~~https://www.overleaf.com/project/61b7316acd0428731aa19448~~

~~An adapted~~ Adapting a deep convolutional RNN model with imbalanced regression loss for improved spatio-temporal ~~prediction~~ forecasting of extreme wind speed ~~extremes~~ events in the short-to-medium range~~for wind energy applications~~.

Daan R. Scheepens[1], Irene Schicker[2], Kateřina Hlaváčková-Schindler[1], and Claudia Plant[1]

[1]Research Group Data Mining and Machine Learning, Faculty of Computer Science, University of Vienna, Währingerstrasse 29, 1090 Vienna, Austria; d.r.scheepens@gmail.com (DRS); katerina.schindlerova@univie.ac.at (KHS); claudia.plant@univie.ac.at (CP).
[2]Zentralanstalt für Meteorologie und Geodynamik (ZAMG), Hohe Warte 38, 1190 Vienna, Austria; irene.schicker@zamg.ac.at (IS)

Correspondence: d.r.scheepens@gmail.com, irene.schicker@zamg.ac.at

Abstract. The amount of wind farms and wind power production in Europe, both on- and off-shore, has increased rapidly in the past years. To ensure grid stability, on-time (re)scheduling of maintenance tasks and to 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. ~~In this work we propose~~ This work explores the possibility of adapting a deep convolutional recurrent neural network (RNN) based regression model ~~,~~ for the spatio-temporal prediction of extreme wind speed events ~~over Europe~~ in the short-to-medium range (12 hour lead-time in one hour intervals) ~~. This is achieved by training~~ through the manipulation of the loss function. To this end, a multi-layered convolutional long short-term memory (ConvLSTM) network ~~with so-called~~ is adapted with a variety of imbalanced regression loss ~~. To this end we investigate three different loss functions : the inversely weightedmean absolute error (W-MAE) loss, the inversely weighted mean squared error (W-MSE) loss and the~~ functions that have been proposed in the literature: Inversely weighted, linearly weighted and squared error-relevance area (SERA) loss. ~~We investigate forecast performance for various high-threshold~~ Forecast performance is investigated for various intensity thresholds of extreme events and ~~for various numbers of network layers, and compare the imbalanced regression loss functions to the~~ a comparison is made with the commonly used mean squared error (MSE) and mean absolute error (MAE) loss. The results indicate ~~superior performance of an ensemble of networks trained with either W-MAE, W-MSE or SERA loss, showing substantial improvements on high intensity extreme events. We conclude~~ the inverse weighting method to most effectively shift the forecast distribution towards the extreme tail, thereby increasing the number of forecasted events in the extreme ranges, considerably boosting the hit rate and reducing the root mean squared error (RMSE) in those ranges. The results also show, however, that such improvements are invariably accompanied by a pay-off primarily in terms of increased overcasting and false alarm ratios, which increase both with lead-time and intensity threshold. The inverse weighting method is judged to most effectively balance this trade-off, with the weighted MAE loss scoring slightly better than the weighted MSE loss. It is concluded that the ConvLSTM ~~trained with imbalanced regression~~ network trained with inversely weighted 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 speed events in the short-to-medium range. ~~This work was performed 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.~~

1 Introduction

Global warming demands ever more urgently that electricity generation is shifted away from fossil fuels and towards renewable energy sources. Although global demands for fossil fuels are not yet showing signs of decreasing, renewables are on the rise. In 2021, more than half of the growth in global electricity supply was provided by renewables, while the share of renewables in global electricity generation reached close to 30 %, having steadily risen over the past decades (IEA, 2021). Possessing the largest market share among 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). Thanks to continued advancements in on- and offshore wind energy technology and the associated continued reduction in costs, wind power capacity could grow from having met 1.8 % of global electricity demand in 2009 to meeting roughly 20 % of demand in 2030 (Darwish and Al-Dabbagh, 2020). Indeed, many countries have already demonstrated that hybrid electric systems with large contributions of wind energy can operate reliably. For example, in 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 numbers have only risen since.

One of the main challenges to the deployment of wind energy, however, is its inherent variability and lower level of predictability than are common for other types of power plants (Lei et al., 2009; Chen and Yu, 2014; 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 and thus ensure grid stability (Wiser et al., 2011). Failing to manage this variability leads to scheduling errors which impact grid reliability and market-based ancillary service costs (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, 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 the aforementioned 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 (Wiser et al., 2011; Li et al., 2018). The development of accurate wind speed forecasts in the short-range has ~~therefore~~ thus become increasingly important.

Short-term wind speed prediction is not just a key element in the successful management of hybrid electric power systems, it is also vital in the 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 of around 25 m s^{-1}, after which the rotation of the blades is brought to a halt and the turbine is essentially turned off (Burton et al., 2001). 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). ~~The prediction of extreme wind speeds poses a considerable challenge to computer science research, however, where heavy-tailed distributions such as those of wind speed (modelled according to a type III extreme value 'Weibull' distribution) pose a serious problem to the statistical prediction of extreme values at the upper and lower tails of the distribution. In the case of regression this problem is referred to as imbalanced regression, which we attempt to tackle in this paper for extreme wind speed prediction in the spatio-temporal setting using an adapted convolutional recurrent neural network (ConvRNN) model and imbalanced regression loss.~~

~~Due to the growing utilisation of wind power as a renewable energy resource, a large amount of research has focused on the development of new and improved methods for reliable forecasting of wind speed and wind power. These methods can be broadly divided into either physical model based methods or statistical modelling methods (Costa et al., 2008; Lei et al., 2009; Jung and Bro~~ ~~. Physical model based methods, such as numerical weather prediction (NWP) models, are highly capable of modelling the state of the atmosphere and have been used extensively for wind speed forecasting (see e.g. Alessandrini et al., 2013; Deppe et al., 2013; Kikuchi~~ ~~. However, due to high computational demands they tend to have a long temporal lag (depending on their domain coverage, spatial resolution and temporal forecast frequency) which means that for the nowcasting and short-time prediction range NWP forecasts are typically not available on time. The physical model based methods, furthermore, suffer drawbacks due to the often laborious acquisition of the site-specific physical data (Jung and Broadwater, 2014) and the fact that the predictive capability of NWP models degrades significantly for highly stochastic variables like wind (Chen and Yu, 2014). In practice, 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.~~

~~Alternatively, statistical modelling (i.e. data-driven) methods have proved to be another viable solution for the problem of weather prediction. Among these,~~ While there has been a particularly strong trend ~~in the past years towards deep artificial neural networks, also termed deep learning (DL). In fact, the artificial neural network (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). The power of the ANN lies in its ability to model highly complex and non-linear relationships between input and output while requiring no prior assumption on the mathematical relationship between them (Jung and Broadwater, 2014) . Deep (i.e. multi-layered) ANNs are capable of automatically and effectively learning hierarchical feature representations from raw input data, where different layers in the network essentially learn 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 given to the model (Wang et al., 2020). The above qualities have made deep learning models particularly attractive to the 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 learned very effectively directly from the data (Wang et al., 2020). For excellent review papers on deep learning applications to STSF we refer the reader to Shi and Yeung (2018), Amato et al. (2020) or Wang et al. (2020).~~

~~Deep learning can be applied to STSF in myriad ways. Srivastava et al. (2015) proposed the usage of a multi-layered, fully connected long short-term memory (FC-LSTM) network for video frame prediction by flattening the input images directly into arrays to be used by the 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, embedding 2D CNNs into the LSTM network structure, which the authors applied to precipitation nowcasting. A similar extension was made to the gated recurrent unit (GRU) network by Shi et al. (2017), the ConvGRU, which was also applied to precipitation nowcasting, where it demonstrated a superior ability to capture rotating precipitation fields. Instead of using a 2D CNN to capture spatial correlations only, 3D CNN models may also be used instead 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. Arguably, the combinations of 2D CNNs with RNN networks into ConvRNNs (such as the ConvLSTM) have been met with the most success, and have been used extensively in the literature as building blocks for DL models for STSF tasks (Shi and Yeung, 2018). Improvements to the ConvRNN network structure have been~~ within the area of weather forecasting research towards data-driven, deep artificial neural networks (Jung and Broadwater, 2014), such forecasting models are faced with a considerable challenge when tasked with the prediction of extreme events. Typically referring to the upper or lower tails of the data distribution, extreme values are inherently underrepresented during data-driven model learning and thus typically suffer from poor predictability and low bias in comparison to the bulk of the distribution. Improving the predictability of extreme values of data-driven models comprises an active area of research and various approaches 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. (2017) 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 and the authors demonstrate the model 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 possible improvements to the ConvRNN approaches. Generative adversarial networks (GANs) can offer yet another alternative to STSF, a thorough review of which is provided by Gao et al. (2020).~~ depending on the nature of the task. Class imbalances within classification tasks, for example, can be mitigated with a wide range of resampling strategies, either resampling the classes themselves (e.g. Batista et al., 2004) or the underlying probability density function (Mohamad and Sapsis, 2018; Hassanaly et al., 2021, e.g.). The task may, furthermore, be treated as one-class classification (e.g. Deng et al., 2018; Goyal et al., 2020) or outlier exposure (e.g. Hendrycks et al., 2019) .

~~More recently, Rasp et al. (2020) created a benchmark data set for data-driven spatio-temporal forecasts which has been used extensively since its publication. Rasp and Thuerey (2021) used a ResNet to predict three parameters (geopotential, temperature and precipitation) with a coarse spatial resolution of 5.625° for up to five days ahead whereas Weyn et al. (2020) used a convolution neural network on a cubed sphere. A different approach was followed by Pathak et al. (2022) who used a Fourier-based neural network for forecasting surface wind speed and total precipitation on a global scale, with a spatial resolution of 0.25° and for lead times of up to ten days. Lastly, Keisler (2022) implemented a graph neural network based approach for prediction of 500 hPa geopotential and 850 hPa temperature, transforming the gridded information on an iconohexadral grid and back to a latitude-longitude grid as output and were able to achieve good results for the first days.~~

~~Deep learning has been used also in the context of extreme weather forecasting. Liu et al. (2016) developed a multi-channel CNN model 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 CNN architecture to classify extreme weather events spatially as well as temporally. 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. 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 definition of the term ’extreme event’ can vary substantially~~ While resampling strategies have also been proposed for imbalanced regression tasks (see e.g. Oliveira et al. (2021) for an application in the spatio-temporal ~~context, however. In the literature, extreme events often refer to hazardous weather *patterns*, present over some spatial or spatio-temporal domain. While these events are certainly extreme within the underlying climatology of the study, they are not usually extreme with respect to the data distribution used for the study. For example, many classification studies of extreme weather patterns ensure that the model is supplied with an equal number of negative and positive samples, so as to avoid any 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), one-class classification approaches (e.g. Deng et al., 2018; Goyal et al., 202 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. highly underrepresented values in the data that are therefore rarely encountered during model training. Regression problems on imbalanced data distributions are termed imbalanced regression problems. Ding et al. (2019) provide a formal analysis on why DL regression models suffer from overfitting and underfitting problems when data is imbalanced and~~ setting), the machine learning literature on imbalanced regression tends to treat the problem as either anomaly detection (see e.g. Schmidl et al. (2022) for a review) or by changing the loss function utilised during model learning. In the latter context Ding et al. (2019) propose a novel loss function based on extreme value theory, 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~~ notion of ’relevance functions’. Yang et al. (2021), on the other hand, proposed the idea of distribution smoothing to 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 can then be used easily for re-weighting methods, where the loss function can be 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.~~ by Shi et al. (2017).

~~In this paper , we propose~~ Furthermore, a lot of work has been done in recent years on probabilistic weather forecasting and many postprocessing methods have been proposed to improve probabilistic forecasts. Postprocessing is typically applied to ensemble weather- or e.g. energy forecasts and attempts to correct biases exhibited by the system and improve overall performance (see e.g. Phipps et al. (2022)) but has been explored to a lesser degree in the context of extreme event prediction. One approach to postprocess ensemble forecasts for extreme events is to utilise extreme-value theory, a review of which can be found in Friederichs et al. (2018). The authors propose separately postprocessing toward the tail distribution and formulate a postprocessing approach for the spatial prediction of wind gusts. Other authors have explored the potential of ML in this context. Ji et al. (2022), for example, investigate two DL-based postprocessing approaches for ensemble precipitation forecasts and compare these against the censored and shifted gamma distribution-based ensemble model output statistics (CSG EMOS) method. The authors report significant improvements of the DL-based approaches over the CSG EMOS and the raw ensemble, particularly for extreme precipitation events. Ashkboos et al. (2022) introduce a 10-ensemble dataset of several atmospheric variables for ML-based postprocessing purposes and compare a set of baselines in their ability to correct forecasts, including extreme events. Alessandrini et al. (2019), on the other hand, demonstrate improved predictions on the right tail of the forecast distribution of analog ensemble (AnEn) wind speed forecasts using a novel bias-correction method based on linear regression analysis, while Williams et al. (2014) show that flexible bias-correction schemes can be incorporated into standard postprocessing methods, yielding considerable improvements in skill when forecasting extreme events.

As data-driven forecasting model this paper investigates an adaptation of a deep convolutional LSTM (ConvLSTM) ~~model for~~ regression model, as proposed by Shi et al. (2015) for precipitation nowcasting in the range of 0–6 hours. The capability of deep ANNs to automatically and effectively learn hierarchical feature representations from raw input data have made DL models particularly attractive to the area of spatio-temporal sequence forecasting, where complex spatial and temporal correlations are typically present in the data (Shi and Yeung, 2018; Amato et al., 2020; Wang et al., 2020). The ConvLSTM is an example of a ConvRNN model, which forms a synthesis of a convolutional neural network (CNN) and a recurrent neural network (RNN). 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), while RNNs are known for their powerful ability to effectively model temporal dependencies (Shi et al., 2015)

. By utilising the strengths of the CNN to capture spatial correlations and the RNN to capture temporal correlations in the data, ConvRNN models have demonstrated very promising forecasting ability in the spatio-temporal setting (Wang et al., 2020), outperforming both non-recurrent convolutional models, as well as non-convolutional LSTM models (Shi et al., 2015, 2017). As a multi-layered ConvRNN model, the deep ConvLSTM thus has the potential to effectively model the complex dynamics of the spatio-temporal wind speed forecasting problem.

In this paper an adaptation of a deep ConvLSTM regression model is applied to the task of extreme wind speed prediction~~,~~ ~~adapted with~~ . The model is adapted with different types of imbalanced regression loss ~~to account for the heavy tails . To this end, we investigate the inversely weighted mean absolute error (W-MAE), the inversely weighted mean squared error (W-MSE) and the squared error relevance area (SERA) loss functions. The performance of our adapted model is compared against the~~ and their efficacy in improving predictions on the tails of the local wind speed distributions at each coordinate is compared. As such, this paper attempts to shed light on how the loss function of a deep learning model may be best adapted to improve forecasting performance on the distributional tails. Such improvement has practical relevance to wind energy applications where obtaining accurate predictions of extreme events are more desirable than accurate predictions of non-extremes e.g. in early-warning systems for wind farm operators. It is important to note, however, that while the local distributional tails in this work do not necessarily denote severe events in the absolute sense, the methodology of this work can be translated directly to cases where distributional tails denote actual hazardous events. The adapted models are, furthermore, compared against two base-line models, trained with standard mean absolute error (MAE) and mean squared error (MSE) ~~losses, as~~ loss. Forecast quality of all models is determined from a ~~spatio-temporal forecast verification using the symmetric extremal dependency index (SEDI).~~ combination of categorical and continuous scores over a variety of intensity thresholds.

