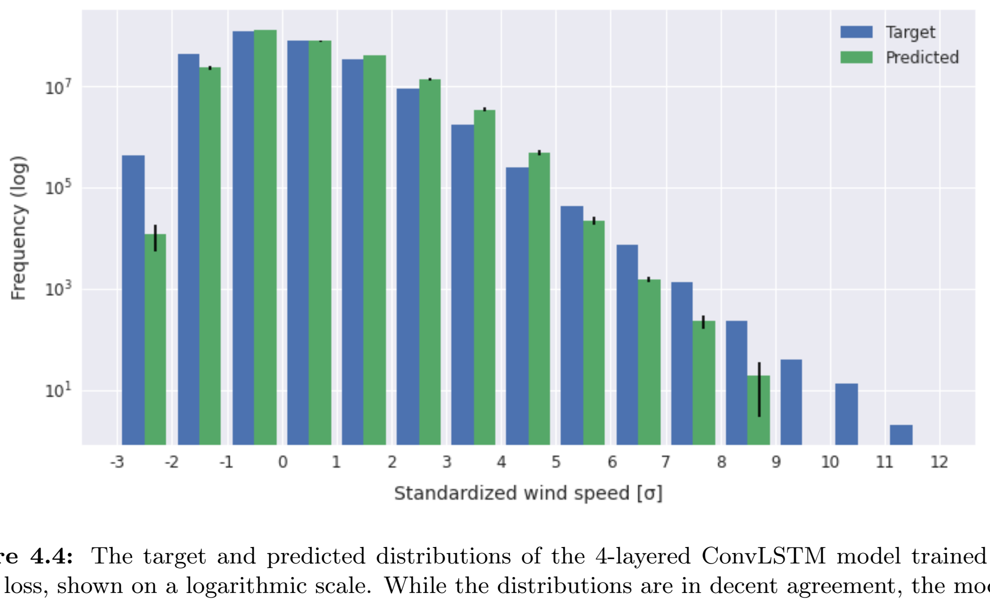

Figure 4.4: The target and predicted distributions of the 4-layered ConvLSTM model trained with W-MAE loss, shown on a logarithmic scale. While the distributions are in decent agreement, the model shows some undercasting on the left tail and overcasting on the right tail of the distribution.

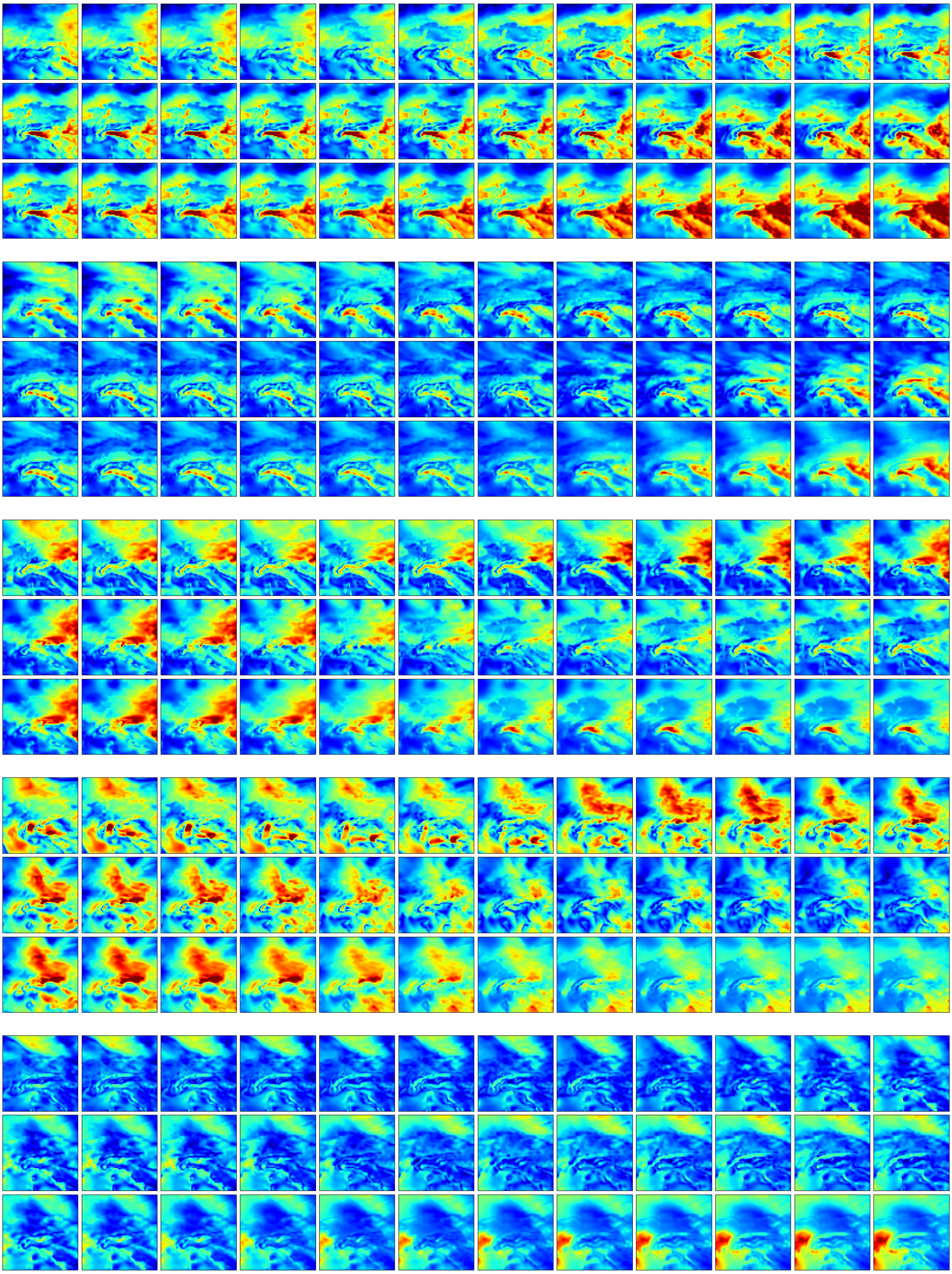

Figure 4.5: Visualisation of the forecasts, as obtained from the 4-layered ConvLSTM model trained with W-MAE loss. The figure shows five selected triples of input (first row), target (second row) and forecast (third row), consisting of 12 frames each. The frames are plotted as heat-maps to visualise the standardised wind speed from blue (-2.0) to red (4.0).

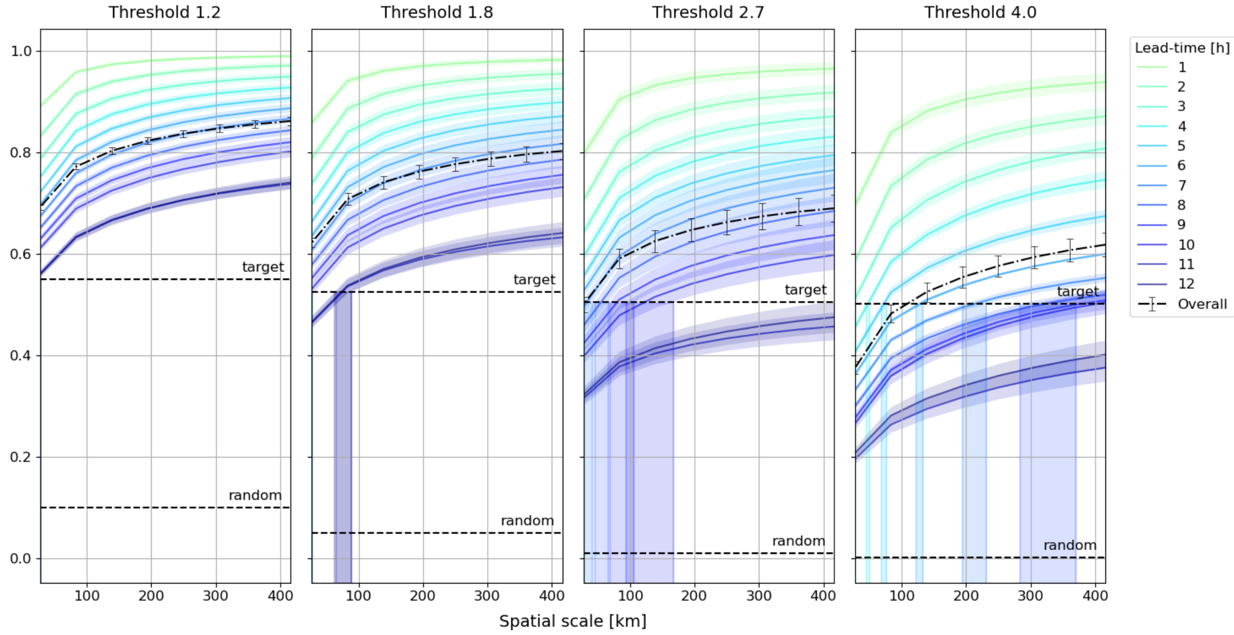

Figure 4.6: The obtained FSS scores as plotted for each threshold and each lead-time against spatial scale (in km). Plotted as dotted lines are the scores as expected for a random forecast ('random') and for a sufficiently skillful forecast ('target'). Comparison of the FSS scores for each lead-time with the 'target' score allows us to determine the minimum spatial scale with which the forecasts must be used in order to be sufficiently skillful on that lead-time.

We proceed to present the spatio-temporal forecast verification of the 4-layered W-MAE model on all four wind intensity thresholds. Following the verification strategy of Roberts (2008), we use the FSS to determine, for each lead-time, the minimum spatial scale at which the forecast can be deemed 'sufficiently skillful'. This is done by plotting the FSS scores for each lead-time against spatial scale and determining the scale at which the scores surpass 'target' skill. The target skill is computed as halfway between random and perfect (unit) skill (more details on this are provided in the methodology, Section 3.3.2). The results of this strategy are shown in Figure 4.6. The figure shows the obtained FSS scores for each intensity threshold, plotted over spatial scale (in km). Each lead-time (in h) is plotted individually and color-coded from green (shorter) to blue (longer) and compared against 'target' skill. The 'overall' scores (black, dash-dotted) show the results as computed independently of lead-time. The results are shown as the mean $\pm$ one standard deviation as obtained from the four-fold cross validation. The intersections of the FSS scores with the target line are computed for the scores + and - their respective uncertainties (linearly interpolated) in order to obtain the proper intervals of minimum spatial scale for sufficient skill.

Firstly, the figure shows that the FSS scores increase with spatial scale, decrease with lead-time and decrease with increasing intensity thresholds. With the lowest threshold of 1.2 we can see that the forecasts are sufficiently skillful on the finest scale of 28km for all lead-times. This is no longer the case for threshold 1.8, where lead-times 11 and 12h require a coarse-graining of spatial scale to ca. 70-90km. Independently of lead-time, however, the forecasts with this threshold are 'overall' sufficiently skillful on the finest scale. The extreme/non-extreme threshold of 2.7 shows that coarse-graining is required on lead-times 8, 9 and 10h, while lead-times 11 and 12h can not be considered sufficiently skillful on any of the investigated scales. 'Overall', however, the forecasts with this threshold may still be considered sufficiently skillful on the finest scale. On the threshold 4.0 lead-times 4-8h must be coarse-grained while lead-times 9-12h can not be considered sufficiently skillful for any of the investigated scales. Although the mean FSS scores on lead-times 9 and 10h do surpass target skill, their lower uncertainties do not and thus these lead-times were not considered sufficiently skillful. 'Overall', the forecasts on this threshold require a coarse-graining to ca. 100km in order to be sufficiently skillful.

The intervals of minimum spatial scale for sufficient skill were then compared to the SEDI, PSS, CSS and ISS scores to determine their value ranges within these intervals. This was done by computing the intersections of the interval edges with lower and upper uncertainties of the respective score. An example of this is provided in Figure 4.7. The figure shows the FSS (top) and PSS (bottom) for threshold 2.7 and lead-time 10h and illustrates how the interval edges, as determined from the FSS, were compared to the uncertainty bars of the PSS to determine the value range of the PSS within the interval. The intersections were computed through linear interpolation between the score values. The score on the interval is then recorded as the midpoint of the value range $\pm$ half the length of the range. The thus obtained scores for each lead-time for which the forecast was determined sufficiently skilful, as well as the respective minimum spatial scale are appended in Figure A.5. A complete presentation of the variation of each scores with intensity threshold, lead-time and spatial skill is furthermore appended in Figure A.4.