2 Methodology

2.1 Data Collection and Preprocessing

The wind speed data used in this work was downloaded from the Copernicus Climate Change Service Climate Data Store (CDS) of the ECMWF (see Hersbach et al., 2018). ~~The reanalysis dataof the~~ Different vertical levels are available of the ERA5 data. In this study, the focus lies on the 1000 hPa pressure level data which typically varies between 100 and 130 m above ground level, corresponding to the most common hub heights in the eastern, flat part of Austria (main wind energy region). Not shown in this study are results of the surface wind speed and other pressure levels. The U and V components of the horizontal wind velocity (in m s^{-1}) were taken at 1000 hPa from the *ERA5 hourly data on pressure levels from 1979 to present* ~~dataset. By computing the square root of the sum of the squares of the two wind velocity components the~~ to calculate the scalar wind speed~~was obtained~~. The data was collected with a temporal resolution of one hour between 01 January 1979 and 01 January 2021 (42 years) on a spatial grid over central Europe. Of these data, the last two years between 2019-2021 were held out as testset. The eight years between 2011-2019 were used for training and validation in the first part of the experiment, ~~dedicated to model optimisation~~using 4-fold cross validation (with six years training and two years validation data) to determining optimal model architecture for each of the investigated loss functions. In the second part of the experiment the optimal ~~models~~ model

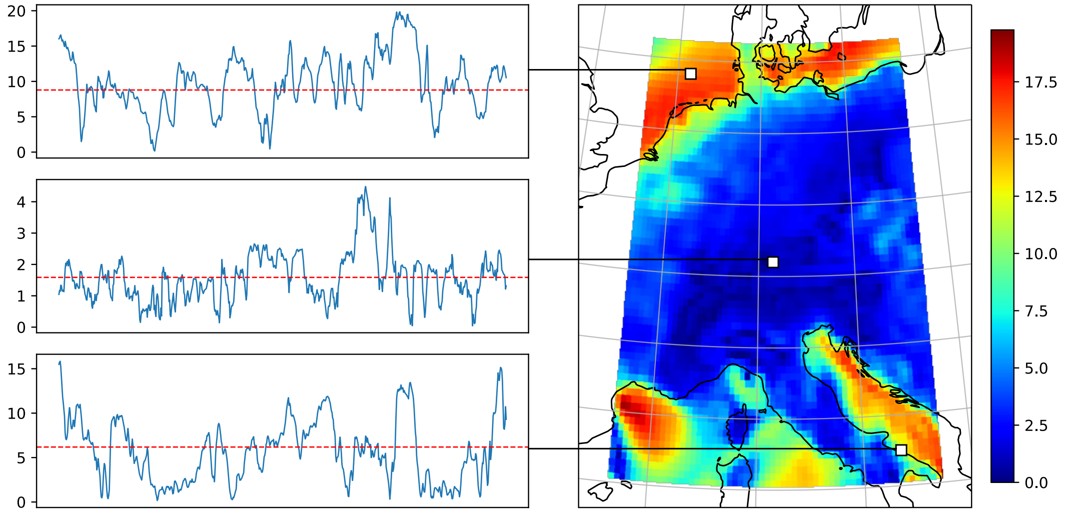


Figure 1. A visualisation of the wind speed data (in m s^{-1}). The right figure shows a color-map of an example data frame, overlaid on a cartographic map (Central Europe) showing the coastlines of the region. On the left, the wind speed time series of three arbitrary locations (white squares) within the frame are plotted for the duration of one month, as well as the climatological means at these locations (dotted red lines).

architectures were then trained and validated on the entire 40 years of data between 1979-2019, using the ~~years between 2017-2019~~ eight years between 2011-2019 as validation.

The spatial grid comprises 64×64 grid points between 40–56° N and 3–19° E, the spatial resolution being 0.25° ($\approx$ 28 km). 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 is expected to be particularly promising. Moreover, ~~we expect~~ the fine spatial resolution of 0.25° is expected to be critical to capturing the complex fine-scale dynamics of a variable like low-level wind, and thus improving forecasting ability~~, while the~~. The resolution also marks an important step forward for data-driven models to be truly competitive with state-of-the-art numerical weather prediction models, which are run at $\approx 0.1°$ resolution (Pathak et al., 2022).

A visualisation of the data is provided in Fig. 1. The figure shows on the right an example time slice and on the left wind speed time series of three arbitrary locations over the duration of one month, including the climatological means at these locations. Evidently, the local climatological means (and by extension, the local wind speed distributions) vary substantially throughout the region, where striking differences in magnitude can be ~~seen~~ observed between the off-shore and on-shore regions. To highlight these spatial differences, Fig. 2 shows the maximum, mean and standard deviation of the wind speed over the region, which unveil a sharp division of the statistics with the underlying coastlines of the region. Indeed, extreme winds

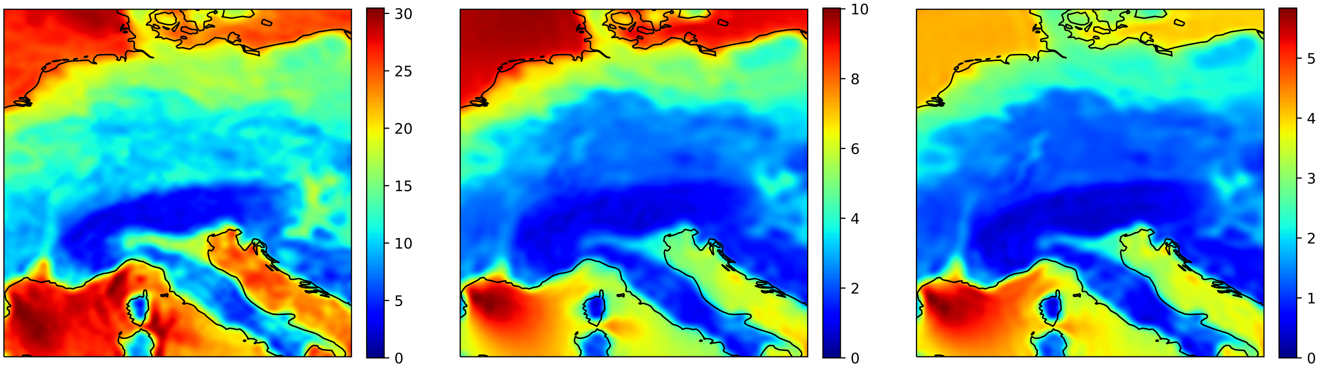


Figure 2. Color-maps of the maximum (left), mean (center) and standard deviation (right) of the wind speeds (in m s^{-1}) over the region. The figures display a sharp division of the statistics along the coastlines.

(e.g. larger than 25 m s^{-1}) seem to occur almost exclusively off-shore. If there were, in fact, stronger winds present over this region of mainland Europe between 1979 and 2021 then they have not been captured by the hourly ERA5 reanalysis.

~~Rather~~ Thus, rather than defining extreme winds in terms of their absolute severity, ~~we proceed to define extreme winds , instead,~~ extreme winds are here defined in terms of their *relative rarity* at each coordinate. This definition focuses the forecasting problem on the tails of the respective distributions at each coordinate, which ensures that the forecasting of extremes is conducted over the entire region, rather than only locally over some particularly dominant area. By selecting a distributional percentile (e.g. the ~~99th~~ 99th percentile), ~~we then define extreme winds~~ extreme winds are then defined as those wind speeds surpassing ~~this~~ the percentile threshold of the wind speed sample distribution at the respective coordinate ~~. To this end, the raw wind speed data were standardised with a local Z-normalisation at~~ i.e. wind speeds that are, indeed, rare at that coordinate (although not necessarily severe or hazardous in a absolute sense). For the remainder of this paper, the term ’p^{th} percentile threshold’ refers always to the p^{th} percentile at each coordinate ~~, which centers each local distribution around zero mean with unit standard deviation according to the following transformation:~~

$$\mathbf{x}'_{i,j} = \frac{\mathbf{x}_{i,j} - \mu_{i,j}}{\sigma_{i,j}}$$

~~where $\mathbf{x}'_{i,j}$ denotes the transformed variable, $\mathbf{x}_{i,j}$ the original variable and $\mu_{i,j}$ and $\sigma_{i,j}$ the mean and standard deviation, respectively, at the coordinate i,j. The Z-normalised data thus represent the wind speed~~ of the target observation field. The above approach allows us to investigate forecasting improvements of extreme events more generally by looking at improvements on the tails of the respective distributions (in terms of ~~the number of standard deviations from the respective mean~~ percentiles), regardless of the absolute values of the tails. Any improvements on the tails that result from the loss function modifications investigated in this paper can be swiftly translated to other cases where the tails of the distributions denote actual hazardous events.

Finally, the data were preprocessed at each coordinate ~~.~~ using a Yeo-Johnson power transform (Yeo and Johnson, 2000) to make the local wind speed distributions more Gaussian-like and were subsequently standardised locally using zero-mean, unit-variance normalisation. The optimal parameter for stabilising variance and minimising skewness in the power transform was estimated through maximum likelihood.

2.2 Model Description

The model implemented ~~and adapted~~ for the task of spatio-temporal forecasting of ~~extreme events~~ wind speed is an adaptation of the convolutional long short-term memory (ConvLSTM) network, as proposed by Shi et al. (2015) for precipitation nowcasting. However, while Shi et al. (2015) trained their ConvLSTM model using cross-entropy loss, ~~we propose adapting the model to~~ the model proposed here adjusts the ConvLSTM to the forecasting of extreme events by utilising two types of loss functions 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, which forms a synthesis of a convolutional neural network (CNN) and a recurrent neural network (RNN). 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), while RNNs are known for their powerful ability to effectively model temporal dependencies (Shi et al., 2015). By utilising the strengths of the CNN to capture spatial correlations and the RNN to capture temporal correlations in the data, ConvRNN models have demonstrated promising forecasting ability in the spatio-temporal setting. As a deep ConvRNN model, the ConvLSTM has the potential to effectively model the complex dynamics of the spatio-temporal wind speed forecasting problem.~~ The focus in this regard is set on providing a comparison of the ConvLSTM adapted with different variants of weighted loss and SERA loss.

~~We adopt the~~ The deep ConvLSTM model architecture is adopted with an encoding–forecasting network structure ~~, as is~~ (common for spatio-temporal sequence forecasting~~,~~) where both encoding and forecasting networks consist of several stacked ConvLSTM layers. As depicted in Fig. 3, 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. ~~We implement the model~~ The model is implemented as a multi-frame forecasting model, with 12 hour input and 12 hour prediction. This means that the model takes in tensors of size $(12 \times 64 \times 64)$ as input, consisting of the previous 12 hours of wind speed over the (64×64) grid, which are then encoded all together through various hidden states of the encoding network and decoded through the decoding network into a subsequent 12-hour prediction tensor of size $(12 \times 64 \times 64)$.

2.2.1 ~~Inversely Weighted Loss~~

The model was implemented and trained using Pytorch (Paszke et al., 2019), the code of which can be found under: https://github.com/dscheepens/Deep-RNN-for-extreme-wind-speed-prediction. In addition to the different loss functions, different model architectures with different numbers of ConvLSTM layers are investigated, ranging from 2–5 layers (in both the encoder and the forecasting networks). The numbers of parameters of all model architectures are shown in Table 1. In line with Shi et al. (2015), all layers utilise 3×3 kernels. The convolution over each successive filter operates such as to

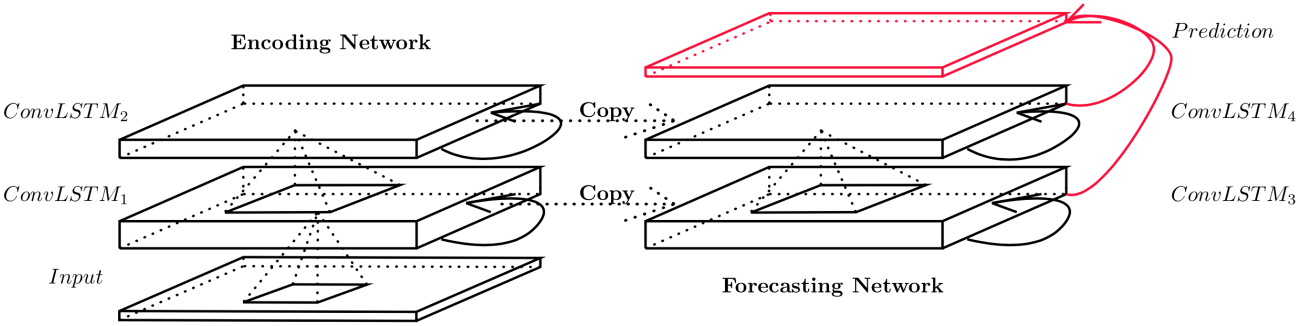


Figure 3. The multi-layered encoding–forecasting ConvLSTM network. The hidden states and cell outputs of the encoding network are copied to the forecasting network, from which the final prediction is made. **© Shi et al. (2015). Used with permission.**

Table 1. The number of parameters of the ConvLSTM model with different numbers of layers.

ConvLSTM layers	Number of parameters
2	2,385,953
3	10,061,025
4	34,201,185
5	62,060,641

successively halve the spatial dimensions of the input, while the number of hidden states (features) are successively doubled (starting from 16 hidden states).

The models are trained using mini-batch gradient descent with a batch-size of 16 and used the adaptive moment estimation (Adam) as 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 (see e.g. Ruder, 2017). As in Shi et al. (2017) , the initial learning rate of the Adam optimiser is set to 10^{-4}. During training, early-stopping is performed on the validation set to ensure that the model with the lowest validation loss is saved as the best model and thus to avoid overfitting the model. The early-stopping mechanism is set up to stop training when the validation loss fails to decrease for 20 consecutive epochs.

These implementation and parameter choices were selected a priori based on the work of Shi et al. (2015) and Shi et al. (2017) . Model performance may certainly be improved by performing a thorough hyper-parameter optimisation but that is not the focus of this paper. The focus is set, instead, on the different loss functions proposed in the literature for spatio-temporal imbalanced regression using deep learning and compare these in terms of their improvement in the prediction of spatio-temporal wind speed extremes using the ConvLSTM model.

2.2.1 Weighted Loss

In order to combat the effects of data imbalance on the imbalanced regression problem, ~~we adapt~~ the ConvLSTM model is adapted with two different types of loss functions that have been proposed for imbalanced regression problems. The first of these is the relatively simple weighted loss, which consists of assigning a weight $w(y)$ to each value in the input frame according to its target wind speed y. For a loss function L of the target y and prediction $\hat{y}$ (consisting of N time-frames of $M \times M$ spatial coordinates) and a weighting function $w(y)$, the weighted loss L_W is computed as ~~:~~

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

in Eq. 1. As weighted loss functions ~~we investigate~~ both the weighted mean squared error (W-MSE) loss and the weighted mean absolute error (W-MAE) loss ~~. We proceed to compute~~ are investigated.

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

As weighting function both an inverse weighting function as well as a simple linear weighting function are investigated. The inverse weighting function computes the weights in proportion to the inverse of of the data distribution for each target, as suggested by Yang et al. (2021). For a continuous target distribution, this typically implies discretising the distribution into intervals (see e.g. Shi et al., 2017), where all predictions within an interval are weighted by the same weight. Due to our definition of extreme events in terms of local percentile thresholds we proceed to discretise the target distribution into intervals spanning the percentage of the distribution between percentile p and 100. For a set of increasing percentiles $\mathcal{P} = \{p_k\}$, all targets $p_k \leq y < p_{k+1}$ are then weighted proportionally to the inverse of the percentage between p_k and 100 i.e. $w(y) \propto 1/(100 - p_k)$. We utilise a range of integer percentiles $\mathcal{P} = \{p_k | k \in [50, 99]\}$ and normalise weights such that the interval between percentiles 50 and 51 is given unit weight. As such, weights increase inversely from 1 up until a weight of 50 (given to target values $y_{99} \leq y \leq y_{100}$). All values smaller than the ~~50th~~ 50th percentile (p_{50}) are also given unit weight. This results in the weighting function shown in Eq. ~~??~~2, which is also presented graphically in Fig. ~~??~~4.

$$w_{inv}(y) = \begin{cases} 1 & \text{if} \quad y < p_{50} \\ 50 \cdot \frac{1}{100-k} & \text{if} \quad p_k \leq y < p_{k+1} \quad \text{for} \quad k \in [50, 99] \end{cases} \qquad (2)$$

The linear weighting function is constructed analogously as shown in Eq. 3. Target values $y < p_{50}$ are similarly given unit weight, while weights for target values $p_k \leq y \leq p_{k+1}$ are increased linearly from 1 to 50 for percentiles $k \in [50, 99]$.

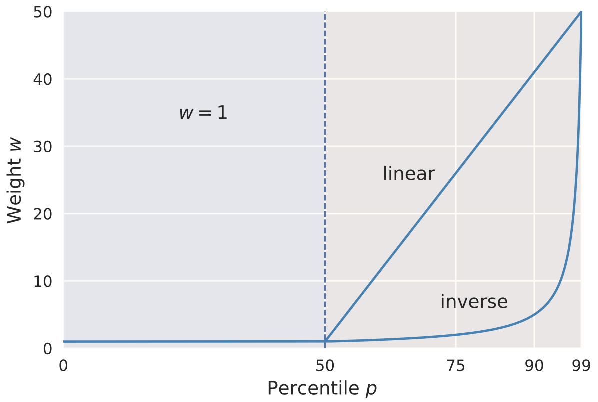


Figure 4. Weighting ~~function w~~ functions used to construct either the inversely weighted mean squared error (W-MSE$_{inv}$) and ~~inversely~~ mean absolute error (W-MAE$_{inv}$) or the linearly weighted mean squared error (W-MSE$_{lin}$) and mean absolute error (W-MAE$_{lin}$).

$$w_{lin}(y) = \begin{cases} 1 & \text{if} \quad y < p_{50} \\ k-49 & \text{if} \quad p_k \leq y < p_{k+1} \quad \text{for} \quad k \in [50,99] \end{cases} \tag{3}$$

2.2.2 Squared Error-Relevance Area Loss

As a second approach to combating data imbalance, ~~we investigate~~ the squared error-relevance area (SERA) loss is investigated, as proposed by Ribeiro and Moniz (2020). 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 90th percentile (p_{90}) of the standardised wind speed distribution~~

In this implementation the method is implemented on a per-coordinate basis and the local 99th percentile (p_{99}) at each coordinate is fixed as the point of ~~minimum relevance at that coordinate $\langle y_1 = p_{90},\ \phi(y_1) = 0.0,\ \phi'(y_1) = 0.0\rangle$ and the 99th percentile (p_{99})as the point of maximum relevance $\langle y_1 = p_{99},\ \phi(y_1) = 1.0,\ \phi'(y_1) = 0.0\rangle$~~maximum relevance, but the point of minimum relevance is varied between either the 90th percentile (p_{90}), the 75th percentile (p_{75}) or the 50th percentile (p_{50}), in order to get a better idea of how this choice affects forecasting performance. The interpolation in all cases is carried out according to Ribeiro and Moniz (2020) by using the piecewise cubic Hermite interpolating polynomials (*pchip*) algorithm~~and the~~. The obtained relevance function for control-points p_{90} and p_{99} is shown in Fig. 5.