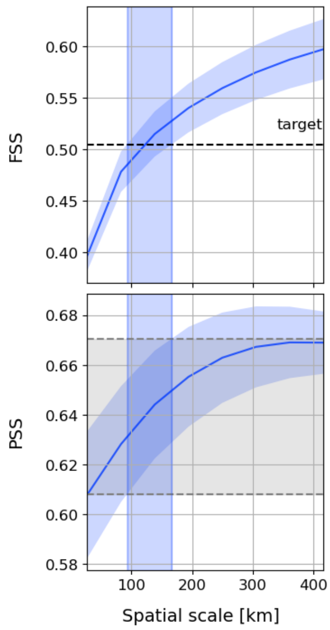

Figure 4.7: Illustration of how the interval of spatial scale for sufficient skill, as determined from the FSS (top), was used to compute the value ranges of the other scores on the respective interval, as illustrated here with the PSS (bottom).

Finally, for each threshold, the maximum lead-time, and the corresponding minimal spatial scale, at which the forecast was determined to be sufficiently skillful are summarised in Table 4.1, along with the score values as determined on these scales. The table shows that when maximum lead-time and minimum spatial scale for sufficient skill are accounted for, performance on the extreme/non-extreme threshold of 2.7 is effectively on par with the lower thresholds. While the forecasting with a threshold of 2.7 thus implies a shortening of maximum lead-time to 10h and a spatial coarse-graining of the forecast to ca. 130km, the forecast can ensure high scoring SEDI and PSS scores, while maintaining CSS and ISS values considerably greater than zero. At a threshold of 4.0 the forecasts must be considerably shortened to 8h and coarse-grained to ca. 330km.

Table 4.1: Summary of spatio-temporal forecast quality. For each intensity threshold, the table presents the maximum lead-time (in h) and the corresponding minimum spatial scale (in km) for which the forecast can be considered sufficiently skillful, as obtained from the FSS scores (Figure 4.6), as well as the obtained scores values at the respective lead-time and spatial scale. Arrows indicate whether higher (↑) or lower (↓) values are superior. The uncertainties of the scores on threshold 1.2 reflect the standard deviations of the four-fold cross validation at the finest scale of 28km, while the remaining uncertainties rather reflect the value range of the scores over the respective spatial interval, as discussed in the text.

Threshold [σ]	1.2	1.8	2.7	4.0
Max. lead-time [h] ↑	12	12	10	8
Min. scale [km] ↓	28	76 ± 13	130 ± 40	330 ± 40
SEDI ↑	0.779 ± 0.005	0.799 ± 0.003	0.833 ± 0.001	0.811 ± 0.003
PSS ↑	0.621 ± 0.006	0.641 ± 0.003	0.662 ± 0.008	0.58 ± 0.01
CSS ↑	0.41 ± 0.01	0.41 ± 0.02	0.42 ± 0.03	0.44 ± 0.01
FSS ↑	0.559 ± 0.07	0.54 ± 0.02	0.53 ± 0.02	0.51 ± 0.01
ISS ↑	0.31 ± 0.02	0.362 ± 0.005	0.28 ± 0.04	0.52 ± 0.05

Chapter 5

Discussion

In Section 4.1 of the results the best ConvLSTM model for the task of spatio-temporal forecasting of extreme winds was determined to be the 4-layered model trained with weighted MAE (W-MAE) loss. The fact that the W-MAE-trained model outperformed the W-MSE-trained model is an interesting result since the MSE loss appears to be the prevalent choice for deep learning applications to wind speed prediction (see: e.g. Marndi et al., 2020; Marzban and Stumpf, 1998; Robert et al., 2013; Wang et al., 2017a; Xie et al., 2021; Zhu et al., 2018). In comparison, we are not aware of any usage of MAE loss for deep learning applications to wind speed prediction (although we note that Amato et al. (2020) used MAE loss for spatio-temporal temperature prediction and Shi et al. (2017) and Wang et al. (2018) proposed the usage of a balanced MAE-MSE loss for wind speed and video frame prediction, respectively). We expect that the W-MSE was outperformed in our work due to a combination of the MSE's sensitivity to outliers and the weights used for weighing the loss function. While the sensitivity of the MSE to outliers should be a favourable property for the prediction of extreme events, we expect that the combination of this property with the heavy weighing of the loss function on the right tail of the data distribution may have led the model to be penalised too heavily for mistakes on wind speed values much larger than necessary for the task at hand and thus impede learning on the actual range of interest (i.e. between thresholds 1.2 and 4.0). When employing the weighted MSE loss for imbalanced regression purposes it may therefore be worthwhile to experiment with weaker weighing methods than the inversely weighted approach followed here (e.g. linear or quadratic weighing). We do note that the standard MSE was indeed observed to be slightly superior to the standard MAE.

The results demonstrated how the performance of a spatio-temporal forecasting model can be improved on imbalanced regression tasks by combining the model with weighted- or SERA loss. It is peculiar, however, that the standard MAE and MSE-trained models attained significantly higher CSS scores than the W-MAE and W-MSE-trained models. We suspect that the reason for this is due to an undercasting bias of events and overcasting bias of non-events which we observe from the respective predicted distributions of these models (appended in the appendix, in Figure A.1 and Figure A.2, respectively). We can see that the forecasts on the right tails of the distributions are consistently undercasted, as opposed to the overcasting that we observed for the W-MAE (Figure 4.4). When a threshold is set on the right tail of the distribution most data-points are, indeed, non-events and thus undercasting events and overcasting non-events may well lead to an enormous reduction in false alarms and thus improve the CSS considerably (due to its similarity to the precision for rare events, see Equation 3.14). As the PSS tells us, however (due to its similarity to the recall for extreme events, see Equation 3.13), the models did an accordingly poor job at forecasting actual extreme events correctly. This is an unexpected result since the CSS is a truly equitable skill score and should thus reflect the performance of a more naive model with a poorer score. We expect that this behaviour is lost due to score degeneracy related to the severe rarity of events.

It is also interesting to note that between the differently layered W-MAE models, it appears that there is no trivial relationship between network depth and forecast quality. From the SEDI and PSS scores in Figure 4.2 on long lead-times (9 and 12h) and fine scales we obtain, in order of superior performance, first the 4-layered, 5-layered, 2-layered and lastly the 3-layered network architecture. This result differs from the observation made in the work of Shi et al. (2015), where deeper models were observed to be superior to shallower models. We note, however, that Shi et al., 2015 based this assessment on the value of the loss function on the validation set after training, while we base our assessment on an extensive spatio-temporal forecast verification. When we compare solely the loss values on the validation set after training we obtain in order of minimal loss first the 2-layered, 4-layered, 5-layered, and lastly the 3-layered model, which still does not suggest a trivial relationship between network depth and forecast quality.

In Section 4.2 of the results a thorough spatio-temporal forecast verification was presented on the 4-layered ConvLSTM model trained with W-MAE loss. Our strategy utilised the FSS to determine for each lead-time the minimum spatial scale at which the forecast can be judged to be sufficiently skillful. We proceeded to present all scores on each lead-time and respective minimum spatial scale for sufficient skill in Figure A.5 and finally created a summarised picture of spatio-temporal forecast quality in Table 4.1. During the verification process we observed a surprising mismatch between the ISS and the PSS, CSS and FSS. Looking at Figure A.5, for example, we can see the ISS scores approach zero at the threshold 2.7σ for lead-times 3-6h and even take on negative values at the threshold 4.0σ for lead-times 1-4h, which appears to suggest that forecasts on these lead-times and scales are effectively as good as random. This is in severe contrast with the fact that forecasts on these lead-times were evidently found to be sufficiently skillful by the FSS. Furthermore, the PSS and CSS should reflect a random forecast with zero skill, as both scores are truly equitable skill scores, while the figure shows both scores to be considerably greater than zero at all times. Although the ISS can be interpreted as a scale-decomposition of the Heidke skill score (HSS), which is a truly equitable skill score, it appears that this does not hold true for the ISS itself.

To investigate this further we provide in Figure A.7 a comparison, for each intensity threshold, of the ISS scores of the 4-layered ConvLSTM model trained with W-MAE loss with the ISS scores obtained by a 'null model' that forecasts 'no-event' at all coordinates and times. Since such a null model forecasts no events, its hits a and false alarms b are zero and thus it is easy to show that such a model scores zero PSS and zero FSS, while the SEDI and CSS encounter division by zero and are thus undefined. From Figure A.7 we observe the ISS scores of the null model are only approximately zero for the lower thresholds of 1.2 and 1.8 and on spatial scales below ca. 400 km. On threshold 2.7 and 4.0, however, the figure shows that a null model can obtain ISS scores significantly greater than zero on coarse spatial scales beyond ca. 400 km. While the figure does validate that the 4-layered ConvLSTM model trained with W-MAE loss is significantly better than a simple null model, the highly non-equitable behaviour of the ISS score on high intensity thresholds does suggest that the score may be hedged by a naive null model for sufficiently rare events, similarly to the accuracy score in the example of Finley's tornado forecasts back in Section 2.2. This is an important result as it helps shed light on the mismatch of the ISS and the other scores employed in this thesis. We do remind the reader, however, that, as a scale-separation method the intensity-scale method provides a very different spatial assessment of the forecasts than the other scores that were employed and its ability to distinguish which features were best captured by the forecasts at different spatial scales certainly remains useful. Thus, while the IS score can certainly be an informative tool for the assessment of spatial forecasts, it appears that caution must be taken when using the score for exceedingly high thresholds i.e. exceedingly rare events.

Returning once more to the PSS and CSS, we provide here a short discussion on the scores' relation to economic value. The PSS and CSS scores obtained in this work are consistently and considerably larger than zero for all investigated intensity thresholds, spatial scales and lead-times. Disregarding any degeneracy these scores may have suffered due to the rarity of events at high thresholds, this suggests that the forecasts from the 4-layered ConvLSTM model trained with W-MAE loss have potential economic value on all intensity thresholds, spatial scales and lead-times. It appears worthwhile, however, to ask to what extent the realisation of this potential economic value is actually economically feasible. For example, the simple cost-loss model (see: Section 2.2.2) to which the scores are related must assume the existence of a preparatory cost C smaller than the potentially suffered loss L in case of an extreme event i.e. $\alpha = C/L < 1$. If such an α does not exist, i.e. the cost is greater than the loss, then it is clearly no longer economically sensible to take any preparatory action whatsoever. While the results show that the PSS and CSS are considerably greater than zero for high thresholds and coarse spatial scales, it is another question entirely whether the preparatory cost on such coarse scales can realistically be smaller than a suffered loss so dispersed over a large region. Indeed, it appears highly unlikely that there exists a preparatory cost on the full scale of the grid (approx. 1800×1800 km^2) that is smaller than the suffered loss of a number of localised wind-extremes across such a vast region. Thus, while a CSS greater than zero implies that there exists a range of α over which the value-curve is positive, where the maximum of that curve (equal to the PSS) may well be substantially larger than zero, we caution that it may be highly unrealistic to assume the existence of a cost C and loss L for the cost-loss ratio α to be within the respective range.

We proceed to make a note on the methodology of our work. Upon a closer inspection of the actual forecast results (examples of which were presented in Figure 4.5 of the results) it was noticed that many of the input-target samples containing extreme winds (in the locally standardised sense) are samples where extreme winds are present in both the input as well as in the target. Examples where there are no extremes present in the input, but the target is showing onsets of extremes, are disproportionately rare in the data although they clearly represent a much more interesting situation (e.g. for early-warning systems of extreme events). The reason for this imbalance appears to be that instances of extreme winds in the data are typically grouped together over a period of a couple of days to about a week and therefore most of the instances of extreme winds in the 12 hour input and target samples are sampled from within the period as opposed to at its onset. Our model therefore appears to be particularly good at the forecasting of the *continuation* of an already-occurring period of extreme winds, while it appears less adapt at correctly forecasting its *onset* (although these observations beg a further, more quantitative analysis). A continuation of this work may, therefore, involve shifting the focus of model learning to the onsets of periods of extreme winds, rather than employing all available instances as was done in this work. It may be worthwhile, to this end, to change the model into a nowcasting model. The WMO (2017) defines nowcasting as the "forecasting with local detail [...] over a period from the present to 6 hours ahead, including a detailed description of the present weather" and states that nowcasting is "particularly concerned with providing accurate warnings and watches" of hazardous extreme events. This would entail reducing the lead-time of the model to below 6 hours while increasing the temporal and spatial resolutions of the data. The local detail of the data may also be enhanced by utilising more precise ground data than raster data from satellites, as recommended by Amato et al. (2020).

This work may, furthermore, be extended by taking a multi-variate approach to wind speed forecasting whereby other atmospheric variables are included into the input of the model, which is an approach that has already been pursued by other researchers (see: e.g. Marndi et al., 2020; Racah et al., 2017; Xie et al., 2021). Marndi et al. (2020) suggest the utilisation of temperature, humidity and pressure into the forecasting task as these have been found to be "significantly more important than other [atmospheric variables]" - a result based on the work done by Cadenas et al. (2016). Xie et al. (2021) use these same three variables, as well as the one-hour minimum and maximum temperature, while Racah et al., 2017 use a much larger set of 16 atmospheric variables, albeit for the classification of large-scale extreme weather events and not for regression of wind speed. Preliminary experiments that we have conducted appear to support the significance of temperature in the forecasting of extreme winds. It may also be worthwhile to consider other atmospheric variables such as the convective available potential energy (CAPE) and deep-layer wind shear (DLS) due to their strong correlation with severe convective storm activity such as the occurrence of thunderstorms and supercells (see: e.g. Chavas and Dawson II, 2021; Rädler et al., 2015; Tsonevsky et al., 2018).

Chapter 6

Conclusion

In this thesis we explored a deep learning approach to the task of spatio-temporal prediction of wind speed extremes in the short-to-medium range. To this end, we investigated the application of a multi-layered convolutional long short-term memory (ConvLSTM) network. We adapted the encoding-forecasting ConvLSTM model proposed by Shi et al. (2015) to imbalanced spatio-temporal regression by training the model with three loss functions from the literature on imbalanced regression: the inversely weighted mean absolute error (W-MAE) loss, the inversely weighted mean squared error (W-MSE) loss and the squared error-relevance area (SERA) loss. Our model provides multi-frame forecasts of horizontal wind speed over Europe at 1000 hPa with a 12 hour lead-time and in 1 hour intervals, using the preceding 12 hours as input. The model was trained and tested on reanalysis wind speed data obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF). By standardising the data based on the local wind speed distributions at each coordinate we focused the definition of an extreme event on its relative rarity rather than its absolute severity.

The forecast quality of the model was assessed using a spatio-temporal forecast verification strategy based on the state of the art of spatial forecast verification methods. In order to provide a varied picture of forecast quality we presented the results of five selected scores over lead-time, various spatial scales and four wind intensity thresholds used to separate extreme events from non-events. We decided to prioritise the symmetric extremal dependence index (SEDI) and the Peirce skill score (PSS) as indicators of forecast skill for extreme events, while the fractions skill score (FSS) was used to determine for each intensity threshold the maximum lead-time as well as the minimum spatial scale with which the forecasts may be used to ensure sufficient skill. The Clayton skill score (CSS) and the intensity-scale score (ISS) were included in the analysis for comparative purposes and we do not recommend prioritising these scores as performance measures for the forecasting of extreme events. We argued the prioritisation of the PSS over the CSS two-fold: one, by viewing the scores as different measures of economic value and, two, by showing that for extremely rare events the CSS approximates the precision and the PSS approximates the recall. We view the latter point as a highly relevant result since forecasts of extreme events are typically desired to have high recall at the cost of precision due to the simple fact that it must be assumed (for there to be any point in the forecasting of extreme events) to be costlier to miss the occurrence of an extreme event than to prepare for a false alarm. We advise against the usage of the ISS for the forecasting of extreme events due to its highly non-equitable behaviour in this context, as considered in detail in the discussion.

The results in this thesis indicate superior performance of the 4-layered ConvLSTM model trained with W-MAE loss. The W-MAE loss shows significant improvements over the SERA loss and the standard MAE and MSE on high intensity thresholds while it shows some superiority over the W-MSE loss on long lead-times and coarse spatial scales. Lastly, the forecast verification of the 4-layered ConvLSTM trained with W-MAE loss shows that the performance of the model on the forecasting of extreme events above the 99th percentile is effectively on par with the forecasting of events above the 90th percentile, granted a shortening of maximum lead-time from 12 to 10 hours and a coarse-graining of the spatial scale from 28km to 130km. We conclude that the W-MAE loss provides an effective way to adapt deep learning to the task of imbalanced spatio-temporal regression and its application to the forecasting of extreme wind events in the short-to-medium range. We thereby contribute to the area of deep learning for wind energy applications as well as the area of imbalanced spatio-temporal regression. In addition, we consider the forecast verification strategy employed here to provide a meaningful contribution towards improved forecast verification of spatio-temporal sequence forecasting.

Appendices

A Additional Figures

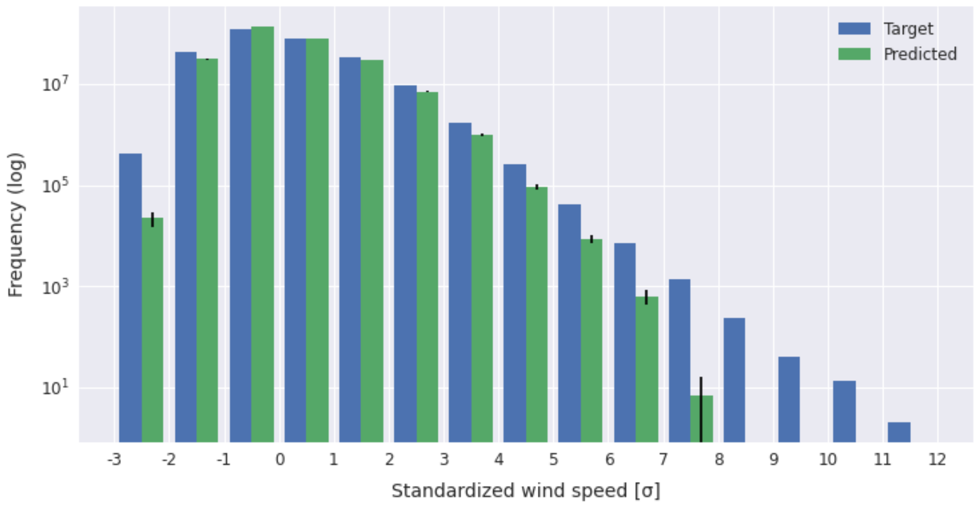

Figure A.1: The target and forecast distributions of the (3-layered) ConvLSTM model trained with standard MAE loss, shown on a logarithmic scale. Results are shown as the mean ± one standard deviation as obtained from the four-fold cross validation.

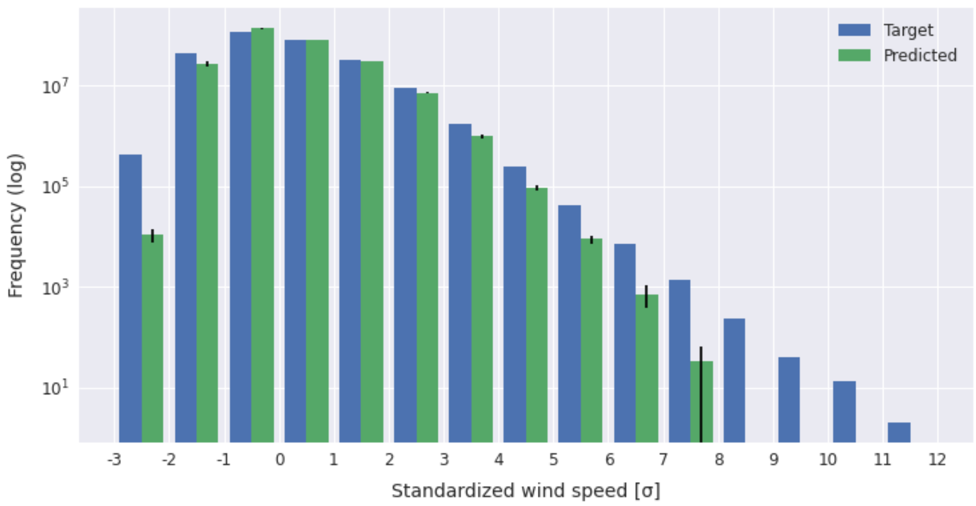

Figure A.2: The target and forecast distributions of the (3-layered) ConvLSTM model trained with standard MSE loss, shown on a logarithmic scale. Results are shown as the mean ± one standard deviation as obtained from the four-fold cross validation.

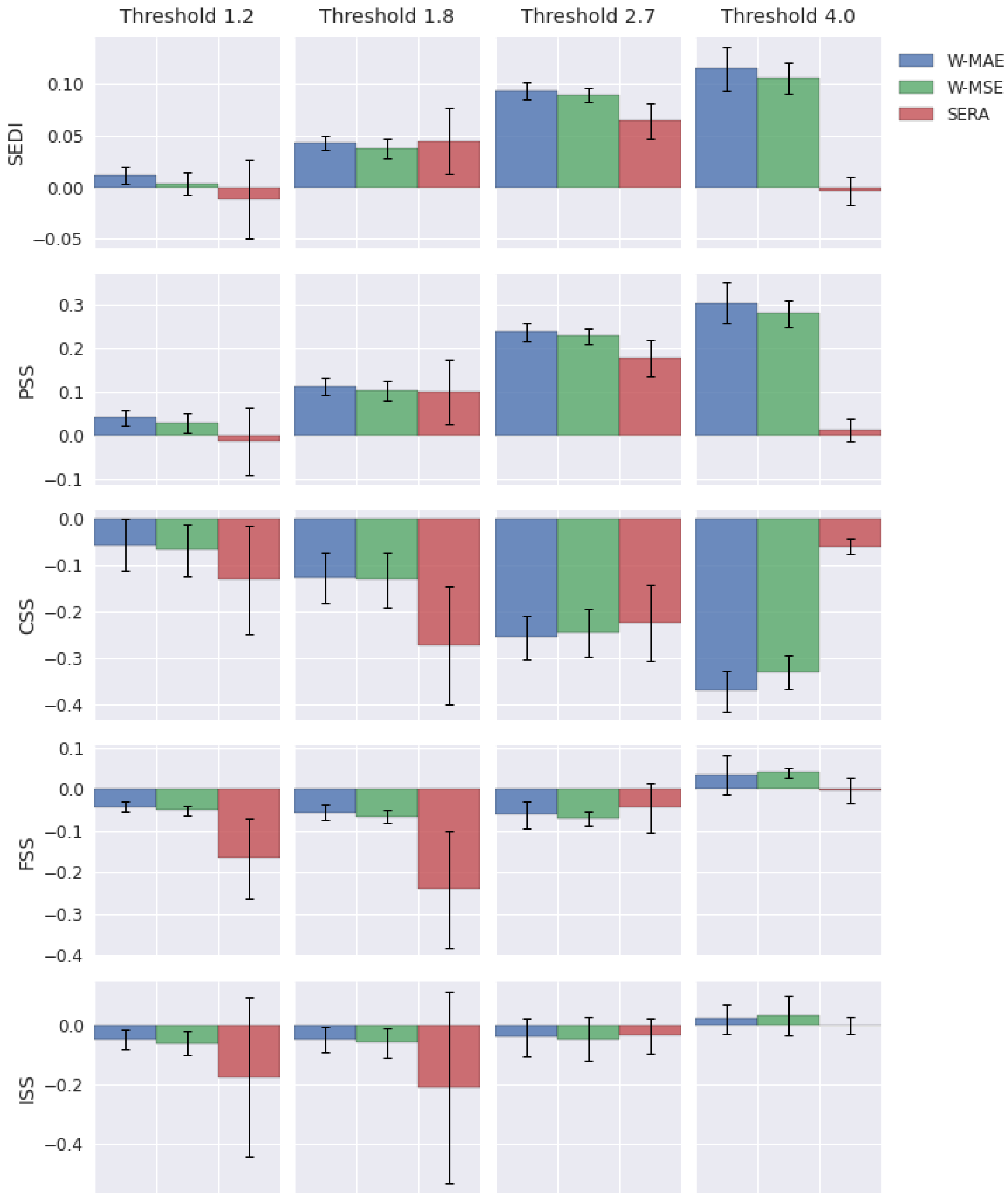

Figure A.3: Differences in scores between each of the imbalanced regression losses investigated in this thesis (W-MAE, W-MSE and SERA) and the standard MSE loss, shown for each wind intensity threshold. Results are shown as the mean of the differences of the scores (computed independent of lead-time) over all respective spatial scales ± one standard deviation.