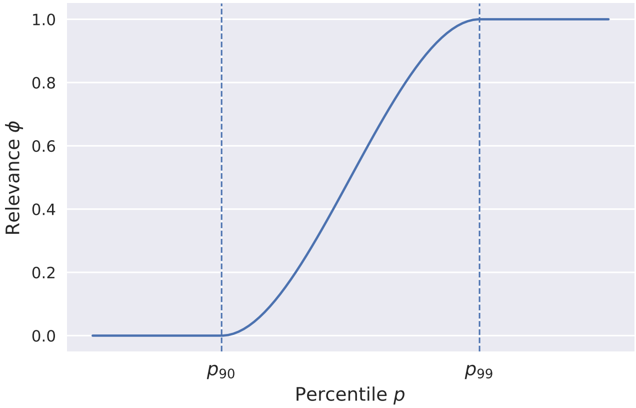


Figure 5. The relevance function ϕ obtained by interpolating the ~~90th~~ 90th percentile (p_{90}) as control-point of minimum relevance and the ~~99th~~ 99th percentile (p_{99}) as control-point of maximum relevance, using the *pchip* interpolation algorithm of Ribeiro and Moniz (2020).

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:

$$SER_t = \sum_{i \in D^t} (\hat{y}_i - y_i)^2 \tag{4}$$

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:

$$SERA = \int_0^1 SER_t \, dt \tag{5}$$

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. We also note that regardless of the choice of relevance function, the SERA loss utilises *all* given samples in its computation, as the integral in Eq. 5 starts at $t = 0$ and $SER_{t=0}$ denotes all samples with relevance values greater or equal than 0 i.e. all samples.

2.2.3 ~~Implementation~~

2.3 Forecast Evaluation

~~The model was implemented and trained using Pytorch, the code of which can be found under: . In addition to 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 numbers of parameters of all model architectures are shown in Table 1. 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 number of parameters of the ConvLSTM model with different numbers of layers. ConvLSTM layers Number of parameters 2 2, 385~~In order to evaluate the predictions of the ConvLSTM against observation, the model's hit rate ($H = \frac{a}{a+c}$), the false alarm ratio (FAR $= \frac{b}{a+b}$), ~~953 3 10,061,025 4 34,201,185 5 62,060,641~~

~~We trained our models using mini-batch gradient descent with a batch-size of 16 and used the adaptive moment estimation (Adam) as 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 (see e.g. Ruder, 2017). The initial learning rate of the 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 and thus to avoid overfitting the model. The early-stopping mechanism was set up to stop training when the validation loss failed to decrease for 20 consecutive epochs.~~

2.4 ~~Verification~~

~~Since the ConvLSTM model investigatedin this study is a spatio-temporal forecasting model, it appears in order to evaluate the model with a verification method that captures forecasting ability at different temporal, as well as spatial scales. In order to evaluate the model at different spatial scales we utilise the minimum coverage method, as proposed by Damrath (2004). As a filtering method, the minimum coverage method works well for verifying 'messy' forecasts that do not contain well-defined features (Ebert, 2009), which we expect to be particularly applicable to wind speed due to its highly stochastic behaviour. Another advantage of the method is that it is parameter-free and easy to implement. While the method is a spatial forecast verification method, it can be applied in a simple manner *per* lead-time for a temporal assessment.~~

~~The minimum coverage 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:~~

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

~~where P_e is the minimum fraction of~~ threat score (TS$= \frac{a}{a+b+c}$) and 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. Due to the scarcity of extreme events in the data, we chose 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 score used here is the symmetric extremal dependence index (SEDI), which is computed as follows:~~

$$\mathrm{SEDI} = \frac{\log F - \log H + \log(1-H) - \log(1-F)}{\log F + \log H + \log(1-H) + \log(1-F)} \in [-1,1].$$

frequency bias ($B = \frac{a+b}{a+c}$) are investigated, where ~~H and F are~~ a denotes the number of hits, b the number of false alarms, c the number of missed hits and d the number of correct negatives obtained by the ~~hit rate and false alarm rate, respectively. The SEDI was chosen for its unique property of non-degeneracy for rare events. Stephenson et al. (2008) have shown that practically all categorical scores degenerate to trivial values such as 0, 1 or infinity for exceedingly rare events, i.e. as the base rate of the event tends to zero. The SEDI was proposed by Ferro and Stephenson (2011) as a remedy to the degeneracy problem and, in fact, combines more desirable properties into one score than any other categorical score (Hogan and Mason, 2012, p. 54).~~ model. The hit rate, false alarm ratio and threat score are routinely used by the UK Met Office to evaluate warnings (Hogan and Mason, 2012) and have also been used by Shi et al. (2015) to evaluate the ConvLSTM model for precipitation nowcasting, while the frequency bias provides valuable information on whether the model tends to overforecasting or underforecasting.

~~As is typical for neighbourhood methods, scores were~~ These scores are computed for a set of ~~scales and a set of intensity thresholds in order to provide a diagnostic assessment of forecast quality on spatial scale and intensity (see: e.g. Ebert, 2009). As such, we computed the SEDI for a set of scales corresponding to approx. 28, 83, 139, 194~~ intensity thresholds corresponding to the local 50th, 75th, 90th, 95th, 99th and ~~250 km, and a set of thresholds corresponding to the local 50th, 75th, 90th, 95th, 99th and~~ 99.9~~th~~ th percentiles of the ~~standardised wind speed distribution~~ observed sample distributions at each coordinate~~: We remind the reader that the models are judged on their ability to forecast extreme events in terms of *relative rarity*, which we measure as an event's percentile with respect to the local frequency distribution at the respective coordinate. Finally, in~~ , which are computed using the training set. In order to obtain an aggregated result over all forecasts made by a model, the elements in the 2 × 2 contingency table are aggregated over all forecasts and the scores are computed subsequently from the aggregated contingency table.

~~In~~ Since the above categorical scores work on the basis of a forecast being correct as long as it surpasses the same threshold t as the observed event, they are able to give an indication of the *frequency* of errors while unable to give an indication of the *magnitude* of the errors between forecast and observation. In order to include in the analysis a comparison of error magnitudes, the root-mean-square error (RMSE) between (continuous-valued) predictions and observations is utilised. Unlike the categorical scores, the RMSE is here computed between two consecutive percentile thresholds: For a particular 12 hour forecast and observation, and thresholds p_1 and p_2, the RMSE is computed between all pairs of forecast and observation values (f, o) where the observation values lie between p_1 and p_2 i.e. $p_1 \leq o < p_2$. The total RMSE for those thresholds is then computed as an aggregate over all forecasts and observations of the model. This approach serves to give an indication of the typical magnitude of errors of the forecasts of a model over a particular value range of the observations.

In the next section the ~~next section we present the~~ results obtained from combining the multi-layered ConvLSTM network with ~~inversely weighted mean absolute error (W-MAE), inversely weighted mean squared error (W-MSE) and squared error-relevance area (SERA) loss and compare these against the standard mean absolute error (MAE) and mean squared error (MSE) loss~~ the various loss functions are presented. The optimal number of layers for each model is determined from the minimum validation-loss obtained by the networks as averaged over the 4-fold cross validation process ~~, as~~ (conducted over the ~~8~~ eight years of data between 2011–2019). The optimal models are then re-trained using the entire 40 years of data between 1979–2019 (using ~~2017-2019~~ the eight years between 2011-2019 as validation) and their results are compared on the held-out test set ~~(comprising the years~~ comprising the two years between 2019-2021~~) using the SEDI. Finally, the best performing model is further analysed with the minimum coverage method and its forecasts are visualised.~~ .

3 Results

~~We begin by showing, in Table 2 ,~~

3.1 Validation-loss

Table 2. Minimum validation loss as obtained by the ConvLSTM network with number of layers ranging from 2–5 (denoted in brackets) and trained with the various different loss functions. Values are presented as the mean ± one standard deviation from the 4-fold cross validation. The lowest minimum validation loss reached, and thus the optimal network architecture, is emphasised in boldface for each loss function. Where multiple architectures obtained the same minimum validation loss, the simpler architecture is preferred.

Loss	ConvLSTM (2)	ConvLSTM (3)	ConvLSTM (4)	ConvLSTM (5)
W-MAE$_{inv}$	$(65.1 \pm 2.2) \cdot 10^{-2}$	$(63.6 \pm 2.0) \cdot 10^{-2}$	$(\mathbf{63.3 \pm 2.1}) \cdot \mathbf{10^{-2}}$	$(63.3 \pm 2.1) \cdot 10^{-2}$
W-MSE$_{inv}$	$(52.0 \pm 1.6) \cdot 10^{-2}$	$(49.9 \pm 1.3) \cdot 10^{-2}$	$(\mathbf{49.5 \pm 1.8}) \cdot \mathbf{10^{-2}}$	$(49.6 \pm 1.6) \cdot 10^{-2}$
W-MAE$_{lin}$	$(249.3 \pm 3.9) \cdot 10^{-2}$	$(243.4 \pm 3.8) \cdot 10^{-2}$	$(243.3 \pm 3.4) \cdot 10^{-2}$	$(\mathbf{242.9 \pm 4.4}) \cdot \mathbf{10^{-2}}$
W-MSE$_{lin}$	$(148.3 \pm 3.3) \cdot 10^{-2}$	$(142.6 \pm 3.6) \cdot 10^{-2}$	$(\mathbf{142.5 \pm 3.0}) \cdot \mathbf{10^{-2}}$	$(142.5 \pm 2.3) \cdot 10^{-2}$
SERA$_{p90}$	$(116.2 \pm 4.1) \cdot 10^{-3}$	$(113.2 \pm 5.6) \cdot 10^{-3}$	$(113.1 \pm 4.5) \cdot 10^{-3}$	$(\mathbf{111.0 \pm 2.9}) \cdot \mathbf{10^{-3}}$
SERA$_{p75}$	$(125.2 \pm 1.6) \cdot 10^{-3}$	$(121.4 \pm 2.4) \cdot 10^{-3}$	$(119.6 \pm 2.8) \cdot 10^{-3}$	$(\mathbf{119.4 \pm 3.2}) \cdot \mathbf{10^{-3}}$
SERA$_{p50}$	$(136.6 \pm 1.8) \cdot 10^{-3}$	$(132.1 \pm 1.1) \cdot 10^{-3}$	$(130.8 \pm 3.1) \cdot 10^{-3}$	$(\mathbf{130.6 \pm 1.7}) \cdot \mathbf{10^{-3}}$
MAE	$(264.7 \pm 2.6) \cdot 10^{-3}$	$(257.4 \pm 2.6) \cdot 10^{-3}$	$(256.9 \pm 3.0) \cdot 10^{-3}$	$(\mathbf{256.1 \pm 2.9}) \cdot \mathbf{10^{-3}}$
MSE	$(213.2 \pm 2.6) \cdot 10^{-3}$	$(204.7 \pm 3.9) \cdot 10^{-3}$	$(204.4 \pm 3.2) \cdot 10^{-3}$	$(\mathbf{204.0 \pm 2.7}) \cdot \mathbf{10^{-3}}$

Table 2 shows the minimum validation-loss obtained by the ConvLSTM network with the number of layers ranging between 2–5~~to determine the optimal number of layers for each loss function, which is~~ , as trained with either inversely weighted loss (W-MAE$_{inv}$ and W-MSE$_{inv}$), linearly weighted loss (W-MAE$_{lin}$ and W-MSE$_{lin}$), SERA loss or standard MAE or MSE loss. The SERA loss is denoted with a subscript denoting the first control-point used, with the second control-point fixed at the local 99th percentile (p_{99}) for each coordinate. Results are shown as the mean ± one standard deviation from the 4-fold cross validation. The minimum validation-loss for each loss function has been emphasised in boldface~~. As described in the~~

~~methodology, the models were then trained once again using the full 40 years of data and the~~ , indicating the optimal number of network layers for each loss function. ~~These optimal models are compared in Table 3 over local intensity thresholds varying between the 50th and 99.9th percentiles, using the symmetric extremal dependency index (SEDI). The~~ In cases where the mean validation loss is equal for multiple numbers of layers, the smallest number of layers, and thus the simplest model, was given precedence.

3.2 Comparison over intensity thresholds

Table 3. Comparison of hit score (H), false alarm ratio (FAR), threat score (TS) and frequency bias (B) of the ConvLSTM network trained with the various different loss functions. Scores are presented for wind forecasts f and observations o exceeding local intensity thresholds varying between the 50th (p_{50}) and 99.9th ($p_{99.9}$) percentiles, aggregated over lead-time. The optimal number of network layers used for each loss function is given in brackets after the name of the loss function. The persistence forecast is included in the table for reference. For each intensity threshold, the best scores are emphasised in boldface (where applicable).

	$H\uparrow$						FAR $\downarrow$					
Loss (layers)	$f,o \geq p_{50}$	$f,o \geq p_{75}$	$f,o \geq p_{90}$	$f,o \geq p_{95}$	$f,o \geq p_{99}$	$f,o \geq p_{99.9}$	$f,o \geq p_{50}$	$f,o \geq p_{75}$	$f,o \geq p_{90}$	$f,o \geq p_{95}$	$f,o \geq p_{99}$	$f,o \geq p_{99.9}$
W-MAE$_{inv}$ (4)	0.866	0.858	0.809	0.761	0.583	0.262	0.178	0.291	0.381	0.432	0.473	0.427
W-MSE$_{inv}$ (4)	0.861	0.846	0.788	0.735	0.531	0.201	0.179	0.285	0.374	0.42	0.45	0.424
W-MAE$_{lin}$ (5)	**0.979**	0.885	0.712	0.612	0.408	0.18	0.351	0.343	0.286	0.292	0.289	**0.306**
W-MSE$_{lin}$ (4)	0.966	0.884	0.689	0.583	0.389	0.187	0.312	0.335	0.272	0.27	0.289	0.362
SERA$_{p90}$ (5)	0.814	0.871	**0.938**	**0.945**	**0.614**	0.215	0.175	0.36	0.602	0.716	0.608	0.419
SERA$_{p75}$ (5)	0.849	0.921	0.924	0.844	0.527	0.225	0.2	0.407	0.571	0.572	0.464	0.421
SERA$_{p50}$ (5)	0.907	**0.932**	0.828	0.712	0.467	0.188	0.245	0.424	0.454	0.436	0.394	0.355
MAE (5)	0.836	0.76	0.656	0.58	0.419	0.215	0.138	**0.177**	**0.214**	**0.242**	**0.279**	0.354
MSE (5)	0.819	0.755	0.652	0.565	0.371	0.142	**0.133**	0.187	0.234	0.257	0.282	0.321
Persistence	0.774	0.678	0.582	0.523	0.408	**0.268**	0.238	0.34	0.441	0.503	0.611	0.741
	TS $\uparrow$						B					
	$f,o \geq p_{50}$	$f,o \geq p_{75}$	$f,o \geq p_{90}$	$f,o \geq p_{95}$	$f,o \geq p_{99}$	$f,o \geq p_{99.9}$	$f,o \geq p_{50}$	$f,o \geq p_{75}$	$f,o \geq p_{90}$	$f,o \geq p_{95}$	$f,o \geq p_{99}$	$f,o \geq p_{99.9}$
W-MAE$_{inv}$ (4)	0.729	0.635	0.54	0.482	**0.383**	**0.219**	1.054	1.209	1.306	1.341	1.108	0.457
W-MSE$_{inv}$ (4)	0.725	0.633	0.536	0.48	0.37	0.175	1.048	1.182	1.258	1.267	0.966	0.348
W-MAE$_{lin}$ (5)	0.64	0.606	0.554	0.488	0.35	0.167	1.51	1.346	0.997	0.864	0.574	0.259
W-MSE$_{lin}$ (4)	0.671	0.612	0.548	0.479	0.336	0.169	1.404	1.328	0.946	0.799	0.546	0.293
SERA$_{p90}$ (5)	0.694	0.585	0.388	0.279	0.314	0.186	0.986	1.361	2.355	3.328	1.567	0.371
SERA$_{p75}$ (5)	0.7	0.565	0.414	0.397	0.362	0.193	1.06	1.552	2.153	1.973	0.983	0.389
SERA$_{p50}$ (5)	0.7	0.553	0.491	0.459	0.359	0.17	1.201	1.618	1.515	1.263	0.771	0.291
MAE (5)	**0.737**	**0.653**	**0.557**	**0.489**	0.361	0.192	0.97	0.924	0.835	0.765	0.582	0.332
MSE (5)	0.727	0.644	0.544	0.473	0.324	0.133	0.944	0.929	0.852	0.761	0.517	0.209
Persistence	0.623	0.503	0.399	0.342	0.248	0.152	1.016	1.027	1.041	1.052	1.049	1.035

The networks were then retrained on the entire dataset with the corresponding optimal number of network layers ~~used with each model is shown in brackets next to~~ (henceforth indicated in brackets after the name of the ~~loss function. It is evident from Table 3 that the usage of imbalanced regression loss results in superior SEDI scores for extreme events between the 75th~~

~~and 99th percentiles. While outperformed by the SERA loss,~~ respective loss function with which the network was trained). Table 3 shows a comparison of the ~~W-MAE and W-MSE also show clear improvements over the standard MAE and MSE loss from the 90th percentile onward. Indeed, the table shows how the usage of imbalanced regression loss manages to shift optimal performance towards the extreme intensity thresholds , as opposed to performance simply decreasing monotonically for increasingly rare events, as is the case for the standard MAE and MSE from the 75th percentile onward. In the table are included, for reference, the SEDI scores achieved by a simple~~ hit score (H), false alarm ratio (FAR), threat score (TS) and frequency bias (B) for wind forecasts f and observations o exceeding local intensity thresholds between the 50th (p_{50}) and the 99.9th ($p_{99.9}$) percentiles, aggregated over all lead-times. The persistence forecast, which ~~is a forecast consisting simply~~ simply consists of a repetition of the final ~~observation (input) frame. It is clear that the improvement offered by the imbalanced regression loss functions ceases around the 99.9th percentile threshold, where SEDI scores are comparable among all models, while being, in addition, only marginally better than persistence.~~ input frame, is included in the table for reference.