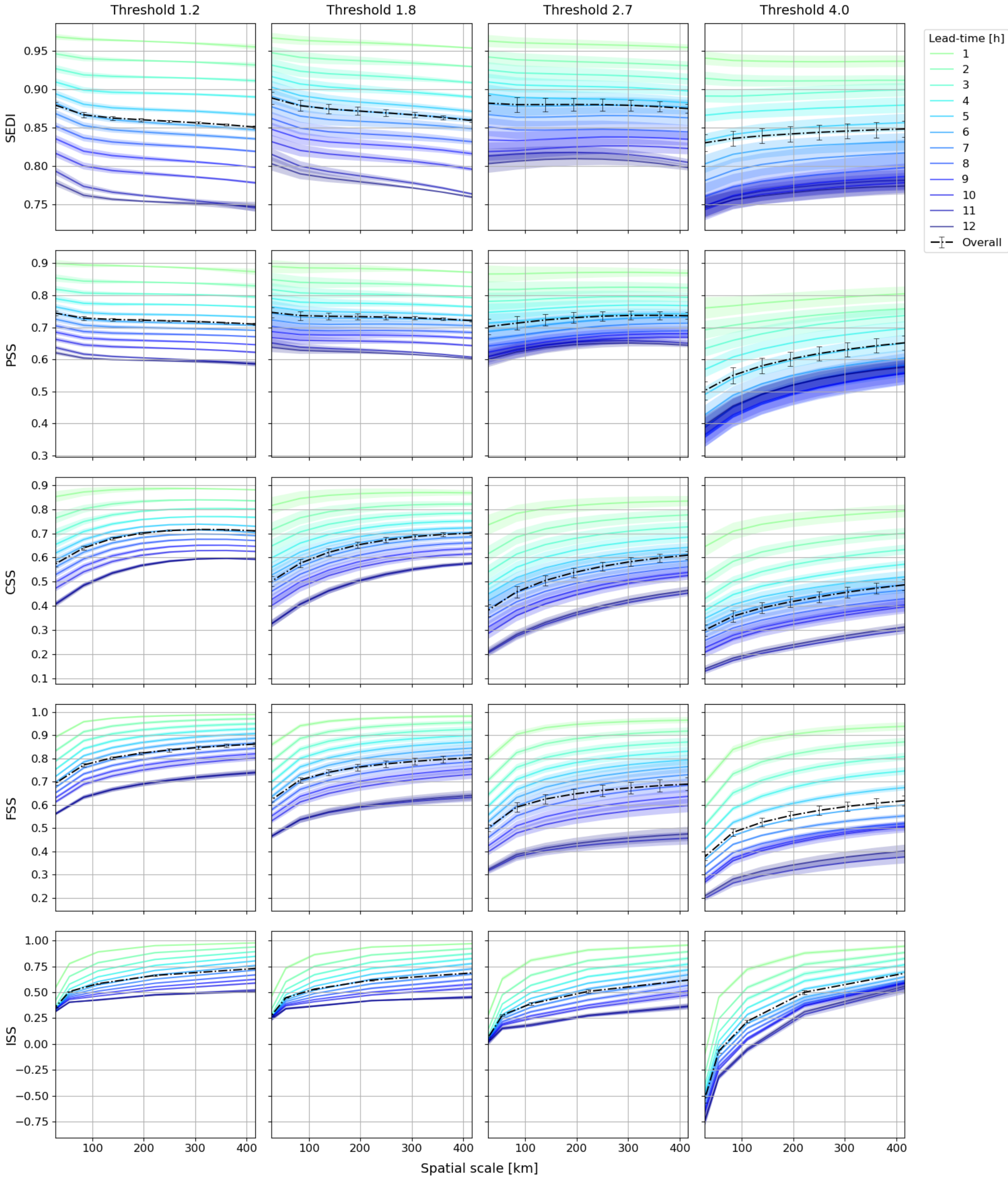

Figure A.4: Forecast verification results for the 4-layered ConvLSTM model trained with W-MAE loss. Scores are displayed over spatial scale (in km) for each wind speed intensity threshold and for each lead-time (in h) separately, color-coded from green (shorter lead-times) to blue (longer lead-times), as well as the 'overall' scores as computed independently of lead-time (black, dash-dotted). Scores are shown as the mean ± one standard deviation as obtained from the four-fold cross validation.

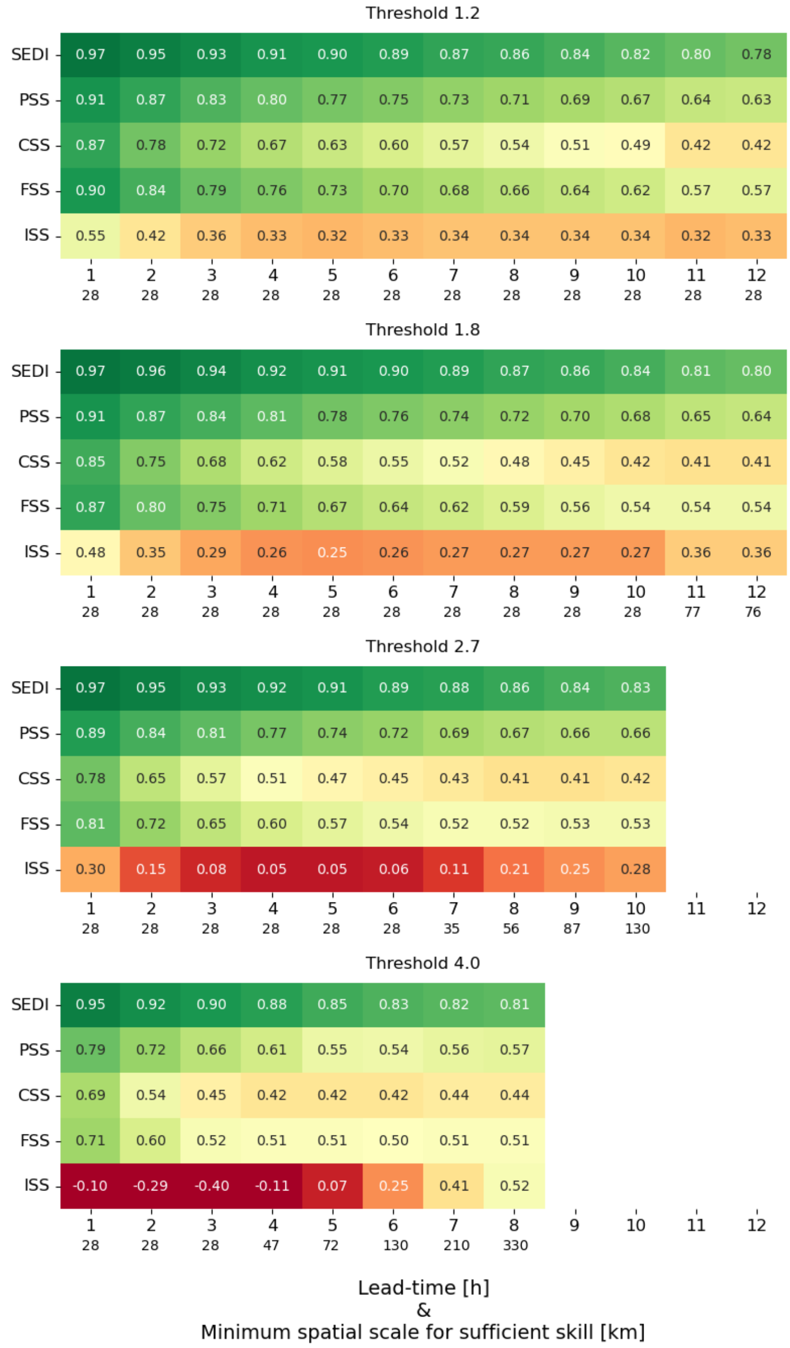

Figure A.5: Comparison of scores against lead-time (in h) for each intensity threshold for the 4-layered ConvLSTM model trained with W-MAE loss. The scores are presented for the minimum spatial scale (in km) for sufficient skill, as determined from the FSS scores in Figure 4.6. No scores are presented for lead-times where the forecasts were determined to be insufficiently skillful for all investigated scales.

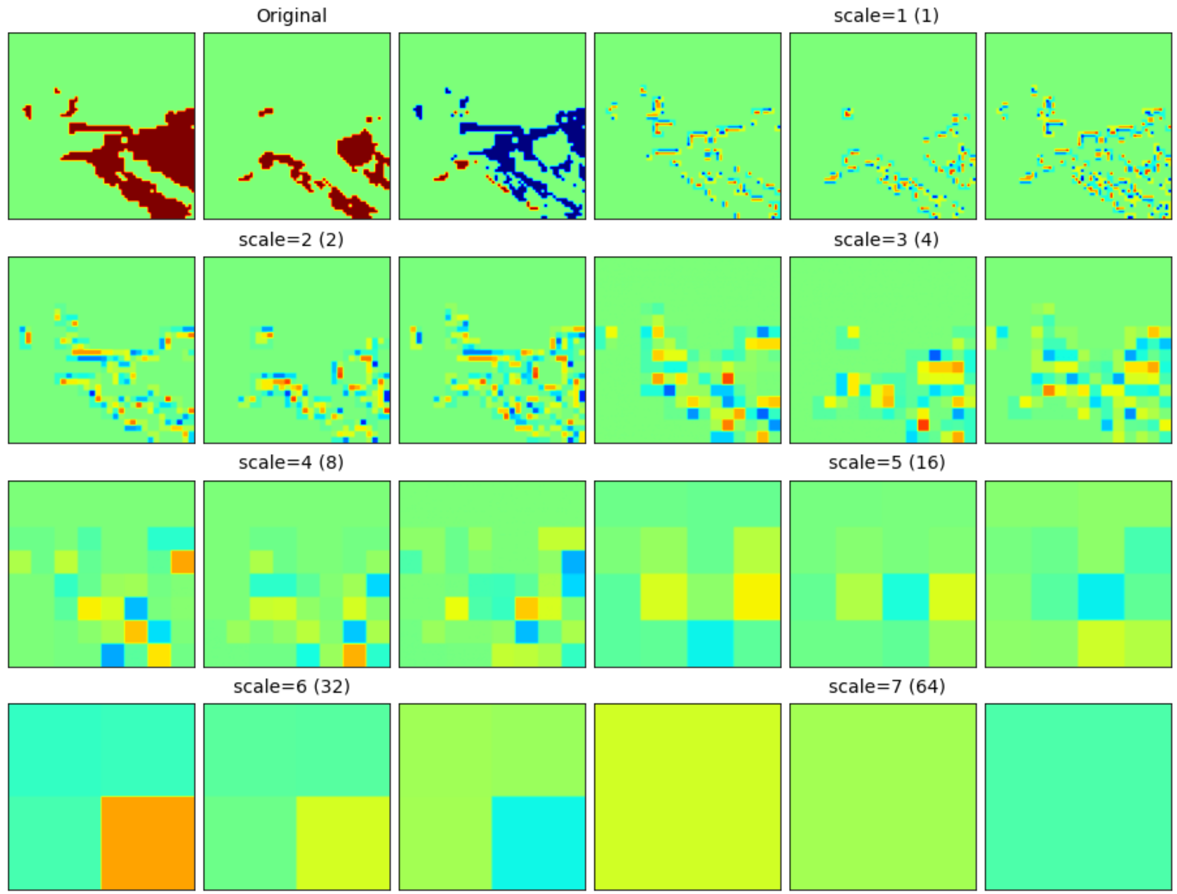

Figure A.6: Scale-decomposition of an example binary forecast and observation frame thresholded at 2.7σ, using the intensity-scale method of Casati et al. (2008). The frames are shown as triplets of the forecast (left), observation (center) and the difference (right). The first triplet in the top left shows the original binary frames. The following triplets, from left-to-right, top-to-bottom show the reconstructions of the frames using the decomposition coefficients from the denoted scale. The number of grid-points corresponding to the scale number are included in brackets.

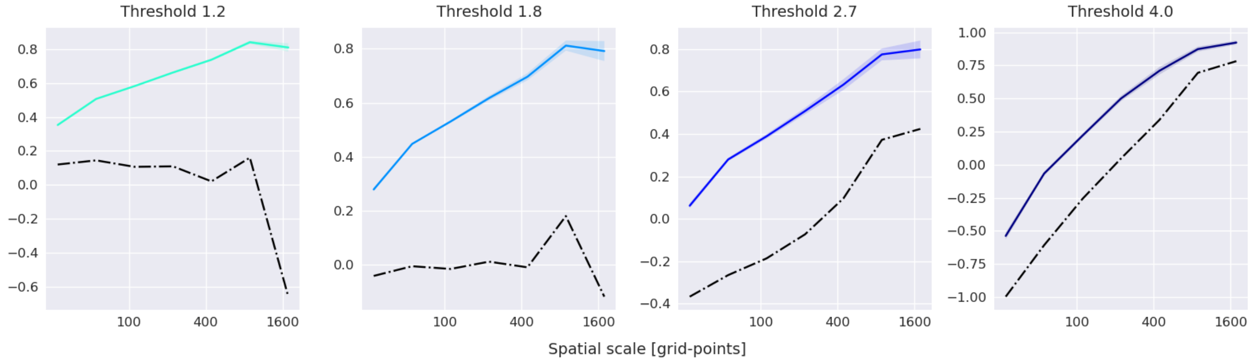

Figure A.7: Comparison of the ISS scores obtained by the 4-layered ConvLSTM model trained with W-MAE loss and a 'null model' (black, dash-dotted) which forecasts 'no event' at all coordinates and times. Scores are lead-time independent and shown over spatial scale (in km) for each wind intensity threshold. Model scores are shown as the mean $\pm$ one standard deviation as obtained from the four-fold cross validation. Although the figure validates that the ConvLSTM model scores significantly better than the null model, the figure also shows that a null model can attain ISS scores substantially greater than zero for high thresholds and coarse scales.

Bibliography

Ahijevych, D., Gilleland, E., Brown, B. G., and Ebert, E. E. (2009). "Application of Spatial Verification Methods to Idealized and NWP-Gridded Precipitation Forecasts". In: *Weather and Forecasting* 24.6, pp. 1485–1497. DOI: 10.1175/2009WAF2222298.1.

Ahmed, M., Maume-Deschamps, V., and Ribereau, P. (2021). *Recognizing a Spatial Extreme dependence structure: A Deep Learning approach.*

Alessandrini, S., Sperati, S., and Pinson, P. (2013). "A comparison between the ECMWF and COSMO Ensemble Prediction Systems applied to short-term wind power forecasting on real data". In: *Applied Energy* 107.C, pp. 271–280. DOI: 10.1016/j.apenergy.2013.0.

Alexander, G. D., Weinman, J. A., Karyampudi, V. M., Olson, W. S., and Lee, A. C. L. (1999). "The Effect of Assimilating Rain Rates Derived from Satellites and Lightning on Forecasts of the 1993 Superstorm". In: *Monthly Weather Review* 127.7, pp. 1433–1457. DOI: 10.1175/1520-0493(1999)127<1433:TEOARR>2.0.CO;2.

Amato, F., Guignard, F., Robert, S., and Kanevski, M. (2020). "A novel framework for spatio-temporal prediction of environmental data using deep learning". In: *Scientific Reports* 10.1, p. 22243. DOI: 10.1038/s41598-020-79148-7.

Atger, F. (2001). "Verification of intense precipitation forecasts from single models and ensemble prediction systems". In: *Nonlinear Processes in Geophysics* 8.6, pp. 401–417. DOI: 10.5194/npg-8-401-2001.

Blonbou, R. (2011). "Very short-term wind power forecasting with neural networks and adaptive Bayesian learning". In: *Renewable Energy* 36.3, pp. 1118–1124. ISSN: 0960-1481. DOI: https://doi.org/10.1016/j.renene.2010.08.026.

Bossanyi, E. and King, J. (2012). "Improving wind farm output predictability by means of a soft cut-out strategy". In: European Wind Energy Conference and Exhibition EWEA.

BP (2021). *Statistical Review of World Energy 2021*. British Petroleum.

Bradley, A. A., Hashino, T., and Schwartz, S. S. (2003). "Distributions-Oriented Verification of Probability Forecasts for Small Data Samples". In: *Weather and Forecasting* 18.5, pp. 903–917. DOI: 10.1175/1520-0434(2003)018<0903:DVOPFF>2.0.CO;2.

Brier, G. W. (1950). "Verification of forecasts expressed in terms of probability". In: *Monthly Weather Review* 78.1, pp. 1–3. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Briggs, W. M. and Levine, R. A. (1997). "Wavelets and Field Forecast Verification". In: *Monthly Weather Review* 125.6, pp. 1329–1341. DOI: 10.1175/1520-0493(1997)125<1329:WAFFV>2.0.CO;2.

Brodersen, K. H., Ong, C. S., Stephan, K., and Buhmann, J. (2010). "The Balanced Accuracy and Its Posterior Distribution". In: *2010 20th International Conference on Pattern Recognition*, pp. 3121–3124.

Burton, T., Sharpe, D., Jenkins, N., and Bossanyi, E. (2001). In: *Wind Engineering* 25.3, pp. 197–199. ISSN: 0309524X, 2048402X.

Cadenas, E., Rivera, W., Campos-Amezcua, R., and Heard, C. (2016). "Wind Speed Prediction Using a Univariate ARIMA Model and a Multivariate NARX Model". In: *Energies* 9.2. ISSN: 1996-1073. DOI: 10.3390/en9020109.

Carpinone, A., Giorgio, M., Langella, R., and Testa, A. (2015). "Markov chain modeling for very-short-term wind power forecasting". In: *Electric Power Systems Research* 122, pp. 152–158. ISSN: 0378-7796. DOI: https://doi.org/10.1016/j.epsr.2014.12.025.

Casati, B. (2010). "New Developments of the Intensity-Scale Technique within the Spatial Verification Methods Intercomparison Project". In: *Weather and Forecasting* 25.1, pp. 113–143. DOI: 10.1175/2009WAF2222257.1.

Casati, B., Ross, G., and Stephenson, D. B. (2004). "A new intensity-scale approach for the verification of spatial precipitation forecasts". In: *Meteorological Applications* 11.2, pp. 141–154. DOI: https://doi.org/10.1017/S1350482704001239.

Casati, B., Wilson, L. J., Stephenson, D. B., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U., Ebert, E. E., Brown, B. G., and Mason, S. (2008). "Forecast verification: current status and future directions". In: *Meteorological Applications* 15.1, pp. 3–18. DOI: https://doi.org/10.1002/met.52.

Castellani, F., Natili, F., Astolfi, D., and Cianetti, F. (2019). "Mechanical behaviour of wind turbines operating above design conditions". In: *Procedia Structural Integrity* 24. AIAS 2019 International Conference on Stress Analysis, pp. 495–509. ISSN: 2452-3216. DOI: https://doi.org/10.1016/j.prostr.2020.02.045.

Celik, A. and Kolhe, M. (Jan. 2013). "Generalized feed-forward based method for wind energy prediction". In: *Applied Energy* 101, pp. 582–588. DOI: 10.1016/j.apenergy.2012.06.040.

Chavas, D. R. and Dawson II, D. T. (2021). "An Idealized Physical Model for the Severe Convective Storm Environmental Sounding". In: *Journal of the Atmospheric Sciences* 78.2, pp. 653–670. DOI: 10.1175/JAS-D-20-0120.1.

Chen, K. and Yu, J. (2014). "Short-term wind speed prediction using an unscented Kalman filter based state-space support vector regression approach". In: *Applied Energy* 113.C, pp. 690–705. DOI: 10.1016/j.apenergy.2013.0.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. (2018). "Neural Ordinary Differential Equations". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Montréal, Canada: Curran Associates Inc., pp. 6572–6583.

Cheng, L., Zang, H., Ding, T., Sun, R., Wang, M., Wei, Z., and Sun, G. (2018). "Ensemble Recurrent Neural Network Based Probabilistic Wind Speed Forecasting Approach". In: *Energies* 11.8. ISSN: 1996-1073. DOI: 10.3390/en11081958.

Cheng, W. Y., Liu, Y., Bourgeois, A. J., Wu, Y., and Haupt, S. E. (2017). "Short-term wind forecast of a data assimilation/weather forecasting system with wind turbine anemometer measurement assimilation". In: *Renewable Energy* 107, pp. 340–351. ISSN: 0960-1481. DOI: https://doi.org/10.1016/j.renene.2017.02.014.

Clayton, H. (1934). “Rating weather forecasts”. In: *American Meteorological Society* 15, pp. 279, 283.

Costa, A., Crespo, A., Navarro, J., Lizcano, G., Madsen, H., and Feitosa, E. (2008). “A review on the young history of the wind power short-term prediction”. In: *Renewable and Sustainable Energy Reviews* 12.6, pp. 1725–1744. ISSN: 1364-0321. DOI: https://doi.org/10.1016/j.rser.2007.01.015.

Cutululis, N., Litong-Palima, M., and Sørensen, P. (2012). “Offshore Wind Power Production in Critical Weather Conditions”. English. In: *Proceedings of EWEA 2012 - European Wind Energy Conference & Exhibition*. EWEC 2012 - European Wind Energy Conference &; Exhibition, EWEC 2012 ; Conference date: 16-04-2012 Through 19-04-2012. European Wind Energy Association (EWEA).

Dabernig, M., Mayr, G. J., Messner, J. W., and Zeileis, A. (2017). “Spatial ensemble post-processing with standardized anomalies”. In: *Quarterly Journal of the Royal Meteorological Society* 143.703, pp. 909–916. DOI: https://doi.org/10.1002/qj.2975.

Damrath, U. (Sept. 2004). “Verification against precipitation observations of a high density network - what did we learn?” In: *International Verification Methods Workshop*.

Darwish, A. S. and Al-Dabbagh, R. (2020). “Wind energy state of the art: present and future technology advancements”. In: *Renew. Energy Environ. Sustain.* 5, p. 7. DOI: 10.1051/rees/2020003.

Davis, C., Brown, B., and Bullock, R. (2006). “Object-Based Verification of Precipitation Forecasts. Part I: Methodology and Application to Mesoscale Rain Areas”. In: *Monthly Weather Review* 134.7, pp. 1772–1784. DOI: 10.1175/MWR3145.1.

DelSole, T. and Shukla, J. (2006). “Specification of Wintertime North American Surface Temperature”. In: *Journal of Climate* 19.12, pp. 2691–2716. DOI: 10.1175/JCLI3704.1.

Deng, X., Li, W., Liu, X., Guo, Q., and Newsam, S. (2018). “One-class remote sensing classification: one-class vs. binary classifiers”. In: *International Journal of Remote Sensing* 39.6, pp. 1890–1910. DOI: 10.1080/01431161.2017.1416697.

Deppe, A. J., Gallus, W. A., and Takle, E. S. (2013). “A WRF Ensemble for Improved Wind Speed Forecasts at Turbine Height”. In: *Weather and Forecasting* 28.1, pp. 212–228. DOI: 10.1175/WAF-D-11-00112.1.

Dickinson, S. and Brown, R. A. (1996). “A Study of Near-Surface Winds in Marine Cyclones Using Multiple Satellite Sensors”. In: *Journal of Applied Meteorology and Climatology* 35.6, pp. 769–781. DOI: 10.1175/1520-0450(1996)035<0769:ASONSW>2.0.CO;2.

Ding, D., Zhang, M., Pan, X., Yang, M., and He, X. (2019). “Modeling Extreme Events in Time Series Prediction”. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD ’19. Anchorage, AK, USA: Association for Computing Machinery, pp. 1114–1122. ISBN: 9781450362016. DOI: 10.1145/3292500.3330896.

do Nascimento Camelo, H., Lucio, P. S., Junior, J. B. V. L., and de Carvalho, P. C. M. (2018). “A hybrid model based on time series models and neural network for forecasting wind speed in the Brazilian northeast region”. In: *Sustainable Energy Technologies and Assessments* 28, pp. 65–72. ISSN: 2213-1388. DOI: https://doi.org/10.1016/j.seta.2018.06.009.

Dorninger, M., Ghelli, A., and Lerch, S. (2020). “Editorial: Recent developments and application examples on forecast verification”. In: *Meteorological Applications* 27.4, e1934. DOI: https://doi.org/10.1002/met.1934.

Dorninger, M., Gilleland, E., Casati, B., Mittermaier, M. P., Ebert, E. E., Brown, B. G., and Wilson, L. J. (2018). “The Setup of the MesoVICT Project”. In: *Bulletin of the American Meteorological Society* 99.9, pp. 1887–1906. DOI: 10.1175/BAMS-D-17-0164.1.

Ebert, E., Wilson, L., Weigel, A., Mittermaier, M., Nurmi, P., Gill, P., Göber, M., Joslyn, S., Brown, B., Fowler, T., and Watkins, A. (2013). “Progress and challenges in forecast verification”. In: *Meteorological Applications* 20.2, pp. 130–139. DOI: https://doi.org/10.1002/met.1392.

Ebert, E. E. and McBride, J. L. (Dec. 2000). “Verification of precipitation in weather systems: determination of systematic errors”. In: *Journal of Hydrology* 239.1, pp. 179–202. DOI: 10.1016/S0022-1694(00)00343-7.

Ebert, E. (Aug. 2002). “Fuzzy verification: giving partial credit to erroneous forecasts”. In: *NCAR/FAA Verification Workshop: Making Verification More Meaningful*. National Center for Atmospheric Research.

Ebert, E. E. (2008). “Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework”. In: *Meteorological Applications* 15.1, pp. 51–64. DOI: https://doi.org/10.1002/met.25.

— (2009). “Neighborhood Verification: A Strategy for Rewarding Close Forecasts”. In: *Weather and Forecasting* 24.6, pp. 1498–1510. DOI: 10.1175/2009WAF2222251.1.