~~Minimum validation loss as obtained by the ConvLSTM network with number of layers ranging from 2–5 (denoted in brackets) and trained with either W-MAE, W-MSE, SERA, MSE or MAE loss. Values are presented as the mean ± one standard deviation from the 4-fold cross-validation. The optimal model for each loss function is emphasised in boldface.~~

	ConvLSTM (2)	ConvLSTM (3)	ConvLSTM (4)	ConvLSTM (5)
W-MAE	$(8.3 \pm 0.6) \cdot 10^{-2}$	$(8.2 \pm 0.6) \cdot 10^{-2}$	$(8.2 \pm 0.6) \cdot 10^{-2}$	$\mathbf{(8.1 \pm 0.5) \cdot 10^{-2}}$
W-MSE	$(8.2 \pm 0.7) \cdot 10^{-2}$	$(8.1 \pm 0.6) \cdot 10^{-2}$	$(8.0 \pm 0.7) \cdot 10^{-2}$	$\mathbf{(7.8 \pm 0.7) \cdot 10^{-2}}$
SERA	$(24.5 \pm 0.8) \cdot 10^{-3}$	$(24.2 \pm 1.0) \cdot 10^{-3}$	$(24.0 \pm 0.8) \cdot 10^{-3}$	$\mathbf{(23.9 \pm 1.1) \cdot 10^{-3}}$
MAE	$(24.3 \pm 0.4) \cdot 10^{-3}$	$(24.0 \pm 0.5) \cdot 10^{-3}$	$\mathbf{(23.6 \pm 0.4) \cdot 10^{-3}}$	$(23.7 \pm 0.4) \cdot 10^{-3}$
MSE	$(18.6 \pm 0.6) \cdot 10^{-3}$	$(18.3 \pm 0.5) \cdot 10^{-3}$	$(17.9 \pm 0.6) \cdot 10^{-3}$	$\mathbf{(17.6 \pm 0.4) \cdot 10^{-3}}$

~~Comparison of SEDI scores obtained by the ConvLSTM network trained with either W-MAE, W-MSE, SERA, MAE or MSE loss, presented for winds (y) exceeding local intensity thresholds varying between the 50th and 99.9th percentiles. The optimal number of network layers used for each loss function is given in brackets after the name of the loss function. The persistence forecast is included in the table for reference. The table shows that the usage of imbalanced regression loss allows to substantially improve forecasts of local wind speeds exceeding the 75th percentile threshold.~~

	$y \geq p_{50}$
W-MAE (5)	0.828
W-MSE (5)	0.801
SERA (5)	0.767
MAE (4)	**0.854**
MSE (5)	0.848
Persistence	0.689

~~We investigate the performance of these models further in Fig. 6, where the SEDI scores obtained by each modelare plotted per lead-time (in hours) for the 99th percentile intensity threshold. We, once again, include in this comparison the persistence~~

~~forecast for reference. The figure shows that the superior performance of the SERA loss over the W-MAE and W-MSE in Table 3 results from improved performance on lead-times beyond ca. eight hours; For lead-times below ca. eight hours performance of the W-MAE and W-MSE loss are very competitive with the SERA. ConvLSTM trained with standard MAE and MSE loss is certainly more informative that the persistence forecast, performance is inferior to all three imbalanced regression losses on all lead-times.~~

~~Comparison of SEDI scores obtained by the ConvLSTM network trained with either W-MAE, W-MSE, SERA, MAE or MSE loss, plotted over lead-time (in hours), for local extreme events of the 99th percentile intensity threshold. The optimal number of network layers used for each loss function is given in brackets after the name of the loss function. The label 'persistence' refers to the persistence forecast. The comparison shows that the superior scores obtained by the SERA loss in Table 3 are due in particular to its better performance on lead-times 6–12 hours.~~

~~In addition, in order to establish a spatial picture of forecast quality, we provide in Fig. **??** an intensity-scale diagram (see e.g. Casati et al., 2004; Ebert, 2008) of the SERA (5) model, which we highlight here due to its superior SEDI scores on intensity thresholds between the 90th and 99th percentiles in Table 3. The figure shows how the SEDI scores change both with varying intensity threshold as well as spatial scale. However, contrary to the expected behaviour~~ The table shows that the imbalanced regression losses generally result in significant increases in the hit rate as compared with the standard MAE or MSE loss, indicating that more of the true occurrences of the events were captured by the model. Any improvement in the hit rate is, however, accompanied by an increase in the false alarm ratio. This suggests that in order to capture more of the events, the models are invariably producing more false alarms. This behavior is particularly pronounced where there is substantial overcasting i.e. a frequency bias substantially greater than 1. This can be best noticed for the SERA$[p_{90}, p_{99}]$ model at the threshold p_{95}, where a massive frequency bias of 332.8% results in the model successfully capturing a spectacular 94.5% of true events (the hit rate) at the cost of 71.6% of forecasted events being false alarms (the false alarm ratio).

The threat scores can give an overall idea of forecasting performance ~~improving with increasingly coarser scale, SEDI scores in the diagram in fact decrease with increasing scale (with the sole exception of the scores obtained with the 50th percentile threshold).~~

~~Intensity-scale diagram of SEDI scores obtained by the 5-layered ConvLSTM network trained with squared error-relevance area (SERA) loss.~~

~~Lastly, we proceed to show two visualisations of the forecasts made by the different ConvLSTM models investigated in this paper, which serve to highlight their respective strengths and weaknesses. Figure 9 shows a target observation of a growing intensification of anomalous winds in the left of the frame, as well as the forecasts made by the respective models . This example highlights the striking difference between the utilisation of~~ and, as such, suggest that that the SERA trained models investigated here can only be considered superior to the MSE trained model for the p_{99} threshold (except for SERA$[p_{90}, p_{99}]$ which scores worse) and for the $p_{99.9}$ threshold. Compared to the MAE loss, however, the SERA trained models typically score worse threat scores for all thresholds, except for the SERA$[p_{75}, p_{99}]$ model which manages to be on-par at thresholds p_{99} and $p_{99.9}$. As a matter of fact, none of the models trained with imbalanced regression loss ~~versus the standard MAE and MSE loss - both of which fail to capture the intensity of the target extremes, although they do manage to capture the general pattern of the~~

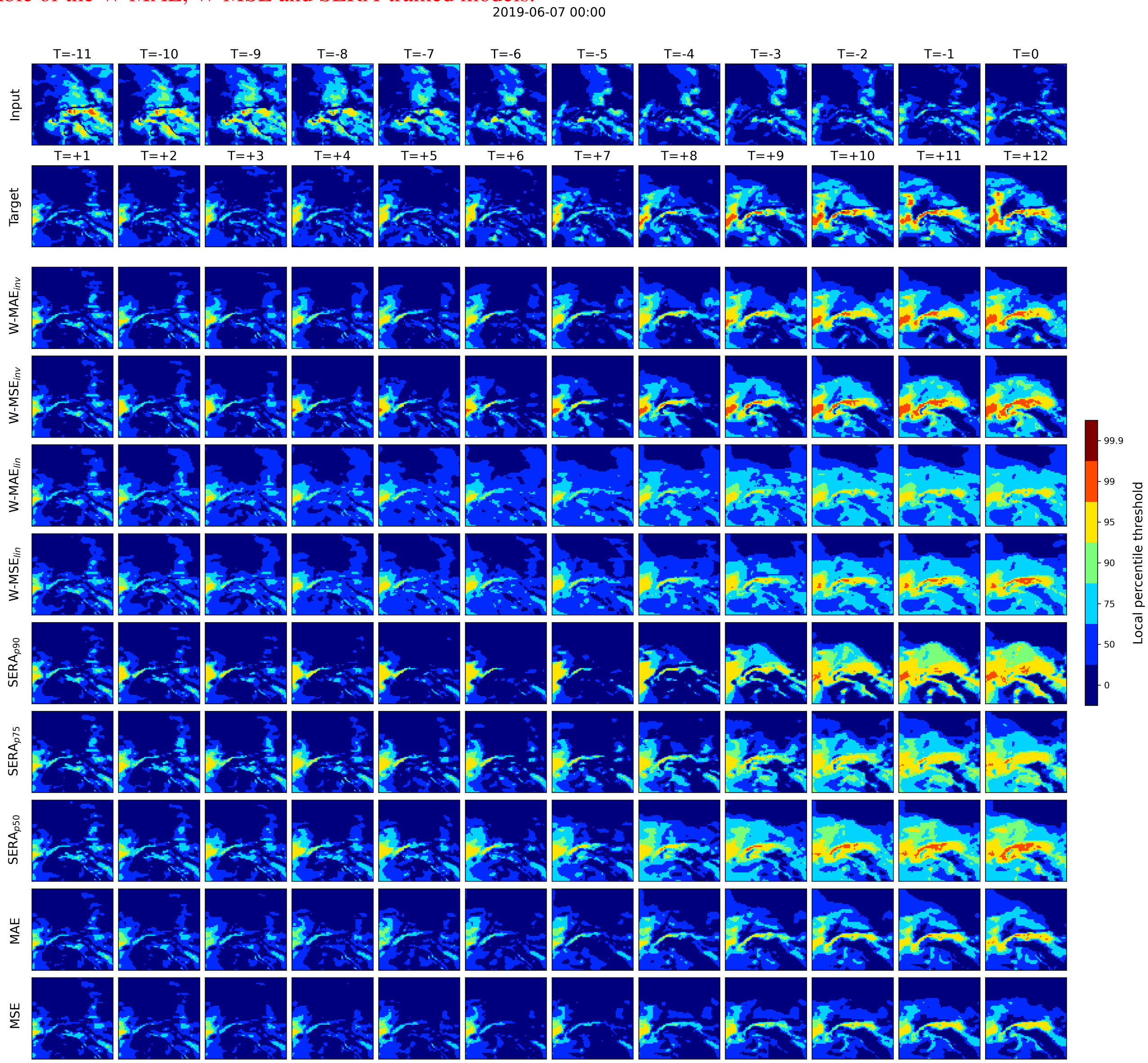
~~target observation. Among the imbalanced regression~~ achieve threat scores superior to the MAE trained model for thresholds p_{50}–p_{95}, although the inversely weighted losses generally achieve comparable scores and the linearly weighted losses achieve comparable scores for p_{90} and p_{95}. Between the linearly weighted losses, the ~~SERA loss provides a substantially coarser forecast of the extreme region than the~~ W-MAE$_{lin}$ achieves better scores for higher thresholds (p_{90} onward). The W-MAE$_{lin}$ also achieves slightly better scores than either inversely weighted losses for thresholds p_{90} and p_{95}. The inversely weighted losses dominate, however, for the extremely high thresholds p_{99} and ~~W-MSE, and, as such, allows for more false alarms in order to capture more of the event. This strategy can be clearly distinguished as well in Fig. 10, where the SERA-trained model severely overshoots its forecastto capture what is only a very minor event in the target observation~~$p_{99.9}$, outperforming all other loss functions on these thresholds. Performance for all models on threshold $p_{99.9}$ must be interpreted with caution, however, since threat scores on this threshold approach those obtained from the naive persistence forecast. Indeed, ~~the SERA loss appears to forecast something of a coarse-grained worst-case scenario, while the W-MAE and W-MSE forecasts are sharper and more conservative. These opposing characteristics lead to a strong suspicion that an ensemble of all three models (forecasting the average of the forecasts made by the models) may be worthwhile investigating further. These ensemble forecasts are included in Fig. 9 and Fig. 10 in~~ in terms of hit rate, none of the models investigated in this paper are able to successfully predict events of the ~~bottom row. While the ensemble forecast in Fig. 9 shows that some of the extreme intensities captured with the SERA lossare lost in averaging process, the ensemble forecast is significantly sharper spatially and continues to provide a substantial improvement over the MAE and MSE. The ensemble forecast in Fig. 10, furthermore, shows how the overshooting of the SERA-trained model is significantly limited and large swaths of false alarms avoided.~~ 99.9th percentile threshold better than persistence. Similarly, for the 99th percentile, the standard MAE and MSE and the linearly weighted MAE and MSE result in hit rates comparable to persistence, highlighting the failure of these loss functions in capturing extremely rare events.

~~A selection of further forecast visualisations can be found in the supplements, or in our GitHub repository: .~~

figures/fig08.png

~~An example forecast from the different ConvLSTM networks trained with either W-MAE, W-MSE, SERA or standard MAE or MSE loss. The first row from the top displays the 12 input frames, the second row the succeeding 12 target frames and the following rows the 12 predicted frames of the models. T refers to the index of the frame (in hours), with $T = 0$ denoting the~~

~~last input frame and $T = +12$ denoting the final target and prediction frames. The final row shows the averaged forecast of an ensemble of the W-MAE, W-MSE and SERA-trained models.~~



~~As Fig. 9~~

~~Another way to highlight the differences in forecasts between the different models is~~ As compared with the standard MAE loss, the W-MAE$_{inv}$ manages to boost the hit rate significantly across all intensity thresholds, while some degree of overcasting and increased false alarms will have to be allowed for. For p_{90}, for example, the usage of the W-MAE$_{inv}$ achieves an increase in H from 0.656 (standard MAE) to 0.809, with the FAR rising from 0.214 to ~~look at frequency bias, which is presented in~~

~~Table ?? for the same set of intensity thresholds as before. The tendency of the SERA-trained model to severely overshoot the target observation with large swaths~~ 0.381. Even for p_{99}, a significant increase in H is achieved, from 0.419 (standard MAE) to 0.583, with the FAR rising more drastically, however, from 0.279 to 0.473. Overcasting and FAR values can be reduced substantially, however, by using the linear weighting method. For p_{90} events, the W-MAE$_{lin}$ increases the FAR more conservatively from 0.214 (standard MAE) to 0.286 while still boosting H from 0.656 to 0.712. The table shows that a small boost in H can still be expected for p_{95} events, but beyond that, the linearly weighted MAE or MSE offer no improvements (with hit rates dropping to values comparable with persistence). Depending then on what magnitude of false alarms ~~is reflected in Table ?? by substantially increased frequency bias for extreme events between the 75th and 99th percentiles, as compared with the other models. While the W-MAE and W-MSE do present higher frequency bias than their unweightedcounterparts, these differences are comparatively minor. As expected, the ensemble forecast significantly limits the frequency bias for all events, as the different biases of the models counteract one another. The low frequency bias of all models in forecasting extreme events of~~ that are acceptable, and for what percentile of extreme events the ~~99.9th percentile threshold offers another clue as to why the SEDI scoresof~~ loss function is desired to offer improvement, either the linear or the inverse weighting methods can be utilised. Between all usages of the MAE and MSE, either weighted or unweighted, the ~~models for such events are so poor: these events are less frequently forecasted at all and are thus more often missed~~MAE returns higher hit rates and higher threat scores at the cost of an increased false alarm ratio and an increased frequency bias.

~~Given its continued widespread usage, we present in Table ?? also the root mean squared~~ Compared to the weighted loss functions, the SERA offers something of an extreme case, allowing hit rates to be boosted spectacularly but at a considerable loss of forecasting performance (as judged by reduced threat scores, increased overcasting and increased false alarms). The first control-point of the SERA loss does offer a way to mitigate this behaviour, however. For example, reducing this control-point from p_{90} to p_{75} and then to p_{50} (while keeping the second control-point fixed at p_{99}) shows a striking reduction in frequency bias, false alarms and hit rates for intensity thresholds between p_{90}–p_{99}, while threat scores, and thus overall forecasting performance, generally improves.

Table 4. As Table 3 but presented for the root-mean-square error (RMSE), which is computed between all pairs of forecast and observation values (f, o) where the observation values lie between p_1 and p_2 i.e. $p_1 \leq o < p_2$.

	RMSE ↓					
	$p_{50} \leq o < p_{75}$	$p_{75} \leq o < p_{90}$	$p_{90} \leq o < p_{95}$	$p_{95} \leq o < p_{99}$	$p_{99} \leq o < p_{99.9}$	$p_{99.9} \leq o < p_{100}$
W-MAE$_{inv}$ (4)	0.508	0.442	0.393	0.369	0.415	0.731
W-MSE$_{inv}$ (4)	0.508	0.446	0.397	0.38	0.445	0.767
W-MAE$_{lin}$ (5)	**0.4**	**0.293**	0.328	0.392	0.532	0.867
W-MSE$_{lin}$ (4)	0.402	0.295	0.338	0.411	0.554	0.892
SERA$_{p90}$ (5)	0.738	0.662	0.475	0.306	**0.29**	**0.642**
SERA$_{p75}$ (5)	0.668	0.51	0.339	**0.276**	0.361	0.706
SERA$_{p50}$ (5)	0.571	0.386	**0.3**	0.314	0.423	0.74
MAE (5)	0.456	0.463	0.475	0.504	0.615	0.933
MSE (5)	0.472	0.484	0.492	0.522	0.639	1.017
Persistence	0.731	0.758	0.786	0.826	0.93	1.243

Similarly to Table 3, Table 4 shows the root-mean-square error (RMSE) obtained ~~by the different models, as aggregated over all forecasts in the testset. The substantial frequency bias of the SERA-trained model in its forecasting of extreme events is reflected here by a substantially increased RMSE , as errors between extremes and non-extremes occur more often due to common misplacement of extreme event forecasts. The RMSE of the ensemble model highlights the effectiveness of the ensemble in limiting these large-scale errors produced by the SERA-trained model. Although the aggregated RMSE scores tell us nothing about the quality of forecasts of extreme events, they do give us an overall picture of the typical magnitude of errors made by the models~~from the continuous forecasts and observations of the different models. Unlike Table 3, the RMSE is computed between all pairs of forecast and observation values (f, o) where the observation values lie between p_1 and p_2 i.e. $p_1 \leq o < p_2$. Again, the persistence forecast is included for reference. The table shows that the imbalanced regression losses tend to result in lower RMSE scores, as compared to the standard MAE and MSE loss, for increasingly rare observation values. It is interesting to note that while the SERA trained models generally appear to result in heavy overcasting and highly inflated false alarm rates (Table 3), the RMSE scores suggest that increasing the first control-point of the SERA loss results in shifting the domain of minimal RMSE towards the higher percentiles. Between the inverse and linear weighting methods, the RMSE scores echo the interpretations from Table 3, with the linear method appearing more adapt (lower RMSE) around the central percentiles and the inverse method more adapt towards the higher percentiles.

3.3 Temporal assessment

~~Finally, analogously to Fig. ??, we provide~~ The performance of the models is investigated further in Fig. ~~?? an intensity-scale diagram of the SEDI~~ 6, where the scores obtained by ~~the ensemble model . From comparing the figures on the finest scale (28 km) it is clear that the ensemble is able to substantially improve the SEDI scores on the 50th–90th percentile thresholds, now comparable with SEDI scores for the W-MAE and W-MSE on these thresholds (see Table 3), while scores between the 95th–99.9th percentile thresholds remain on-par with the SERA (5)model. It is also interesting to note that, while scores do degrade with increasing scale on all thresholds except the 50th percentile , as in Fig. ??, they degrade substantially less than in Fig. ??.~~

~~Intensity-scale diagram of SEDI scores obtained by the ensemble model consisting of the W-MAE, W-MSE and SERA-trained ConvLSTM networks.~~

4 ~~Discussion~~

~~The results presented in this paper indicate that the multi-layered convolutional long short-term memory (ConvLSTM) network can be adapted well to the task of spatio-temporal forecasting of extreme wind events by training the network with imbalanced regression loss. From Table 3, it is clear that the utilisation of the W-MAE~~each model are plotted over lead-time (in hours) for the 75$^{\text{th}}$ (p_{75}), 90$^{\text{th}}$ (p_{90}), ~~W-MSE or SERA loss results in substantially improved forecasts of extreme events of intensity thresholds between the 75th and 99th percentiles, as measured with the symmetric extremal dependency index (SEDI), as compared with either the standard mean squared error (MSE) or mean absolute error (MAE).~~

~~While superior SEDI scores were obtained by the SERA-trained model on intensity thresholds between the 90th–99th percentiles, we have shown that this is in large part due to a severely increased frequency bias, as well as increased coarseness, for extreme events in this range. As shown in Table ??, this bias can be greatly mitigated by merging the SERA (5) model with the W-MAE~~ 95th (~~5~~p_{95}) and ~~W-MSE~~ 99th (~~5)models into a joined ensemble. The ensemble limits frequency bias, limiting the tendency of the SERA (5) model to overshoot and forecast false alarms; it, furthermore, improves SEDI scores between the 50th–90th percentiles while remaining on-par with the SERA (5) model between the 95th–99.9th percentiles; overall RMSE is reduced (Table ??) and forecasts are spatially sharper (Fig. 9 and 10). It must be noted, however, that some of the high intensity hits made~~ p_{99}) percentile intensity thresholds in particular. Once again the persistence forecast is included for reference. Not only does the figure provide a temporal picture of forecasting performance but the different models can readily be compared to the 'baseline' persistence forecast (black-dotted line). The figure clearly shows that the imbalanced regression losses result in sustained hit rates H over lead-time, while false alarm rates FAR and frequency bias B suffer large increases, as compared with the standard MAE and MSE. Indeed, none of the imbalanced regression loss functions succeed in increasing either H or TS without inflating FAR or B to some degree; in fact, typically the stark improvements in H result in degraded TS scores (most clearly visible for the SERA models on thresholds p_{75}, p_{90} and p_{95}).

Although the inverse weighting, the linear weighting and the SERA loss each provide a different way to balance forecasting performance towards improved hit rates, they achieve this aim with varying success. For example, not only are the heavily inflated B and FAR scores produced by the SERA ~~(5) model are inevitably lost in the averaging process of the ensemble forecasts.~~

~~Although the inversely weighted~~ models not typically qualities of reliable forecasting systems, the models also do not succeed in keeping TS scores on-par with the standard MAE and MSE loss~~showed themselves to be less capable than~~ ; spectacular improvements in H thus seem to be primarily a result of extreme overcasting bias, not actually improved predictive power. Compared with the SERA loss~~in shifting performance towards the extreme thresholds, we do note that other weighting methods may yield better results. Shi et al. (2017), for example, utilised a linear weighting method for precipitation nowcasting using a trajectory gated recurrent unit (TrajGRU)network and reported improved performance at higher rain-rate thresholds~~ , the inversely weighted losses, W-MAE$_{inv}$ and W-MSE$_{inv}$ sustain H scores over lead-time to a lesser degree, but nevertheless show strong improvements over the standard MAE and MSE while showing no apparent loss in TS over lead-time (in fact, showing substantial improvement for threshold p_{99}), and inflating FAR and B scores much more conservatively. Although the linear weighted losses display more conservative FAR and B scores still, it is evident that any improvements in H cease rather quickly beyond a threshold of p_{95}.

Lastly, the RMSE scores show clearly how forecasts gradually lose precision with increasing lead-time. It is interesting to note that the relationship appears to be roughly linear. The imbalanced regression losses consistently show improved RMSE scores over lead-time as compared with the standard ~~MSE and MAE (although this conclusion is based on the measuring of performance using the critical success index (CSI) and the Heidke skill score (HSS), neither of which is recommended in the literature for the forecast verification of rare events due to score degeneracy (Stephenson et al., 2008))~~MAE, MSE and persistence, with lowest scores on the higher percentiles between p_{95}–$p_{99.9}$ achieved by the SERA losses, followed by the inversely weighted losses and lastly the linearly weighted losses.

3.1 Forecast distributions

Figure 7 provides a set of bar charts showing the forecast distributions of the ConvLSTM trained with the various different loss functions. The figure is split up into the inversely weighted losses (top-left), the linearly weighted losses (top-right), the SERA loss with different primary control-points (bottom-left) and the standard MAE and MSE losses (bottom-right). Included in the bar charts is the underlying distribution of the observations in the test set, labelled as 'Target' (black dotted line). The distributions were sampled with a step-size of 0.5 (standardised wind speed).

~~We proceed to make a note on the spatial verification that was conducted using the minimum coverage method. Figure ?? suggests that upscaling the forecasts of the SERA (5) model offers no improvement to the forecasting of extreme events beyond the 75th percentile threshold. For non-extreme events of the 50th percentile threshold, forecasts improve at coarser scales as it becomes easier for the model to make correct predictions, which is the expected behaviour. This behaviour ceases, however, for events beyond the 75th percentile threshold. We suspect that this is due to the fact that when a spatio-temporal region of extreme events is missed by the model, it is typically missed completely and hence upscaling the forecast will do nothing to improve it. It may, furthermore, be explained by the fact that when forecasts of extreme events are made by the SERA (5) model, they are typically substantially coarser than the observation, and often forecasted when there are, in fact, no extremes present in the observation. We can see this well in~~ The figure clearly shows how the different types of loss functions result in the forecast distribution being shifted towards the right tail in rather distinct fashions. Comparison of the ~~forecast visualisations provided in Fig. 9 and Fig. 10, or, additionally, in the supplementary material. Since we are using the minimum coverage criterion to spatially upscale the forecasts, a group of false alarms can easily result in an entire upscaled region being labelled as a false alarm, increasing the false alarm rate F and thus degrading the SEDI skill score (see Eq. ??).~~ different forecast distributions with the target distribution highlights the different undercasting and overcasting behaviour as observed from the frequency bias (B) in Table 3. While all imbalanced regression loss functions appear to shift more predictions towards the right tail of the target distribution, they evidently conserve the shape of the target distribution to varying degrees, with the SERA loss and the linear weighting resulting in rather large distortions and heavy overcasting on the right tail. In fact, the SERA loss shifts predictions towards the right tail of the target distribution with such severity that this results in an additional peak on the right side of the forecast distribution, the peak evidently shifted further towards the right tail as the primary control-point varies from p_{50} to p_{90}.

~~On another note, we wish to provide some insight~~

3.2 Permutation tests

In this section, some more insight is given into the predictions made by the ~~multi-layered~~ ConvLSTM network by discussing feature importance. In order to determine the importance of each of the 12 input frames that are used by the ConvLSTM to make its prediction, ~~we proceeded to carry out~~ a permutation test was carried out on the input data. For each input frame at time

T (-11–0), ~~we randomly shuffle~~ all input frames from the testset were randomly shuffled (full fields) at time T~~(~~, essentially nullifying the information flow from this input frame~~), gather~~. Then the model predictions from these permuted inputs ~~and compute~~ were obtained and a skill score S (in %) was computed between the RMSE of the original prediction and target ($\mathrm{RMSE}_{\mathrm{org}}$) and the permuted prediction and target ($\mathrm{RMSE}_{\mathrm{perm}}$) i.e. $S = (1 - \mathrm{RMSE}_{\mathrm{org}}/\mathrm{RMSE}_{\mathrm{perm}}) * 100$. A score of 0 % indicates no change in RMSE, a score of 100 % indicates maximum increase in RMSE ~~due to the permuted inputs~~ and negative scores indicate decrease in RMSE due to the permuted inputs. Not only does this offer insight into the importance that each input frame carries in the ultimate prediction but it also helps to ensure that the model is, in fact, basing its predictions on the information flow between consecutive input frames rather than simply resorting to forecasting climatology. ~~Figure ??~~

Figure 8 presents the RMSE skill scores ~~obtained by the ensemble of the W-MAE, W-MSE and SERA-trained models~~ as aggregated over the testset, obtained by the different models. The figure shows that ~~root mean squared errors tend to get increasingly larger with~~ scores for all models get particularly large as the permuted input frame T ~~approaching~~ approaches 0 hours. This shows clearly that the last input frame at time $T = 0$ hours bears most importance to the predictions, which is to be expected from a regression model predicting the continuation of a sequence from time frame $T = 0$ onward. The standard MAE and MSE loss show a fairly steady rise in RMSE skill score from time $T = -11$ towards $T = 0$, showing that more 'recent' frames of the input tend to bear more importance in the predictions (with the exception of a slight drop (for the MAE loss) or stagnation (for the MSE loss) at $T = -2$ and $T = -1$). The imbalanced regression losses, however, show an additional jump in RMSE skill scores peaking around ca. $T = -8$ and $T = -7$, where-after scores fall considerably before gradually climbing again to peak at $T = 0$. Interestingly, with these earlier frames in the inputs bearing more importance on the predictions, it appears that the models trained with the various imbalanced regression losses have learned to utilise more of the long-term information flow in the inputs to improve the forecasting on the extremes.

3.3 Forecast examples

Finally, Fig. 9, Fig. 10 and Fig. 11 present visualisations of three selected, example forecasts made by the different ConvLSTM models investigated in this paper, which serve to highlight their respective strengths and weaknesses. In each figure, the first row from the top displays the 12 input frames, the second row displays the succeeding 12 target frames and the following rows display the 12 predicted frames of the various models. T refers to the index of the frame (in hours), with ~~an intermediate jump between -9 and -6 hours and a large jump from -3 to 0 hours, where a score of approx. 60% is reached. This shows clearly that the predictions of the ensemble model are heavily dependant on the input frames, with the the most recent frames carrying most importance to the final prediction, but also, interestingly, the frames -9 to -6 hours, perhaps utilised by the ConvLSTM for the more long term dynamics . Most important to the final prediction is clearlythe last input frame at time~~ $T = 0$ ~~hours, which comes as no surprise since the model is a regression model tasked to predict the continuation of spatio-temporal sequence from time frame $T = 0$~~ denoting the last input frame and $T = +12$ denoting the final target and prediction frames. Rather than showing the raw predictions, the predictions are categorised into binary events using local percentile intensity thresholds. In this fashion, the figures show precisely where the different types of events were predicted and where not.

All three examples show a target observation of an intensification of extreme winds, each resulting in a patch of 99th percentile events between from ca. $T = +8$ onward. In each case, the standard MAE and MSE loss either forecast the intensification to some degree but largely fail to capture the 99th percentile events (Fig. 9 and Fig. 11) or they fail to forecast the event completely (Fig. 10). In comparison, inversely weighted losses (W-MAE$_{inv}$ and W-MSE$_{inv}$) show a much improved ability to forecast the right intensification and the right degree of extreme events, with the W-MAE$_{inv}$ performing clearly better in Fig. 11 than the W-MSE$_{inv}$. From the forecasts of the linear weighted losses, the heavy frequency bias on lower percentile events (seen in Table 3, Fig. 6 and Fig. 7) can be easily distinguished, although some 95th and 99th percentile events are captured. Between the SERA$_{p90}$, SERA$_{p75}$ and SERA$_{p50}$ models, the examples clearly reflect the heavily inflated frequency bias towards the higher percentiles, with bias increasing more towards the 99th percentile as the primary control-point varies from p_{50} to p_{90} (in line with the behaviour discussed in Fig. 7).

~~We finish by briefly discussing~~

4 Discussion

The results presented in this paper indicate that the multi-layered convolutional long short-term memory (ConvLSTM) network can be adapted to the task of spatio-temporal forecasting of extreme wind events through the manipulation of the loss function. By analysing the forecasts of the ConvLSTM trained with the various imbalanced regression loss functions investigated in this work, utilising various different scores and intensity thresholds, as well as comparing forecast distributions and visualised forecast examples, it is clear that inverse weighting, linear weighting and squared error-relevance area (SERA) loss each provide viable ways of shifting predictive performance of the ConvLSTM towards the tail of the target distribution. Furthermore, from the permutation tests it is clear that all ConvLSTM models utilise the information flow from the inputs to compute the forecasts and, interestingly, it appears that networks trained with imbalanced regression loss may be utilising more information flow from long-term dynamics than the baseline models trained with standard MAE and MSE loss.