Erdem, E. and Shi, J. (2011). “ARMA based approaches for forecasting the tuple of wind speed and direction”. In: *Applied Energy* 88.4, pp. 1405–1414. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2010.10.031.

Feng, B. and Fox, G. C. (2021). *Spatiotemporal Pattern Mining for Nowcasting Extreme Earthquakes in Southern California*.

Ferreira, M., Santos, A., and Lucio, P. (2019). “Short-term forecast of wind speed through mathematical models”. In: *Energy Reports* 5, pp. 1172–1184. ISSN: 2352-4847. DOI: https://doi.org/10.1016/j.egyr.2019.05.007.

Ferro, C. A. T. and Stephenson, D. B. (2011). “Extremal Dependence Indices: Improved Verification Measures for Deterministic Forecasts of Rare Binary Events”. In: *Weather and Forecasting* 26.5, pp. 699–713. DOI: 10.1175/WAF-D-10-05030.1.

Finley, J. (1884). “Tornado Prediction”. In: *American Meteorological Journal* 1, pp. 85–88.

Focken, U., Lange, M., and Waldl, H.-P. (2001). “Previento – a wind power prediction system with an innovative upscaling algorithm”. In: *Proceedings of the 2001 European wind energy conference & exhibition*.

Fyrippis, I., Axaopoulos, P. J., and Panayiotou, G. (Feb. 2010). “Wind energy potential assessment in Naxos Island, Greece”. In: *Applied Energy* 87.2, pp. 577–586.

Gan, Z., Li, C., Henao, R., Carlson, D. E., and Carin, L. (2015). “Deep Temporal Sigmoid Belief Networks for Sequence Modeling.” In: *NIPS*. Ed. by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, pp. 2467–2475.

Gandin, L. S. and Murphy, A. H. (1992). “Equitable Skill Scores for Categorical Forecasts”. In: *Monthly Weather Review* 120.2, pp. 361–370. DOI: 10.1175/1520-0493(1992)120<0361:ESSFCF>2.0.CO;2.

Gao, N., Xue, H., Shao, W., Zhao, S., Qin, K. K., Prabowo, A., Rahaman, M. S., and Salim, F. D. (2020). *Generative Adversarial Networks for Spatio-Temporal Data: A Survey.*

Germann, U. and Zawadzki, I. (2004). "Scale Dependence of the Predictability of Precipitation from Continental Radar Images. Part II: Probability Forecasts". In: *Journal of Applied Meteorology* 43.1, pp. 74–89. DOI: 10.1175/1520-0450(2004)043<0074:SDOTPO>2.0.CO;2.

Gerrity, J. P. (1992). "A Note on Gandin and Murphy's Equitable Skill Score". In: *Monthly Weather Review* 120.11, pp. 2709–2712. DOI: 10.1175/1520-0493(1992)120<2709:ANOGAM>2.0.CO;2.

Ghil, M., Yiou, P., Hallegatte, S., Malamud, B., Naveau, P., Soloviev, A., Friederichs, P., Keilis-Borok, V., Kondrashov, D., and Kossobokov, V. (May 2011). "Extreme events: Dynamics, statistics and prediction". In: *Nonlin. Processes Geophys., accepted.* DOI: 10.5194/npg-18-295-2011.

Ghosh, A., Sufian, A., Sultana, F., Chakrabarti, A., and De, D. (Jan. 2020). "Fundamental Concepts of Convolutional Neural Network". In: pp. 519–567. ISBN: 978-3-030-32643-2. DOI: 10.1007/978-3-030-32644-9_36.

Gilbert, G. (1884). "Finley's tornado predictions". In: *American Meteorological Journal* 1, pp. 166, 172.

Gilleland, E., Ahijevych, D., Brown, B., Casati, B., and Ebert, E. (2010a). "Spatial Forecast Verification". In: *ASP Colloquium.* National Center for Atmospheric Research.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., and Ebert, E. E. (2009). "Intercomparison of Spatial Forecast Verification Methods". In: *Weather and Forecasting* 24.5, pp. 1416–1430. DOI: 10.1175/2009WAF2222269.1.

Gilleland, E., Ahijevych, D. A., Brown, B. G., and Ebert, E. E. (2010b). "Verifying Forecasts Spatially". In: *Bulletin of the American Meteorological Society* 91.10, pp. 1365–1376. DOI: 10.1175/2010BAMS2819.1.

Gilleland, E., Lee, T. C. M., Gotway, J. H., Bullock, R. G., and Brown, B. G. (2008). "Computationally Efficient Spatial Forecast Verification Using Baddeley's Delta Image Metric". In: *Monthly Weather Review* 136.5, pp. 1747–1757. DOI: 10.1175/2007MWR2274.1.

Gilleland, E., Skok, G., Brown, B. G., Casati, B., Dorninger, M., Mittermaier, M. P., Roberts, N., and Wilson, L. J. (2020). "A Novel Set of Geometric Verification Test Fields with Application to Distance Measures". In: *Monthly Weather Review* 148.4, pp. 1653–1673. DOI: 10.1175/MWR-D-19-0256.1.

Goldin, D. and Kanellakis, P. C. (1995). "On Similarity Queries for Time-Series Data: Constraint Specification and Implementation". In: pp. 137–153.

Goyal, S., Raghunathan, A., Jain, M., Simhadri, H. V., and Jain, P. (2020). "DROCC: Deep Robust One-Class Classification." In: *CoRR* abs/2002.12718.

Grams, J. S., Gallus, W. A., Koch, S. E., Wharton, L. S., Loughe, A., and Ebert, E. E. (2006). "The Use of a Modified Ebert–McBride Technique to Evaluate Mesoscale Model QPF as a Function of Convective System Morphology during IHOP 2002". In: *Weather and Forecasting* 21.3, pp. 288–306. DOI: 10.1175/WAF918.1.

Graves, A. (2013). "Generating Sequences With Recurrent Neural Networks". In: *ArXiv* abs/1308.0850.

Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S., and Hoefler, T. (2021). "Deep learning for post-processing ensemble weather forecasts". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379.2194, p. 20200092. DOI: 10.1098/rsta.2020.0092.

Heidke, P. (1926). “Berechnung Des Erfolges Und Der Güte Der Windstärkevorhersagen Im Sturmwarnungsdienst”. In: *Geografiska Annaler* 8.4, pp. 301–349. DOI: 10.1080/20014422.1926.11881138.

Hendrycks, D., Mazeika, M., and Dietterich, T. (2019). “Deep Anomaly Detection with Outlier Exposure”. In: *International Conference on Learning Representations.*

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers D. abd Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N. (2018). *ERA5 hourly data on pressure levels from 1979 to present.* (Accessed on 20-AUG-2021). Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: 10.24381/cds.bd0915c6.

Hoffman, R. N., Liu, Z., Louis, J.-F., and Grassoti, C. (1995). “Distortion Representation of Forecast Errors”. In: *Monthly Weather Review* 123.9, pp. 2758–2770. DOI: 10.1175/1520-0493(1995)123<2758:DROFE>2.0.CO;2.

Hogan, R. J., Ferro, C. A. T., Jolliffe, I. T., and Stephenson, D. B. (2010). “Equitability Revisited: Why the “Equitable Threat Score” Is Not Equitable”. In: *Weather and Forecasting* 25.2, pp. 710–726. DOI: 10.1175/2009WAF2222350.1.

Hornik, K., Stinchcombe, M., and White, H. (1989). “Multilayer feedforward networks are universal approximators”. In: *Neural Networks* 2.5, pp. 359–366. ISSN: 0893-6080. DOI: https://doi.org/10.1016/0893-6080(89)90020-8.

Horváth, L., Panza, T., and Karadža, N. (Jan. 2007). “The influence of high wind hysteresis effect on wind turbine power production at Bura-dominated site”. In: 2, pp. 1017–1022.

Huang, C., Li, F., and Jin, Z. (Apr. 2015). “Maximum Power Point Tracking Strategy for Large-Scale Wind Generation Systems Considering Wind Turbine Dynamics”. In: *IEEE Transactions on Industrial Electronics* 62.4, pp. 2530–2539. ISSN: 1557-9948. DOI: 10.1109/TIE.2015.2395384.

Huang, G.-B., Zhu, Q.-Y., and Siew, C.-K. (2006). “Extreme learning machine: Theory and applications”. In: *Neurocomputing* 70.1. Neural Networks, pp. 489–501. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2005.12.126.

Huttenlocher, D., Lilien, R., and Olson, C. (Aug. 1996). “Object Recognition Using Subspace Methods”. In: vol. 1064. ISBN: 978-3-540-61122-6. DOI: 10.1007/BFb0015564.

IEA (2021). *Global Energy Review 2021.* International Energy Agency.

IPCC (2011). *Summary for Policymakers. In: IPCC Special Report on Renewable Energy Sources and Climate Change Mitigation* [O. Edenhofer, R. Pichs-Madruga, Y. Sokona, K. Seyboth, P. Matschoss, S. Kadner, T. Zwickel, P. Eickemeier, G. Hansen, S. Schlömer, C. von Stechow (eds)]. Cambridge University Press.

— (2012). *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change* [Field, C.B., V. Barros, T.F. Stocker, D. Qin, D.J. Dokken, K.L. Ebi, M.D. Mastrandrea, K.J. Mach, G.-K. Plattner, S.K. Allen, M. Tignor and P.M. Midgley (eds.)]. Cambridge University Press. DOI: 10.1017/CBO9781139177245.

— (2018). *Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty* [Masson-Delmotte, V., P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J.B.R.

Matthews, Y. Chen, X. Zhou, M.I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, and T. Waterfield (eds.)]. In Press.

IPCC (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* [Masson-Delmotte, V., P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu and B. Zhou (eds.)]. Cambridge University Press.

IRENE (2019). *Future of wind: Deployment, investment, technology, grid integration and socio-economic aspects (A Global Energy Transformation paper)*. International Renewable Energy Agency.

— (2021). *Renewable Power Generation Costs in 2020*. International Renewable Energy Agency.

Jelavić, M., Petrović, V., Barišić, M., and Ivanović, I. (Jan. 2013). "Wind turbine control beyond the cut-out wind speed". In: *European Wind Energy Conference and Exhibition, EWEC 2013* 1, pp. 343–349.

Jiang, P., Wang, Y., and Wang, J. (2017). "Short-term wind speed forecasting using a hybrid model". In: *Energy* 119.C, pp. 561–577. DOI: 10.1016/j.energy.2016.10..

Jolliffe, I. T. and Stephenson, D. B. (2012). In: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley & Sons, Ltd. ISBN: 9781119960003.

Jung, J. and Broadwater, R. P. (2014). "Current status and future advances for wind speed and power forecasting". In: *Renewable and Sustainable Energy Reviews* 31.C, pp. 762–777. DOI: 10.1016/j.rser.2013.12.05.

Kavasseri, R. G. and Seetharaman, K. (2009). "Day-ahead wind speed forecasting using f-ARIMA models". In: *Renewable Energy* 34.5, pp. 1388–1393. DOI: 10.1016/j.renene.2008.09..

Keil, C. and Craig, G. C. (2007). "A Displacement-Based Error Measure Applied in a Regional Ensemble Forecasting System". In: *Monthly Weather Review* 135.9, pp. 3248–3259. DOI: 10.1175/MWR3457.1.

— (2009). "A Displacement and Amplitude Score Employing an Optical Flow Technique". In: *Weather and Forecasting* 24.5, pp. 1297–1308. DOI: 10.1175/2009WAF2222247.1.

Keller, K., Robinson, A., Bradford, D. F., and Oppenheimer, M. (Apr. 2007). "The regrets of procrastination in climate policy". In: *Environmental Research Letters* 2.2, p. 024004. DOI: 10.1088/1748-9326/2/2/024004.

Kikuchi, R., Misaka, T., Obayashi, S., Inokuchi, H., Oikawa, H., and Misumi, A. (Dec. 2017). "Nowcasting algorithm for wind fields using ensemble forecasting and aircraft flight data". In: *Meteorological Applications* 25. DOI: 10.1002/met.1704.

Lack, S. A., Limpert, G. L., and Fox, N. I. (2010). "An Object-Oriented Multiscale Verification Scheme". In: *Weather and Forecasting* 25.1, pp. 79–92. DOI: 10.1175/2009WAF2222245.1.

Landberg, L. (2001). "Short-term prediction of local wind conditions". In: *Journal of Wind Engineering and Industrial Aerodynamics* 89.3. 10th International Conference on Wind Engineering, pp. 235–245. ISSN: 0167-6105. DOI: https://doi.org/10.1016/S0167-6105(00)00079-9.