The results indicate that hit rates and RMSE scores can be greatly improved for extreme events up until the 99th percentile threshold, where-after hit rates drop considerably and cease to surpass persistence scores. Table 3 and Fig. 6 demonstrate clearly, however, that improvements in hit rate are accompanied by proportionate increases in frequency bias and false alarm ratios. When this trade-off is particularly extreme, as in the case of the SERA loss with the here-investigated control-points, not only do threat scores begin to suffer considerably but the model also loses its viability as a reliable forecasting model with false alarm ratios massively inflated. Lowering the primary control-point, however, from the 90th percentile (p_{90}) to the 50th percentile (p_{50}) limits this behaviour for extreme events between the 90–99th percentiles (see Fig. 6).

The linear weighting method, instead, shows minimal improvement in hit rate over the standard MAE and MSE on intensity thresholds above p_{90}, as it increases forecasting bias mostly closer to the median and not the tails (see Fig. 7), which means that it does not appear to put enough relative weight on the extreme tails. It should be noted, however, that the linear weighing method tested here was tested only with one slope and other slopes may yield better results. Shi et al. (2017), for example, utilised a linear weighting method for precipitation nowcasting using a trajectory gated recurrent unit (TrajGRU) network and reported improved performance at higher rain-rate thresholds as compared with the standard MSE and MAE (based on the threat score and the Heidke skill score).

Between the three types of imbalanced regression loss investigated in this work, the inverse weighting method appears to strike the best balance between improved hit rate versus increased frequency bias and false alarm ratio. Not only does the inverse weighting method sample the target distribution more accurately (see Fig. 7), but frequency bias and false alarm rates are substantially less inflated than the SERA losses over all percentile thresholds between p_{75}–p_{99}, while hit rates nevertheless greatly improve over the standard MAE and MSE and threat scores remain on-par (see Fig. 6). The W-MAE$_{inv}$ appears to strike this balance slightly better than the W-MSE$_{inv}$.l

This discussion will finish by mentioning a number of possible extensions of this work. One disadvantage of utilising the entirety of available data in ~~the context of~~ this work is that many of the input-target samples containing extreme winds are samples where extreme winds are present in both the input as well as 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 more interesting problem ~~(e.g.~~ for early-warning systems~~)~~. Improving our model as an early-warning system of onsets of extreme winds may thus be obtained by focusing model-learning on precisely such training samples, rather than employing all available samples. To this end, it could be worthwhile to change the model into a nowcasting model~~. This~~, which would entail reducing the lead-time of the model to below 6 hours while increasing the temporal and spatial resolutions of the data, possibly 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 is already being pursued in the community (see: e.g. Racah et al., 2017; Marndi et al., 2020; 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 "~~, as Cadenas et al. (2016) has found these to be significantly more important than other ~~atmospheric variables " - a result based on the work done by Cadenas et al. (2016)~~atmospheric variables to the task of wind forecasting. Xie et al. (2021) use these same three variables, as well as the 1-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. 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. Rädler et al., 2015; Tsonevsky et al., 2018; Chavas and Dawson II, 2021). Another possible extension would be to implement categorical scores directly into the loss function (see e.g. Lagerquist and Ebert-Uphoff, 2022) or even combine the ConvLSTM with a so-called physics-aware loss function (see e.g. Schweri et al., 2021; Cuomo et al., 2022).

Finally, ~~we note~~ it should be noted that while the ConvLSTM has proven itself to be highly effective at modelling complex spatio-temporal patterns, other models have since been proposed as promising improvements to the ConvLSTM for the task of spatio-temporal sequence forecasting. Most notably, the PredRNN and its ~~predecessor~~ successor PredRNN++, proposed by Wang et al. (2017) and Wang et al. (2018), respectively, have been demonstrated to be superior to the ConvLSTM for the task of video frame prediction by maintaining a global memory state rather than constraining memory states to each

ConvLSTM module individually. Other alternative approaches include the usage of functional neural networks (FNNs) (see: Rao et al., 2020) or generative adversarial networks (GANs) (see: Gao et al., 2020). Such models may ~~well~~ be of interest to the meteorological community pursuing data-driven, spatio-temporal forecasting.

5 Conclusions

In this paper ~~we explored~~ a deep learning approach to the task of spatio-temporal prediction of wind speed extremes ~~in the short-to-medium range~~was explored and, in particular, the role of the loss function was investigated. To this end, ~~we investigated the application of~~ a multi-layered convolutional long short-term memory (ConvLSTM) network ~~, which we adapted to imbalanced~~ was adapted to the task of spatio-temporal imbalanced regression by training the model with ~~either inversely weighted mean absolute error (W-MAE), inversely weighted mean squared error (W-MSE) or squared~~ a number of different imbalanced regression loss functions proposed in the literature: Inversely weighted loss, linearly weighted loss and squared error-relevance area (SERA) loss. The models were trained and tested on reanalysis wind speed data from the European Centre for Medium-Range Weather Forecasts (ECMWF) at 1000 ~~hP~~hPa, providing multi-frame forecasts of horizontal near-surface wind speed over Europe with a 12 hour lead-time and in one hour intervals, using the preceding 12 hours as input. By standardising the data based on the local wind speed distributions at each coordinate ~~we focused~~ the definition of an extreme event was focused on its relative rarity rather than its absolute severity~~and considered extreme winds .~~, with extreme winds thus considered in terms of their local distributional percentile.

The model forecasts were ~~verified by computing the symmetric extremal dependence index (SEDI)~~ analysed and compared with a variety of scores, over various lead-times ~~, spatial scales~~ and intensity thresholds. After determining the optimal number of network layers for ~~each of the models (trained with either W-MAE, W-MSE, SERA or standard MAE and MSE loss), a comparison~~ the ConvLSTM trained with the various different loss functions, an extensive comparison was made between the different loss functions ~~was made in Table 3~~and two baseline models trained with either mean absolute (MAE) or mean squared (MSE) loss. The results show that the imbalanced regression loss functions investigated in this paper ~~(W-MAE, W-MSE and SERA loss)~~ can be used ~~effectively to improve forecasting performance~~ to substantially improve hit rates and RMSE scores over the baseline models, however, at the cost of increased frequency bias and false alarm ratios. The SERA loss provides an extreme case of this behaviour, typically at the additional cost of reductions in threat score, although results are heavily dependent on the loss function's so-called control-points. The linear weighting method shows some ability to boost hit rates while keeping frequency bias and false alarm ratio comparatively low, although the utility of the method is lost for extreme events beyond the ~~75th percentile threshold. While the results indicate superior performance of the SERA loss over the W-MAE and W-MSE loss in forecasting extreme wind events of intensity thresholds between to the 90–99th percentiles, we observed that this goes hand-in-hand with a severe~~ 90th percentile intensity threshold, with predictions heavily biased towards the median of the distributions rather than the right tail. Inverse weighting is concluded to strike the best trade-off between improved hit rates and sustained threat scores versus increased frequency bias and ~~an increased coarseness of~~ false alarm ratio, across various thresholds of extreme events up until the 99th percentile intensity threshold - with the weighted MAE loss scoring slighly better than the weighted MSE loss. The inverse weighting method, furthermore, results in a better sampling of the target distribution as compared with the linear weighting or the SERA loss. Out of the different imbalanced regression loss functions investigated in this work, the inverse weighting loss is thus concluded to be most effective at adapting the ConvLSTM to the ~~forecasts. While the SERA loss thus tends to produce worst-case scenarios, we observe greatly improved results when combining the W-MAE, W-MSE and SERA-trained models into an ensemble. Table **??** and Fig. **??** show this quantitatively, while the forecast visualisations in Fig. 9 and 10 show qualitatively that the ensemble is able to model the complex spatio-temporal dynamics of both extreme and non-extreme wind speeds very effectively as far as 12 hours into the future. We conclude that the inversely weighted loss and the squared error-relevance area loss provide relatively easy and effective ways 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~~, and may be best utilised as an ensemble. With this work we hope~~ . With these results, this work is hoped to provide a valuable contribution to the area of deep learning for ~~wind energy applications as well as the area of imbalanced~~ spatio-temporal imbalanced regression and its ~~verification as a forecasting problem~~application to wind energy forecasting research.

Code and data availability. The current version of model is available at the project repository on Github: https://github.com/dscheepens/Deep-RNN-for-extreme-wind-speed-prediction under the MIT license. The exact version of the model used to produce the results used in this paper is archived on Zenodo (DOI: 10.5281/zenodo.7369015), as are scripts to run the model and produce the plots for all the simulations presented in this paper. The data used in this paper can be downloaded from the Copernicus Climate Change Service Climate Data Store (CDS) of the ECMWF (see Hersbach et al., 2018), where the reanalysis data of the U and V components of the horizontal wind velocity were taken at 1000 hPa from the *ERA5 hourly data on pressure levels from 1979 to present* dataset between years 1979-2021 (42 years) and between 40-56° N and 3-19° E. Scalar wind speed was obtained by computing the square root of the sum of the squares of the two wind velocity components. Scripts to generate the data as such are available in the project repository.

Sample availability. Sample forecasts are available at https://github.com/dscheepens/Deep-RNN-for-extreme-wind-speed-prediction/example_forecasts.

6 ~~Figures~~

~~Results from the permutation test as carried out for the ensemble model consisting of the W-MAE, W-MSE and SERA-trained ConvLSTM networks. The figure shows the RMSE skill score (in %) between the targets and the normal predictions of the ensemble, and the targets and the predictions resulting from randomly permuting the inputs at time-frame T. A score of 0% indicates no change in RMSE, a score of 100% indicates maximum increase in RMSE due to the permuted inputs and negative scores indicate decrease in RMSE due to the permuted inputs.~~

Author contributions. DRS and KHS conceptualised the research. DRS carried out the data curation, formal analysis, investigation, methodology, programming, validation, visualisation and writing of the paper. KHS and IS provided supervision, scientific discussion and guidance, and reviewed and revised the work. IS and CP carried out project administration and CP also provided computing resources.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. ~~Our special thanks go to Markus Dabernig and Aitor Atencia (Zentralanstalt für Meteorologie und Geodynamik, Vienna) who provided invaluable and extensive feedback to the paper at its final stages.~~ This work was funded within the Austrian Climate ~~and Research Programme~~ Research Program (ACRP) under the project name MEDEA (KR19AC0K17614) to further research on renewable energy and meteorologically induced extreme events. Special thanks goes to Markus Dabernig and Aitor Atencia (Zentralanstalt für Meteorologie und Geodynamik, Vienna) who provided invaluable and extensive feedback to the paper at its final stages.

References

Alessandrini, S., Sperati, S., and Pinson, P.: A comparison between the ECMWF and COSMO Ensemble Prediction Systems applied to short-term wind power forecasting on real data, Applied Energy, 107, 271–280, https://doi.org/10.1016/j.apenergy.2013.0, 2013.

Alessandrini, S., Sperati, S., and Monache, L. D.: Improving the Analog Ensemble Wind Speed Forecasts for Rare Events, Monthly Weather Review, 147, 2677 – 2692, https://doi.org/10.1175/MWR-D-19-0006.1, 2019.

Amato, F., Guignard, F., Robert, S., and Kanevski, M.: A novel framework for spatio-temporal prediction of environmental data using deep learning, Scientific Reports, 10, 22 243, https://doi.org/10.1038/s41598-020-79148-7, 2020.

Ashkboos, S., Huang, L., Dryden, N., Ben-Nun, T., Dueben, P., Gianinazzi, L., Kummer, L., and Hoefler, T.: ENS-10: A Dataset For Post-Processing Ensemble Weather Forecast, https://doi.org/10.48550/ARXIV.2206.14786, 2022.

Batista, G., Prati, R., and Monard, M.-C.: A Study of the Behavior of Several Methods for Balancing machine Learning Training Data, SIGKDD Explorations, 6, 20–29, https://doi.org/10.1145/1007730.1007735, 2004.

Burton, T., Sharpe, D., Jenkins, N., and Bossanyi, E.: Reviewed Work: 'Wind Energy Handbook', Wind Engineering, 25, 197–199, http://www.jstor.org/stable/43749820, 2001.

Cadenas, E., Rivera, W., Campos-Amezcua, R., and Heard, C.: Wind Speed Prediction Using a Univariate ARIMA Model and a Multivariate NARX Model, Energies, 9, https://doi.org/10.3390/en9020109, 2016.

Casati, B., Ross, G., and Stephenson, D. B.: A new intensity-scale approach for the verification of spatial precipitation forecasts, Meteorological Applications, 11, 141–154, https://doi.org/https://doi.org/10.1017/S1350482704001239, 2004.

Chavas, D. R. and Dawson II, D. T.: An Idealized Physical Model for the Severe Convective Storm Environmental Sounding, Journal of the Atmospheric Sciences, 78, 653 – 670, https://doi.org/10.1175/JAS-D-20-0120.1, 2021.

Chen, K. and Yu, J.: Short-term wind speed prediction using an unscented Kalman filter based state-space support vector regression approach, Applied Energy, 113, 690–705, https://doi.org/10.1016/j.apenergy.2013.0, 2014.

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

Costa, A., Crespo, A., Navarro, J., Lizcano, G., Madsen, H., and Feitosa, E.: A review on the young history of the wind power short-term prediction, Renewable and Sustainable Energy Reviews, 12, 1725–1744, https://doi.org/https://doi.org/10.1016/j.rser.2007.01.015, 2008.

Cuomo, S., di Cola, V. S., Giampaolo, F., Rozza, G., Raissi, M., and Piccialli, F.: Scientific Machine Learning through Physics-Informed Neural Networks: Where we are and What's next, https://doi.org/10.48550/ARXIV.2201.05624, 2022.

Cutululis, N., Litong-Palima, M., and Sørensen, P.: Offshore Wind Power Production in Critical Weather Conditions, in: Proceedings of EWEA 2012 - European Wind Energy Conference & Exhibition, European Wind Energy Association (EWEA), http://events.ewea.org/annual2012/, eWEC 2012 - European Wind Energy Conference &; Exhibition, EWEC 2012 ; Conference date: 16-04-2012 Through 19-04-2012, 2012.

Dabernig, M., Mayr, G. J., Messner, J. W., and Zeileis, A.: Spatial ensemble post-processing with standardized anomalies, Quarterly Journal of the Royal Meteorological Society, 143, 909–916, https://doi.org/https://doi.org/10.1002/qj.2975, 2017.

Damrath, U.: Verification against precipitation observations of a high density network - what did we learn?, in: International Verification Methods Workshop, 2004.

Darwish, A. S. and Al-Dabbagh, R.: Wind energy state of the art: present and future technology advancements, Renew. Energy Environ. Sustain., 5, 7, https://doi.org/10.1051/rees/2020003, 2020.

Deng, X., Li, W., Liu, X., Guo, Q., and Newsam, S.: One-class remote sensing classification: one-class vs. binary classifiers, International Journal of Remote Sensing, 39, 1890–1910, https://doi.org/10.1080/01431161.2017.1416697, 2018.

Deppe, A. J., Gallus, W. A., and Takle, E. S.: A WRF Ensemble for Improved Wind Speed Forecasts at Turbine Height, Weather and Forecasting, 28, 212 – 228, https://doi.org/10.1175/WAF-D-11-00112.1, 2013.

Ding, D., Zhang, M., Pan, X., Yang, M., and He, X.: Modeling Extreme Events in Time Series Prediction, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, p. 1114–1122, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3292500.3330896, 2019.

Ebert, E. E.: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework, Meteorological Applications, 15, 51–64, https://doi.org/https://doi.org/10.1002/met.25, 2008.

Ebert, E. E.: Neighborhood Verification: A Strategy for Rewarding Close Forecasts, Weather and Forecasting, 24, 1498 – 1510, https://doi.org/10.1175/2009WAF2222251.1, 2009.

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

Ferro, C. A. T. and Stephenson, D. B.: Extremal Dependence Indices: Improved Verification Measures for Deterministic Forecasts of Rare Binary Events, Weather and Forecasting, 26, 699 – 713, https://doi.org/10.1175/WAF-D-10-05030.1, 2011.

Friederichs, P., Wahl, S., and Buschow, S.: Chapter 5 - Postprocessing for Extreme Events, in: Statistical Postprocessing of Ensemble Forecasts, edited by Vannitsem, S., Wilks, D. S., and Messner, J. W., pp. 127–154, Elsevier, https://doi.org/https://doi.org/10.1016/B978-0-12-812372-0.00005-4, 2018.

Fyrippis, I., Axaopoulos, P. J., and Panayiotou, G.: Wind energy potential assessment in Naxos Island, Greece, Applied Energy, 87, 577–586, https://ideas.repec.org/a/eee/appene/v87y2010i2p577-586.html, 2010.

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

Ghosh, A., Sufian, A., Sultana, F., Chakrabarti, A., and De, D.: Fundamental Concepts of Convolutional Neural Network, pp. 519–567, Springer, https://doi.org/10.1007/978-3-030-32644-9_36, 2020.

Goyal, S., Raghunathan, A., Jain, M., Simhadri, H. V., and Jain, P.: DROCC: Deep Robust One-Class Classification., CoRR, abs/2002.12718, https://arxiv.org/abs/2002.12718, 2020.

Hassanaly, M., Perry, B. A., Mueller, M. E., and Yellapantula, S.: Uniform-in-Phase-Space Data Selection with Iterative Normalizing Flows, https://doi.org/10.48550/ARXIV.2112.15446, 2021.

Hendrycks, D., Mazeika, M., and Dietterich, T.: Deep Anomaly Detection with Outlier Exposure, in: International Conference on Learning Representations, https://openreview.net/forum?id=HyxCxhRcY7, 2019.

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.: ERA5 hourly data on pressure levels from 1979 to present, https://doi.org/10.24381/cds.bd0915c6, (Accessed on 20-AUG-2021), 2018.

Hogan, R. J. and Mason, I. B.: Deterministic forecasts of binary events, chap. 3, pp. 31–59, John Wiley & Sons, Ltd, 2012.

Huang, C., Li, F., and Jin, Z.: Maximum Power Point Tracking Strategy for Large-Scale Wind Generation Systems Considering Wind Turbine Dynamics, IEEE Transactions on Industrial Electronics, 62, 2530–2539, https://doi.org/10.1109/TIE.2015.2395384, 2015.

IEA: Global Energy Review 2021, https://www.iea.org/reports/global-energy-review-2021, 2021.

Ji, Y., Zhi, X., Ji, L., Zhang, Y., Hao, C., and Peng, T.: Deep-learning-based post-processing for probabilistic precipitation forecasting, Frontiers in Earth Science, 10, https://doi.org/10.3389/feart.2022.978041, 2022.

Jung, J. and Broadwater, R. P.: Current status and future advances for wind speed and power forecasting, Renewable and Sustainable Energy Reviews, 31, 762–777, https://doi.org/10.1016/j.rser.2013.12.05, 2014.

Kavasseri, R. G. and Seetharaman, K.: Day-ahead wind speed forecasting using f-ARIMA models, Renewable Energy, 34, 1388–1393, https://doi.org/10.1016/j.renene.2008.09., 2009.

Keisler, R.: Forecasting Global Weather with Graph Neural Networks, https://doi.org/10.48550/ARXIV.2202.07575, 2022.

Kikuchi, R., Misaka, T., Obayashi, S., Inokuchi, H., Oikawa, H., and Misumi, A.: Nowcasting algorithm for wind fields using ensemble forecasting and aircraft flight data, Meteorological Applications, 25, https://doi.org/10.1002/met.1704, 2017.

Lagerquist, R. and Ebert-Uphoff, I.: Can we integrate spatial verification methods into neural-network loss functions for atmospheric science?, https://doi.org/10.48550/ARXIV.2203.11141, 2022.

Lei, M., Shiyan, L., Chuanwen, J., Hongling, L., and Yan, Z.: A review on the forecasting of wind speed and generated power, Renewable and Sustainable Energy Reviews, 13, 915–920, https://doi.org/https://doi.org/10.1016/j.rser.2008.02.002, 2009.

Leva, S., Dolara, A., Grimaccia, F., Mussetta, M., and Ogliari, E.: Analysis and validation of 24 hours ahead neural network forecasting of photovoltaic output power, Mathematics and Computers in Simulation (MATCOM), 131, 88–100, https://EconPapers.repec.org/RePEc:eee:matcom:v:131:y:2017:i:c:p:88-100, 2017.

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

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

Marndi, A., Patra, G. K., and Gouda, K. C.: Short-term forecasting of wind speed using time division ensemble of hierarchical deep neural networks, Bulletin of Atmospheric Science and Technology, 1, 91–108, https://doi.org/10.1007/s42865-020-00009-2, 2020.

Mohamad, M. A. and Sapsis, T. P.: Sequential sampling strategy for extreme event statistics in nonlinear dynamical systems, Proceedings of the National Academy of Sciences, 115, 11 138–11 143, https://doi.org/10.1073/pnas.1813263115, 2018.

Oh, J., Guo, X., Lee, H., Lewis, R., and Singh, S.: 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, p. 2863–2871, MIT Press, Cambridge, MA, USA, 2015.

Oliveira, M., Moniz, N., Torgo, L., and Santos Costa, V.: Biased resampling strategies for imbalanced spatio-temporal forecasting, International Journal of Data Science and Analytics, 12, 205–228, https://doi.org/10.1007/s41060-021-00256-2, 2021.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Advances in Neural Information Processing Systems 32, pp. 8024–8035, Curran Associates, Inc., http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf, 2019.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., Hassanzadeh, P., Kashinath, K., and Anandkumar, A.: FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators, https://doi.org/10.48550/ARXIV.2202.11214, 2022.

Petrović, V. and Bottasso, C. L.: Wind turbine optimal control during storms, Journal of Physics: Conference Series, 524, 012 052, https://doi.org/10.1088/1742-6596/524/1/012052, 2014.

Phipps, K., Lerch, S., Andersson, M., Mikut, R., Hagenmeyer, V., and Ludwig, N.: Evaluating ensemble post-processing for wind power forecasts, Wind Energy, 25, 1379–1405, https://doi.org/https://doi.org/10.1002/we.2736, 2022.

Racah, E., Beckham, C., Maharaj, T., Kahou, S. E., Prabhat, and Pal, C.: 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, p. 3405–3416, Curran Associates Inc., Red Hook, NY, USA, 2017.

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

Rasp, S. and Thuerey, N.: Data-Driven Medium-Range Weather Prediction With a Resnet Pretrained on Climate Simulations: A New Model for WeatherBench, Journal of Advances in Modeling Earth Systems, 13, e2020MS002 405, https://doi.org/https://doi.org/10.1029/2020MS002405, e2020MS002405 2020MS002405, 2021.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N.: WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting, Journal of Advances in Modeling Earth Systems, 12, e2020MS002 203, https://doi.org/https://doi.org/10.1029/2020MS002203, e2020MS002203 10.1029/2020MS002203, 2020.

Ribeiro, R. and Moniz, N.: Imbalanced regression and extreme value prediction, Machine Learning, 109, 1–33, https://doi.org/10.1007/s10994-020-05900-9, 2020.

Ruder, S.: An overview of gradient descent optimization algorithms, 2017.

Rädler, A., Groenemeijer, P., Pistotnik, G., Sausen, R., and Faust, E.: Identification of favorable environments for thunderstorms in reanalysis data, Meteorologische Zeitschrift, 26, https://doi.org/10.1127/metz/2016/0754, 2015.

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

Scheuerer, M. and Hamill, T. M.: Statistical Postprocessing of Ensemble Precipitation Forecasts by Fitting Censored, Shifted Gamma Distributions, Monthly Weather Review, 143, 4578 – 4596, https://doi.org/10.1175/MWR-D-15-0061.1, 2015.

Schmidl, S., Wenig, P., and Papenbrock, T.: Anomaly Detection in Time Series: A Comprehensive Evaluation, Proceedings of the VLDB Endowment (PVLDB), 15, 1779–1797, https://doi.org/10.14778/3538598.3538602, 2022.

Schweri, L., Foucher, S., Tang, J., Azevedo, V. C., GÃ¼nther, T., and Solenthaler, B.: A Physics-Aware Neural Network Approach for Flow Data Reconstruction From Satellite Observations, Frontiers in Climate, 3, 23, https://doi.org/10.3389/fclim.2021.656505, 2021.

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

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-k., and Woo, W.-c.: 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, p. 802–810, MIT Press, Cambridge, MA, USA, 2015.

Shi, X., Gao, Z., Lausen, L., Wang, H., and Yeung, D.-Y.: 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, p. 5617–5627, MIT Press, Cambridge, MA, USA, 2017.

Srivastava, N., Mansimov, E., and Salakhudinov, R.: Unsupervised Learning of Video Representations using LSTMs, in: Proceedings of the 32nd International Conference on Machine Learning, edited by Bach, F. and Blei, D., vol. 37 of *Proceedings of Machine Learning Research*, pp. 843–852, PMLR, Lille, France, https://proceedings.mlr.press/v37/srivastava15.html, 2015.

Stephenson, D., Casati, B., Ferro, C., and Wilson, C.: The Extreme Dependency Score: A non-vanishing measure for forecasts of rare events, Meteorological Applications, 15, 41 – 50, https://doi.org/10.1002/met.53, 2008.

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

Tsonevsky, I., Doswell, C. A., and Brooks, H. E.: Early Warnings of Severe Convection Using the ECMWF Extreme Forecast Index, Weather and Forecasting, 33, 857 – 871, https://doi.org/10.1175/WAF-D-18-0030.1, 2018.

Vondrick, C., Pirsiavash, H., and Torralba, A.: Generating Videos with Scene Dynamics, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, p. 613–621, Curran Associates Inc., Red Hook, NY, USA, 2016.

Wang, J., Zong, Y., You, S., and Træholt, C.: A review of Danish integrated multi-energy system flexibility options for high wind power penetration, Clean Energy, 1, https://doi.org/10.1093/ce/zkx002, 2017.

Wang, S., Cao, J., and Yu, P.: Deep Learning for Spatio-Temporal Data Mining: A Survey, IEEE Transactions on Knowledge and Data Engineering, pp. 1–1, https://doi.org/10.1109/TKDE.2020.3025580, 2020.

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

Weyn, J. A., Durran, D. R., and Caruana, R.: Improving Data-Driven Global Weather Prediction Using Deep Convolutional Neural Networks on a Cubed Sphere, Journal of Advances in Modeling Earth Systems, 12, e2020MS002 109, https://doi.org/https://doi.org/10.1029/2020MS002109, e2020MS002109 10.1029/2020MS002109, 2020.

Williams, R., Ferro, C., and Kwasniok, F.: A comparison of ensemble post-processing methods for extreme events, Quarterly Journal of the Royal Meteorological Society, 140, https://doi.org/10.1002/qj.2198, 2014.

Wiser, R., Yang, Z., Hand, M., Hohmeyer, O., Infield, D., Jensen, P. H., Nikolaev, V., O'Malley, M., Sinden, G., and Zervos, A.: 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, 2011.

Xie, A., Yang, H., Chen, J., Sheng, L., and Zhang, Q.: A Short-Term Wind Speed Forecasting Model Based on a Multi-Variable Long Short-Term Memory Network, Atmosphere, 12, https://www.mdpi.com/2073-4433/12/5/651, 2021.

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

Yeo, I.-K. and Johnson, R. A.: A New Family of Power Transformations to Improve Normality or Symmetry, Biometrika, 87, 954–959, http://www.jstor.org/stable/2673623, 2000.

Yu, H., Uy, W. I. T., and Dauwels, J.: Modeling Spatial Extremes via Ensemble-of-Trees of Pairwise Copulas, IEEE Transactions on Signal Processing, 65, 571–586, https://doi.org/10.1109/TSP.2016.2614485, 2017.

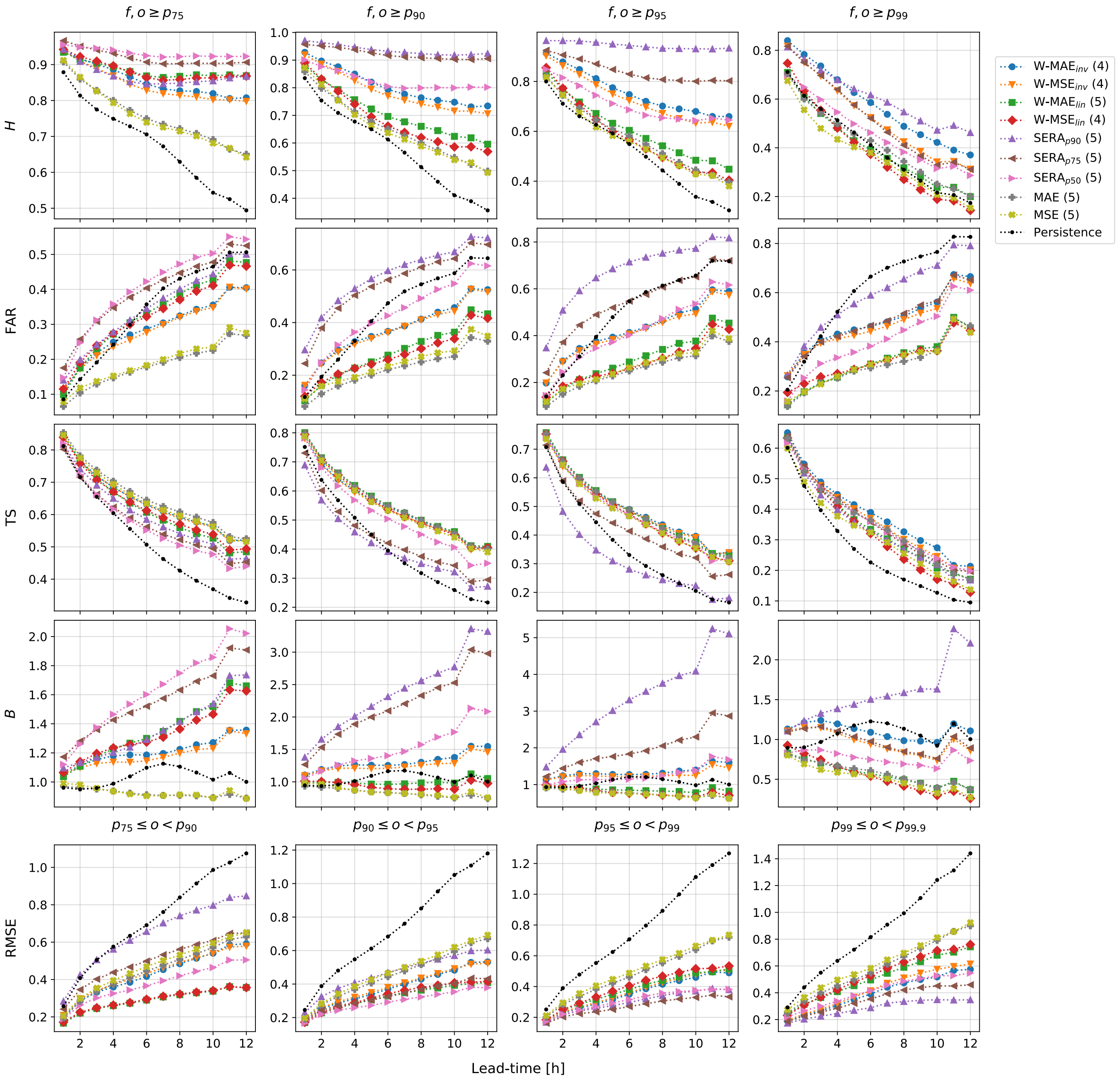


Figure 6. Comparison of the hit score (H), false alarm ratio (FAR), threat score (TS), frequency bias (~~in %~~B) and root mean squared error (RMSE) of the ConvLSTM network trained with ~~either W-MAE, W-MSE, SERA, MAE or MSE~~ the various different loss ~~. Frequency bias is presented for winds~~ functions, plotted over lead-time (~~y~~in hours) ~~exceeding local~~ and various percentile intensity thresholds~~varying between the 50th and 99.9th percentiles~~. The optimal number of network layers used for each loss function is given in brackets after the name of the loss function. ~~Also included in~~ The label 'persistence' refers to the ~~table is the ensemble model, consisting of the W-MAE, W-MSE and SERA-trained networks~~persistence forecast.