Lange, M. and Focken, U. (2006). *Physical Approach to Short-Term Wind Power Prediction*. Springer-Verlag Berlin Heidelberg. ISBN: 978-3-540-31106-5. DOI: 10.1007/3-540-31106-8.

Lee, D. and Baldick, R. (Jan. 2014). "Short-Term Wind Power Ensemble Prediction Based on Gaussian Processes and Neural Networks". In: *Smart Grid, IEEE Transactions on* 5, pp. 501–510. DOI: 10.1109/TSG.2013.2280649.

Lei, M., Shiyan, L., Chuanwen, J., Hongling, L., and Yan, Z. (2009). "A review on the forecasting of wind speed and generated power". In: *Renewable and Sustainable Energy Reviews* 13.4, pp. 915–920. ISSN: 1364-0321. DOI: https://doi.org/10.1016/j.rser.2008.02.002.

Leva, S., Dolara, A., Grimaccia, F., Mussetta, M., and Ogliari, E. (2017). "Analysis and validation of 24 hours ahead neural network forecasting of photovoltaic output power". In: *Mathematics and Computers in Simulation (MATCOM)* 131.C, pp. 88–100.

Li, C., Xiao, Z., Xia, X., Zou, W., and Zhang, C. (2018). "A hybrid model based on synchronous optimisation for multi-step short-term wind speed forecasting". In: *Applied Energy* 215, pp. 131–144. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2018.01.094.

Li, G. and Shi, J. (July 2010). "On comparing three artificial neural networks for wind speed forecasting". In: *Applied Energy* 87, pp. 2313–2320. DOI: 10.1016/j.apenergy.2009.12.013.

Liu, H., Mi, X.-w., and Li, Y.-f. (2018a). "Wind speed forecasting method based on deep learning strategy using empirical wavelet transform, long short term memory neural network and Elman neural network". In: *Energy Conversion and Management* 156, pp. 498–514. ISSN: 0196-8904. DOI: https://doi.org/10.1016/j.enconman.2017.11.053.

Liu, H., Mi, X., and Li, Y. (2018b). "Smart deep learning based wind speed prediction model using wavelet packet decomposition, convolutional neural network and convolutional long short term memory network". In: *Energy Conversion and Management* 166, pp. 120–131. ISSN: 0196-8904. DOI: https://doi.org/10.1016/j.enconman.2018.04.021.

Liu, H., Tian, H.-q., and Li, Y.-f. (2015). "Comparison of new hybrid FEEMD-MLP, FEEMD-ANFIS, Wavelet Packet-MLP and Wavelet Packet-ANFIS for wind speed predictions". English. In: *Energy Conversion and Management* 89.Complete, pp. 1–11. DOI: 10.1016/j.enconman.2014.09.060.

Liu, Y., Racah, E., Prabhat, Correa, J., Khosrowshahi, A., Lavers, D., Kunkel, K., Wehner, M., and Collins, W. (2016). *Application of Deep Convolutional Neural Networks for Detecting Extreme Weather in Climate Datasets*.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. (2017). *The Expressive Power of Neural Networks: A View from the Width*.

M. E. Baldwin, S. L. (2003). "Development of an events-oriented verification system using data mining and image processing algorithms". In: *Third Conference on Artificial Intelligence*. American Meteorological Society.

Magnusson, L., Richardson, D., and Haiden, T. (Sept. 2014). "Verification of extreme weather events: Discrete predictands". In: 731. DOI: 10.21957/1iql31n2c.

Markou, H. and Larsen, T. (2009). "Control strategies for operation of pitch regulated turbines above cut-out wind speeds". English. In: *EWEC 2009 Proceedings online*. 2009 European Wind Energy Conference and Exhibition, EWEC 2009 ; Conference date: 16-03-2009 Through 19-03-2009. EWEC.

Marndi, A., Patra, G. K., and Gouda, K. C. (Mar. 2020). "Short-term forecasting of wind speed using time division ensemble of hierarchical deep neural networks". In: *Bulletin of Atmospheric Science and Technology* 1.1, pp. 91–108. DOI: 10.1007/s42865-020-00009-2.

Marzban, C. and Sandgathe, S. (2006). “Cluster Analysis for Verification of Precipitation Fields”. In: *Weather and Forecasting* 21.5, pp. 824–838. DOI: 10.1175/WAF948.1.

— (2009). “Verification with Variograms”. In: *Weather and Forecasting* 24.4, pp. 1102–1120. DOI: 10.1175/2009WAF2222122.1.

— (2010). “Optical Flow for Verification”. In: *Weather and Forecasting* 25.5, pp. 1479–1494. DOI: 10.1175/2010WAF2222351.1.

Marzban, C., Sandgathe, S., Lyons, H., and Lederer, N. (Feb. 2009). “Three Spatial Verification Techniques: Cluster Analysis, Variogram, and Optical Flow”. In: *Weather and Forecasting - WEATHER FORECAST* 24. DOI: 10.1175/2009WAF2222261.1.

Marzban, C. and Stumpf, G. J. (1998). “A Neural Network for Damaging Wind Prediction”. In: *Weather and Forecasting* 13.1, pp. 151–163. DOI: 10.1175/1520-0434(1998)013<0151:ANNFDW>2.0.CO;2.

Mason, I. (1989). “Dependence of the Critical Sucess Index on sample climate and threshold probability”. In: *Australian Meteorological Magazine* 37, pp. 75–81.

Mass, C. F., Ovens, D., Westrick, K., and Colle, B. A. (2002). “Does increasing horizontal produce more skillful forecasts?: The Results of Two Years of Real-Time Numerical Weather Prediction over the Pacific Northwest”. In: *Bulletin of the American Meteorological Society* 83.3, pp. 407–430. DOI: 10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2.

McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., Smith, T., and Williams, J. K. (2017). “Using Artificial Intelligence to Improve Real-Time Decision-Making for High-Impact Weather”. In: *Bulletin of the American Meteorological Society* 98.10, pp. 2073–2090. DOI: 10.1175/BAMS-D-16-0123.1.

Mi, X.-w., Liu, H., and Li, Y.-f. (2017). “Wind speed forecasting method using wavelet, extreme learning machine and outlier correction algorithm”. In: *Energy Conversion and Management* 151, pp. 709–722. ISSN: 0196-8904. DOI: https://doi.org/10.1016/j.enconman.2017.09.034.

Micheas, A. C., Fox, N. I., Lack, S. A., and Wikle, C. K. (2007). “Cell identification and verification of QPF ensembles using shape analysis techniques”. In: *Journal of Hydrology* 343.3, pp. 105–116. ISSN: 0022-1694. DOI: https://doi.org/10.1016/j.jhydrol.2007.05.036.

Mittelman, R., Kuipers, B., Savarese, S., and Lee, H. (June 2014). “Structured Recurrent Temporal Restricted Boltzmann Machines”. In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by E. P. Xing and T. Jebara. Vol. 32. Proceedings of Machine Learning Research 2. Bejing, China: PMLR, pp. 1647–1655.

Miyakoda, K., Hembree, G. D., Strickler, R. F., and Shulman, I. (1972). “Cumulative Results of Extended Forecast Experiments I. Model Performance for Winter Cases”. In: *Monthly Weather Review* 100.12, pp. 836–855. DOI: 10.1175/1520-0493(1972)100<0836:CROEFE>2.3.CO;2.

Murphy, A. H. (1991). “Probabilities, Odds, and Forecasts of Rare Events”. In: *Weather and Forecasting* 6.2, pp. 302–307. DOI: 10.1175/1520-0434(1991)006<0302:POAFOR>2.0.CO;2.

— (1993). “What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting”. In: *Weather and Forecasting* 8.2, pp. 281–293. DOI: 10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Murphy, A. H. and Epstein, E. S. (1989). “Skill Scores and Correlation Coefficients in Model Verification”. In: *Monthly Weather Review* 117.3, pp. 572–582. DOI: 10.1175/1520-0493(1989)117<0572:SSACCI>2.0.CO;2.

Murphy, A. H. and Winkler, R. L. (1987). "A General Framework for Forecast Verification". In: *Monthly Weather Review* 115.7, pp. 1330–1338. DOI: 10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2.

Nachamkin, J. E. (2004). "Mesoscale Verification Using Meteorological Composites". In: *Monthly Weather Review* 132.4, pp. 941–955. DOI: 10.1175/1520-0493(2004)132<0941:MVUMC>2.0.CO;2.

Namias, J. (1952). "The Annual Course of Month-to-Month Persistence in Climatic Anomalies". In: *Bulletin of the American Meteorological Society* 33.7, pp. 279–285. DOI: 10.1175/1520-0477-33.7.279.

Nicholls, R., Marinova, N., Lowe, J., Brown, S., Vellinga, P., Gusmão, D. de, Hinkel, J., and Tol, R. (2011). "Sea-level rise and its possible impacts given a 'beyond 4°C world' in the twenty-first century". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 369, pp. 161–181.

Noorollahi, Y., Jokar, M. A., and Kalhor, A. (May 2016). "Using artificial neural networks for temporal and spatial wind speed forecasting in Iran". In: *Energy Conversion and Management* 115, pp. 17–25. DOI: 10.1016/j.enconman.2016.02.041.

Oh, J., Guo, X., Lee, H., Lewis, R., and Singh, S. (2015). "Action-Conditional Video Prediction Using Deep Networks in Atari Games". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'15. Montreal, Canada: MIT Press, pp. 2863–2871.

Oliveira, M., Moniz, N., Torgo, L., and Santos Costa, V. (2021). "Biased resampling strategies for imbalanced spatio-temporal forecasting". In: *International Journal of Data Science and Analytics* 12.3, pp. 205–228. DOI: 10.1007/s41060-021-00256-2.

Oppenheimer, M., Glavovic, B., Hinkel, J., Wal, R. van de, Magnan, A., Abd-Elgawad, A., Cai, R., Cifuentes-Jara, M., DeConto, R., Ghosh, T., Hay, J., Isla, F., Marzeion, B., Meyssignac, B., and Sebesvari, Z. (2019). *Sea Level Rise and Implications for Low-Lying Islands, Coasts and Communities. In: IPCC Special Report on the Ocean and Cryosphere in a Changing Climate* [H.-O. Pörtner, D.C. Roberts, V. Masson-Delmotte, P. Zhai, M. Tignor, E. Poloczanska, K. Mintenbeck, A. Alegría, M. Nicolai, A. Okem, J. Petzold, B. Rama, N.M. Weyer (eds.)]. In press.

Peirce, C. S. (Nov. 1884). "The numerical measure of the success of predictions." eng. In: *Science* 4.93, pp. 453–454. ISSN: 0036-8075 (Print); 0036-8075 (Linking). DOI: 10.1126/science.ns-4.93.453-a.

Petrović, V. and Bottasso, C. L. (June 2014). "Wind turbine optimal control during storms". In: *Journal of Physics: Conference Series* 524, p. 012052. DOI: 10.1088/1742-6596/524/1/012052.

Petrović, V. and Bottasso, C. L. (2017). "Wind turbine envelope protection control over the full wind speed range". In: *Renewable Energy* 111.C, pp. 836–848.

Pourmousavi Kani, S. and Ardehali, M. (2011). "Very short-term wind speed prediction: A new artificial neural network–Markov chain model". English. In: *Energy Conversion and Management* 52.1, pp. 738–745. DOI: 10.1016/j.enconman.2010.07.053.

Racah, E., Beckham, C., Maharaj, T., Kahou, S. E., Prabhat, and Pal, C. (2017). "Extreme Weather: A Large-Scale Climate Dataset for Semi-Supervised Detection, Localization, and Understanding of Extreme Weather Events". In: *Proceedings of the 31st International Conference on Neural*

Information Processing Systems. NIPS'17. Long Beach, California, USA: Curran Associates Inc., pp. 3405–3416. ISBN: 9781510860964.

Rädler, A., Groenemeijer, P., Pistotnik, G., Sausen, R., and Faust, E. (Nov. 2015). "Identification of favorable environments for thunderstorms in reanalysis data". In: *Meteorologische Zeitschrift* 26. DOI: 10.1127/metz/2016/0754.

Rao, A. R., Wang, Q., Wang, H., Khorasgani, H., and Gupta, C. (2020). "Spatio-Temporal Functional Neural Networks". In: *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 81–89. DOI: 10.1109/DSAA49011.2020.00020.

REN21 (2021). *Renewables 2021 Global Status Report*.

Rezacova, D., Sokol, Z., and Pesice, P. (Feb. 2007). "A radar-based verification of precipitation forecast for local convective storms". In: *Atmospheric Research* 83.2-4, pp. 211–224. DOI: 10.1016/j.atmosres.2005.08.011.