	Frequency bias						
	$y \geq p_{50}$	$y \geq p_{75}$	$y \geq p_{90}$	$y \geq p_{95}$	$y \geq p_{99}$	$y \geq p_{99.9}$	RMSE
W-MAE (5)	104.2	113.6	121.1	123.6	101.9	39.8	0.551
W-MSE (5)	111.5	124.2	136.6	140.2	103.6	29.4	0.610

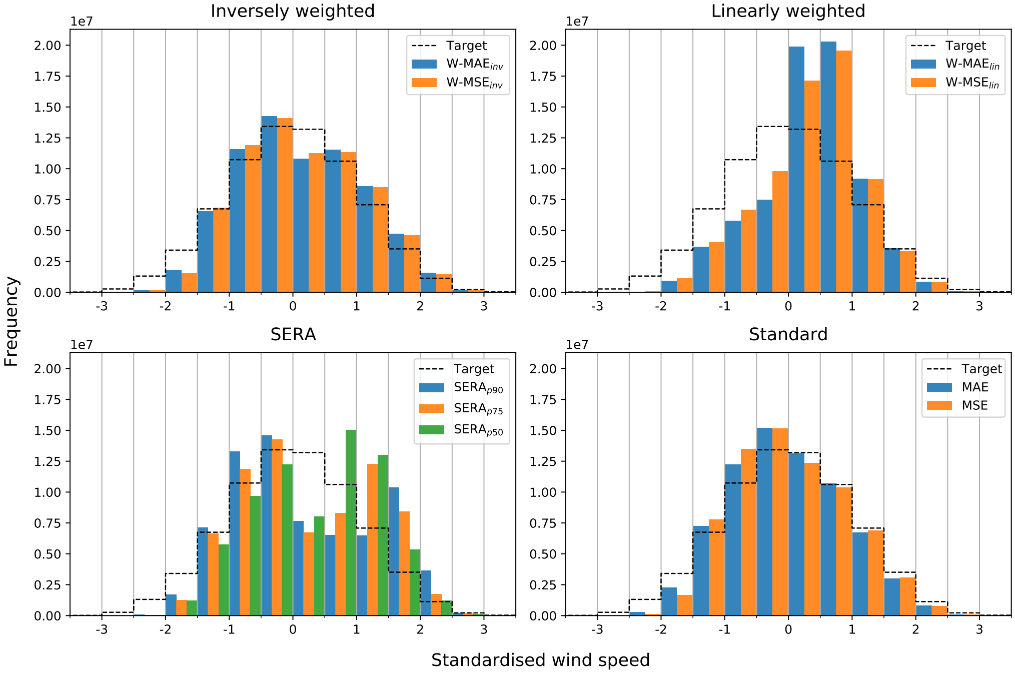


Figure 7. Bar charts of forecast distributions of the ConvLSTM trained with the various different loss functions, compared to the underlying distribution of the observations in the testset, labelled as 'Target' (black dotted line). Top left: The inversely weighted losses. Top right: The linearly weighted losses. Bottom left: The SERA loss with different primary control-points (with the secondary control-point fixed at p_{99}). Bottom right: The standard MAE and MSE losses. The distributions were sampled with a step-size of 0.5 (standardised wind speed).

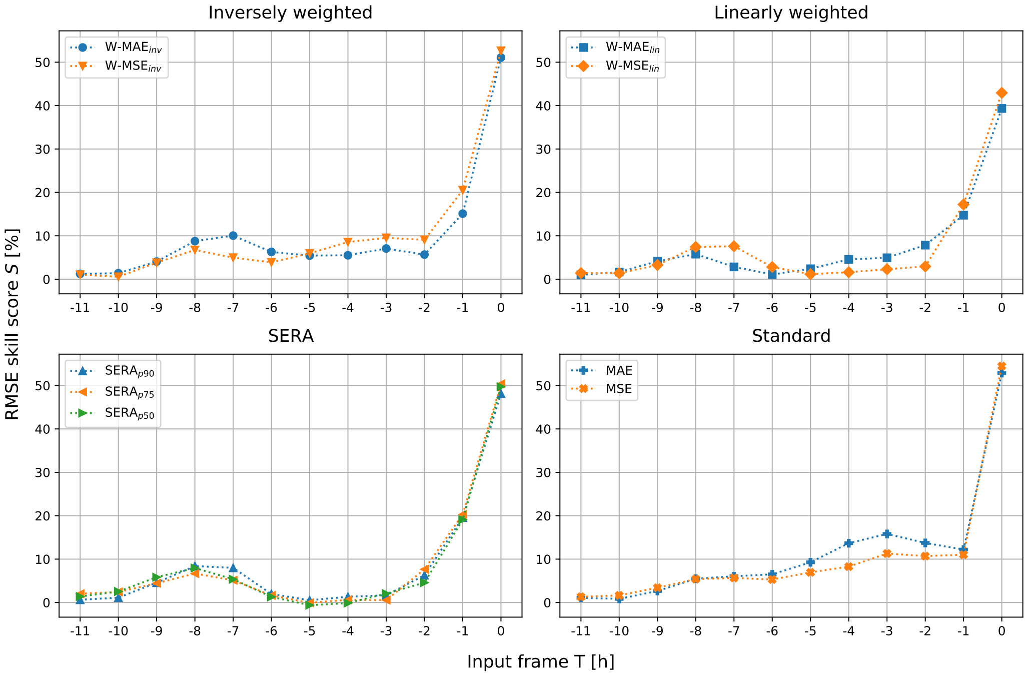


Figure 8. Results from the permutation tests. The figure shows the RMSE skill score (in %) between the targets and the normal predictions of each of the models and the targets and the predictions resulting from randomly permuting the inputs at time-frame T. A score of 0 % indicates no change in RMSE, a score of 100 % indicates maximum increase in RMSE and negative scores indicate decrease in RMSE due to the permuted inputs. Top right: The linearly weighted losses. Bottom left: The SERA loss with different primary control-points (with the secondary control-point fixed at p_{99}). Bottom right: The standard MAE and MSE losses.

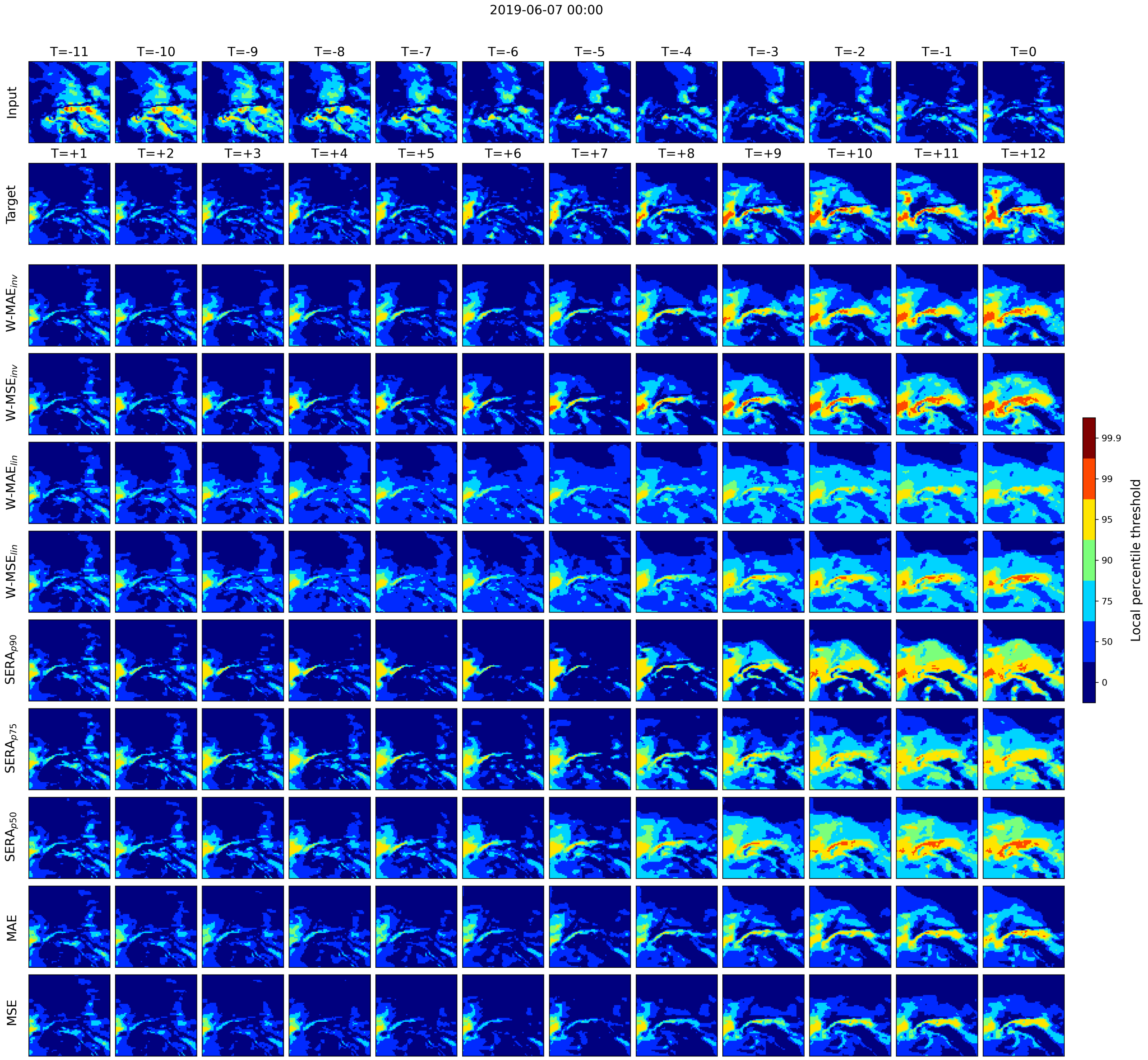


Figure 9. An example forecast from the ConvLSTM network trained with the various different loss functions. The first row from the top displays the 12 input frames, the second row the succeeding 12 target frames and the following rows the 12 predicted frames of the models. T refers to the index of the frame (in hours), with $T = 0$ denoting the last input frame and $T = +12$ denoting the final target and prediction frames. Rather than showing the raw predictions, the predictions are categorised into binary events using percentile intensity thresholds.

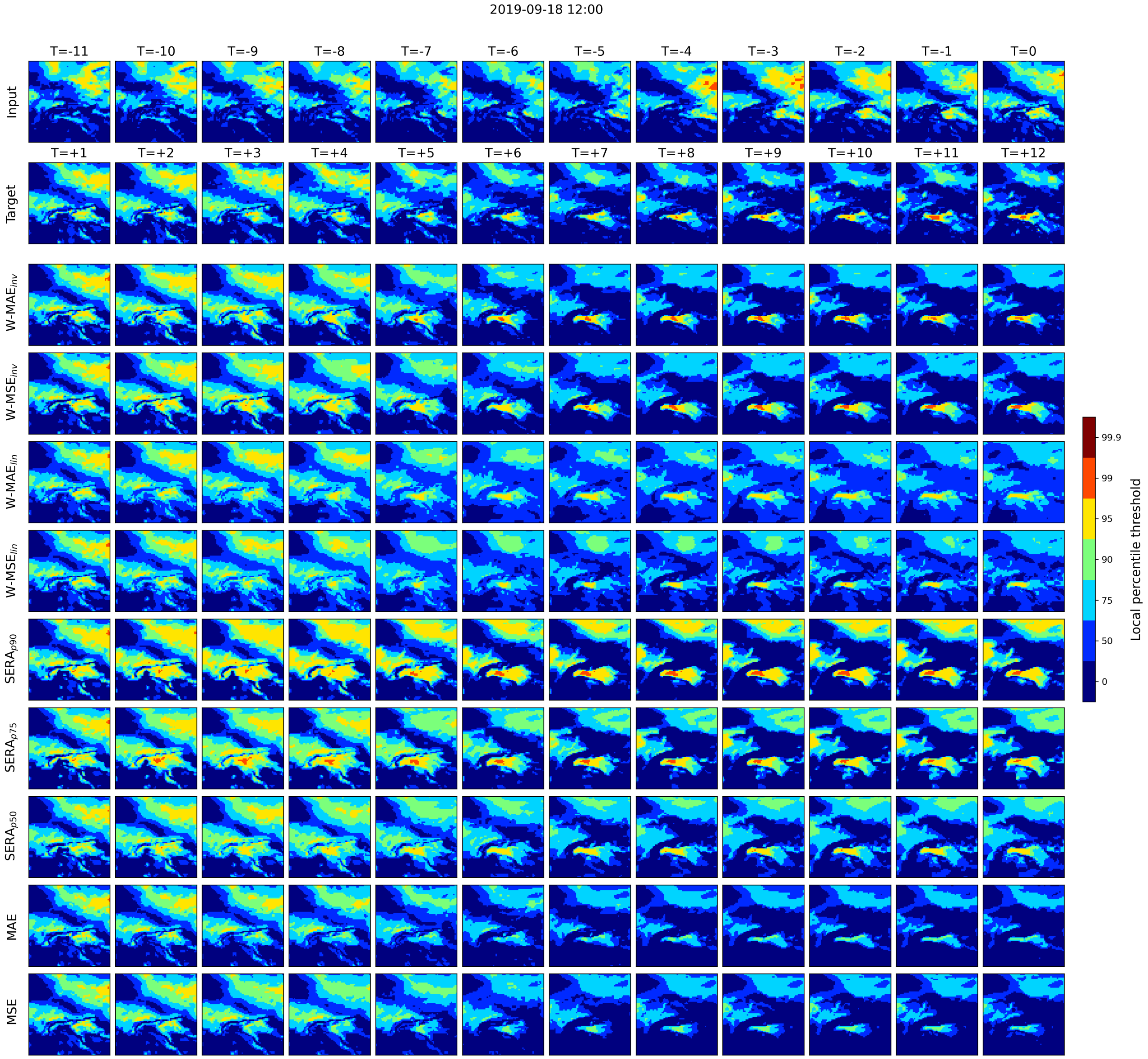


Figure 10. As Fig. 9

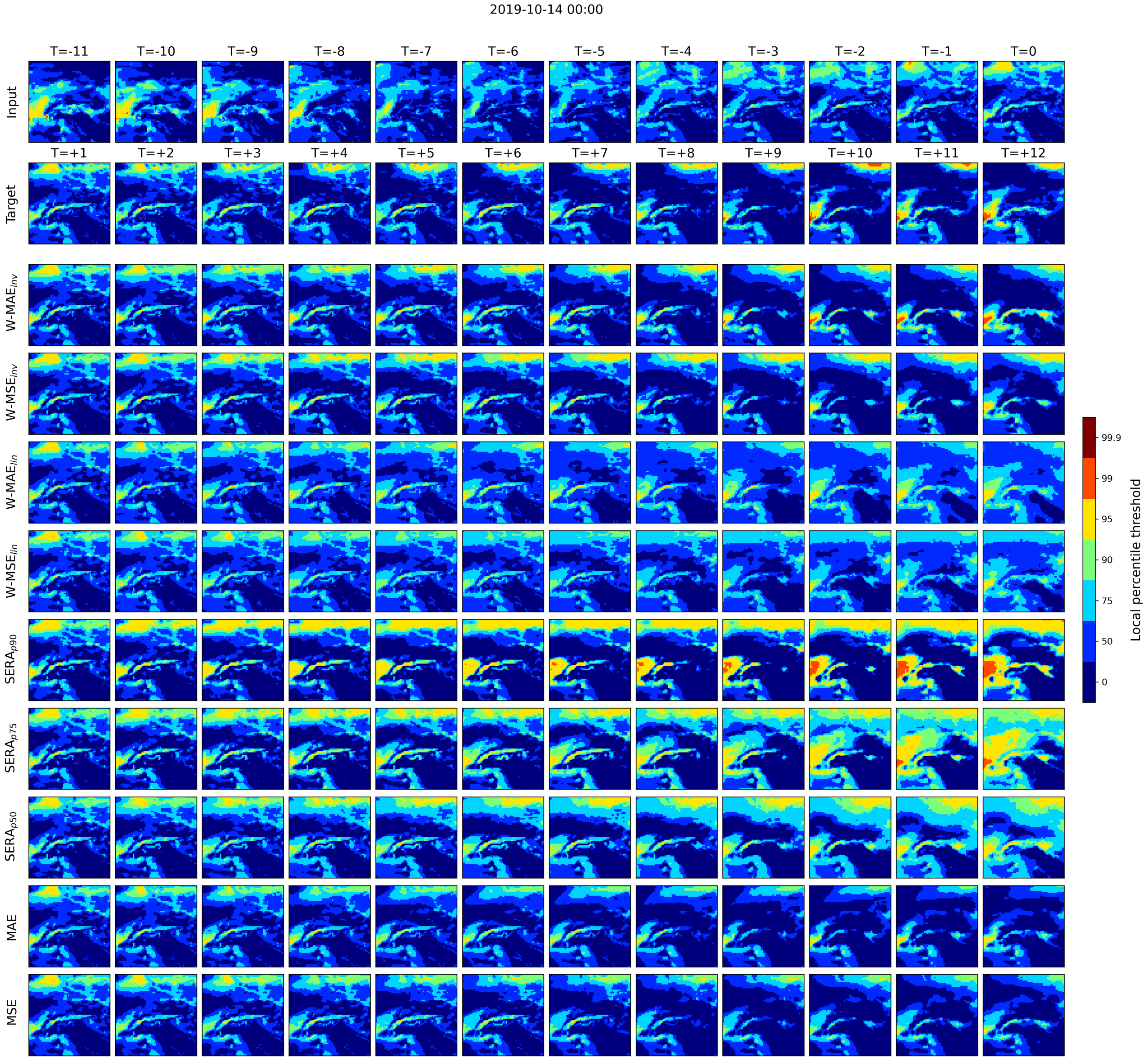


Figure 11. As Fig. 9