Riahy, G. and Abedi, M. (Jan. 2008). "Short term wind speed forecasting for wind turbine applications using linear prediction method". In: *Renewable Energy* 33, pp. 35–41. DOI: 10.1016/j.renene.2007.01.014.

Ribeiro, R. and Moniz, N. (Sept. 2020). "Imbalanced regression and extreme value prediction". In: *Machine Learning* 109, pp. 1–33. DOI: 10.1007/s10994-020-05900-9.

Robert, S., Foresti, L., and Kanevski, M. (2013). "Spatial prediction of monthly wind speeds in complex terrain with adaptive general regression neural networks". In: *International Journal of Climatology* 33.7, pp. 1793–1804. DOI: https://doi.org/10.1002/joc.3550.

Roberts, N. M. (2008). "Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model". In: *Meteorological Applications* 15.1, pp. 163–169. DOI: https://doi.org/10.1002/met.57.

Roberts, N. M. and Lean, H. W. (2008). "Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events". In: *Monthly Weather Review* 136.1, pp. 78–97. DOI: 10.1175/2007MWR2123.1.

Rossa, A., Nurmi, P., and Ebert, E. (2008). "Overview of methods for the verification of quantitative precipitation forecasts". In: *Precipitation: Advances in Measurement, Estimation and Prediction*. Ed. by S. Michaelides. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 419–452. ISBN: 978-3-540-77655-0. DOI: 10.1007/978-3-540-77655-0_16.

Ruder, S. (2017). *An overview of gradient descent optimization algorithms*.

Salcedo-Sanz, S., Pérez-Bellido, Á., Ortiz-García, E., Portilla-Figueras, A., Prieto, L., and Correoso, F. (Jan. 2009). "Accurate Short-Term Wind Speed Prediction by Exploiting Diversity in Input Data using Banks of Artificial Neural Networks". In: *Neurocomputing* 72, pp. 1336–1341. DOI: 10.1016/j.neucom.2008.09.010.

Sanderson, B. M. and O'Neill, B. C. (2020). "Assessing the costs of historical inaction on climate change". In: *Scientific Reports* 10.1, p. 9173. DOI: 10.1038/s41598-020-66275-4.

Scher, S. and Messori, G. (2021). "Ensemble Methods for Neural Network-Based Weather Forecasts". In: *Journal of Advances in Modeling Earth Systems* 13.2. DOI: https://doi.org/10.1029/2020MS002331.

Scheuerer, M. and Hamill, T. M. (2015). "Statistical Postprocessing of Ensemble Precipitation Forecasts by Fitting Censored, Shifted Gamma Distributions". In: *Monthly Weather Review* 143.11, pp. 4578–4596. DOI: 10.1175/MWR-D-15-0061.1.

Scheuerer, M., Switanek, M. B., Worsnop, R. P., and Hamill, T. M. (2020). "Using Artificial Neural Networks for Generating Probabilistic Subseasonal Precipitation Forecasts over California". In: *Monthly Weather Review* 148.8, pp. 3489–3506. DOI: 10.1175/MWR-D-20-0096.1.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-k., and Woo, W.-c. (2015). "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, pp. 802–810.

Shi, X., Gao, Z., Lausen, L., Wang, H., and Yeung, D.-Y. (2017). "Deep Learning for Precipitation Nowcasting: A Benchmark and A New Model". In: *Proceedings of the 31th International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, CA, USA: MIT Press, pp. 5617–5627.

Shi, X. and Yeung, D.-Y. (2018). *Machine Learning for Spatiotemporal Sequence Forecasting: A Survey*.

Skok, G. and Roberts, N. (2016). "Analysis of Fractions Skill Score properties for random precipitation fields and ECMWF forecasts". In: *Quarterly Journal of the Royal Meteorological Society* 142.700, pp. 2599–2610. DOI: https://doi.org/10.1002/qj.2849.

Srivastava, N., Mansimov, E., and Salakhudinov, R. (July 2015). "Unsupervised Learning of Video Representations using LSTMs". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by F. Bach and D. Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, pp. 843–852.

Stephenson, D., Casati, B., Ferro, C., and Wilson, C. (Mar. 2008). "The Extreme Dependency Score: A non-vanishing measure for forecasts of rare events". In: *Meteorological Applications* 15, pp. 41–50. DOI: 10.1002/met.53.

Sutskever, I. and Hinton, G. (Mar. 2007). "Learning Multilevel Distributed Representations for High-Dimensional Sequences". In: *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*. Ed. by M. Meila and X. Shen. Vol. 2. Proceedings of Machine Learning Research. San Juan, Puerto Rico: PMLR, pp. 548–555.

Sutskever, I., Hinton, G. E., and Taylor, G. W. (2009). "The Recurrent Temporal Restricted Boltzmann Machine". In: *Advances in Neural Information Processing Systems*. Ed. by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou. Vol. 21. Curran Associates, Inc.

Teweles, S. and Wobus, H. B. (1954). "Verification of prognostic charts". In: *American Meteorological Society* 35, pp. 455–463.

Thomas, S. R., Martínez-Alvarado, O., Drew, D., and Bloomfield, H. (2021). "Drivers of extreme wind events in Mexico for windpower applications". In: *International Journal of Climatology* 41.S1, E2321–E2340. DOI: https://doi.org/10.1002/joc.6848.

Thompson, J. C. and Carter, G. M. (1972). "On Some Characteristics of the S1 Score". In: *Journal of Applied Meteorology and Climatology* 11.8, pp. 1384–1385. DOI: 10.1175/1520-0450(1972)011<1384:OSCOTS>2.0.CO;2.

Tsonevsky, I., Doswell, C. A., and Brooks, H. E. (2018). "Early Warnings of Severe Convection Using the ECMWF Extreme Forecast Index". In: *Weather and Forecasting* 33.3, pp. 857–871. DOI: 10.1175/WAF-D-18-0030.1.

Venugopal, V., Basu, S., and Foufoula-Georgiou, E. (Apr. 2005). "A new metric for comparing precipitation patterns with an application to ensemble forecasts". In: *Journal of Geophysical Research (Atmospheres)* 110.D8, D08111, p. D08111. DOI: 10.1029/2004JD005395.

Visconti, G. (2001). "Chaos in the Atmosphere". In: *Fundamentals of Physics and Chemistry of the Atmosphere*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 507–524. ISBN: 978-3-662-04540-4. DOI: 10.1007/978-3-662-04540-4_19.

Vondrick, C., Pirsiavash, H., and Torralba, A. (2016). "Generating Videos with Scene Dynamics". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., pp. 613–621. ISBN: 9781510838819.

Wan, C., Xu, Z., Pinson, P., Dong, Z., and Wong, K. (2014). "Probabilistic forecasting of wind power generation using extreme learning machine." English. In: *I E E E Transactions on Power Systems* 29.3. ISSN: 0885-8950. DOI: 10.1109/TPWRS.2013.2287871.

Wang, H.-z., Li, G.-q., Wang, G.-b., Peng, J.-c., Jiang, H., and Liu, Y.-t. (2017a). "Deep learning based ensemble approach for probabilistic wind power forecasting". In: *Applied Energy* 188.C, pp. 56–70. DOI: 10.1016/j.apenergy.2016.1.

Wang, S., Cao, J., and Yu, P. (2020). "Deep Learning for Spatio-Temporal Data Mining: A Survey". In: *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1. DOI: 10.1109/TKDE.2020.3025580.

Wang, Y., Gao, Z., Long, M., Wang, J., and Yu, P. (Apr. 2018). "PredRNN++: Towards A Resolution of the Deep-in-Time Dilemma in Spatiotemporal Predictive Learning". In: *ICML*.

Wang, Y., Long, M., Wang, J., Gao, Z., and Yu, P. S. (2017b). "PredRNN: Recurrent Neural Networks for Predictive Learning Using Spatiotemporal LSTMs". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., pp. 879–888. ISBN: 9781510860964.

Wilks, D. S. (2019). "Forecast Verification". In: *Statistical Methods in the Atmospheric Sciences*. Ed. by D. S. Wilks. Fourth Edition. Elsevier. Chap. 9, pp. 369–483. ISBN: 978-0-12-815823-4. DOI: https://doi.org/10.1016/B978-0-12-815823-4.00009-2.

Wiser, R., Yang, Z., Hand, M., Hohmeyer, O., Infield, D., Jensen, P. H., Nikolaev, V., O'Malley, M., Sinden, G., and Zervos, A. (2011). *Wind Energy. In IPCC Special Report on Renewable Energy Sources and Climate Change Mitigation* [O. Edenhofer, R. Pichs-Madruga, Y. Sokona, K. Seyboth, P. Matschoss, S. Kadner, T. Zwickel, P. Eickemeier, G. Hansen, S. Schlömer, C. von Stechow (eds)]. Cambridge University Press.

WMO (2017). *Guidelines for Nowcasting Techniques*. English. World Meteorological Organization. ISBN: 978-92-63-11198-2.

Wunderling, N., Donges, J. F., Kurths, J., and Winkelmann, R. (2021). "Interacting tipping elements increase risk of climate domino effects under global warming". In: *Earth System Dynamics* 12.2, pp. 601–619. DOI: 10.5194/esd-12-601-2021.

Xie, A., Yang, H., Chen, J., Sheng, L., and Zhang, Q. (2021). "A Short-Term Wind Speed Forecasting Model Based on a Multi-Variable Long Short-Term Memory Network". In: *Atmosphere* 12.5. ISSN: 2073-4433. DOI: 10.3390/atmos12050651.

Yang, L., He, M., Zhang, J., and Vittal, V. (2015). "Support-Vector-Machine-Enhanced Markov Model for Short-Term Wind Power Forecast". In: *IEEE Transactions on Sustainable Energy* 6.3, pp. 791–799. DOI: 10.1109/TSTE.2015.2406814.

Yang, Y., Zha, K., Chen, Y.-C., Wang, H., and Katabi, D. (2021). *Delving into Deep Imbalanced Regression.*

Yates, E., Anquetin, S., Ducrocq, V., Creutin, J.-D., Ricard, D., and Chancibault, K. (2006). "Point and areal validation of forecast precipitation fields". In: *Meteorological Applications* 13.1, pp. 1–20. DOI: https://doi.org/10.1017/S1350482705001921.

Yu, H., Uy, W. I. T., and Dauwels, J. (2017). "Modeling Spatial Extremes via Ensemble-of-Trees of Pairwise Copulas". In: *IEEE Transactions on Signal Processing* 65.3, pp. 571–586. DOI: 10.1109/TSP.2016.2614485.

Zell, A. (1994). *Simulation neuronaler Netze*. ger. 1. Aufl. Bonn [u.a.]: Addison-Wesley, 624 S.

Zepeda-Arce, J., Foufoula-Georgiou, E., and Droegemeier, K. (Apr. 2000). "Space-time rainfall organization and its role in validating quantitative precipitation forecasts". English (US). In: *Journal of Geophysical Research A: Space Physics* 105.D8, pp. 10129–10146. ISSN: 2169-9380. DOI: 10.1029/1999JD901087.

Zhai, M., Xiang, X., Lv, N., and Kong, X. (2021). "Optical flow and scene flow estimation: A survey". In: *Pattern Recognition* 114, p. 107861. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2021.107861.

Zhao, B. and Zhang, B. (2018). "Assessing Hourly Precipitation Forecast Skill with the Fractions Skill Score". In: *Journal of Meteorological Research* 32.1, pp. 135–145. DOI: 10.1007/s13351-018-7058-1.

Zhao, E., Zhao, J., Liu, L., Su, Z., and An, N. (2016a). "Hybrid Wind Speed Prediction Based on a Self-Adaptive ARIMAX Model with an Exogenous WRF Simulation". In: *Energies* 9.1. ISSN: 1996-1073. DOI: 10.3390/en9010007.

Zhao, Y., Ye, L., Li, Z., Song, X., Lang, Y., and Su, J. (2016b). "A novel bidirectional mechanism based on time series model for wind power forecasting". In: *Applied Energy* 177.C, pp. 793–803. DOI: 10.1016/j.apenergy.2016.0.

Zhu, Q., Chen, J., Zhu, L., Duan, X., and Liu, Y. (2018). "Wind Speed Prediction with Spatio–Temporal Correlation: A Deep Learning Approach". In: *Energies* 11.4. ISSN: 1996-1073. DOI: 10.3390/en11040705.

Zucatelli, P. J., Nascimento, E. G. S., Santos, A. Á. B., and Moreira, D. M. (2020). "Nowcasting prediction of wind speed using computational intelligence and wavelet in Brazil". In: *International Journal for Computational Methods in Engineering Science and Mechanics* 21.6, pp. 343–369. DOI: 10.1080/15502287.2020.1841335.