Machine learning applications for weather and climate predictions need greater focus on extremes: 2023 update

Peter Watson
University of Bristol, UK
Peter.watson@bristol.ac.uk

Abstract

Multiple studies have now demonstrated that machine learning (ML) can give improved skill for predicting or simulating fairly typical weather events, for tasks such as short-term and seasonal weather forecasting, downscaling simulations to higher resolution and emulating and speeding up expensive model parameterisations. Many of these used ML methods with very high numbers of parameters, such as neural networks, which are the focus of the discussion here. Not much attention has been given to the performance of these methods for extreme event severities of relevance for many critical weather and climate prediction applications, with return periods of more than a few years. This leaves a lot of uncertainty about the usefulness of these methods, particularly for general purpose prediction systems that must perform reliably in extreme situations. ML models may be expected to struggle to predict extremes due to there usually being few samples of such events. However, there are some studies that do indicate that ML models can have reasonable skill for extreme weather, and that it is not hopeless to use them in situations requiring extrapolation. This paper reviews these studies, updating an earlier review, and argues that this is an area that needs researching more. Ways to get a better understanding of how well ML models perform at predicting extreme weather events are discussed.

This paper is based on and updated from Watson (2022) [1]. Key updates are summarized in sec. 2.1.

1 Introduction

It has been shown that machine learning (ML) can perform well at making predictions for events of typical intensities for tasks such as short-term and seasonal weather forecasting e.g. [2]–[6], downscaling simulations to higher resolution [7]–[10] and emulating and speeding up expensive model parameterisations e.g. [11], [12]. However, evaluation of the performance of these methods for extreme events is relatively neglected. This paper discusses the need for improving understanding of how ML methods perform in extreme situations, relevant results from the small number of studies that have evaluated this to date, and approaches that can be used in future work to accelerate progress. This is relevant for addressing climate change as predictions of changes in extreme event severities are required to inform mitigation and adaptation, and identifying ML methods that perform well for this will help to speed up advances.

The lack of evaluation on extreme events leaves a lot of uncertainty about the usefulness of these methods, particularly for general purpose prediction systems that must perform reliably in extreme situations. ML models may be expected to struggle to predict extremes due to there usually being few samples of such events. However, as will be discussed below, there are some studies that do

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2023.

indicate that ML models can have reasonable skill for extreme weather, and that it is not hopeless to use them in situations requiring extrapolation. This makes it an area worth researching more.

Some clarity is needed about the use of the term “extreme”. One useful metric to represent the degree to which an event is extreme is the return period, the average time between events with a magnitude at least as large as for the event in question. A large number of studies use the term “extreme” to describe events around the 90-99th percentile of daily data, which correspond to only a 10-100 day return period. It is indeed useful to assess the performance of ML models around such thresholds. However, these are far from event severities that are relevant to many applications of weather and climate models, and studies typically do not demonstrate how their methods would perform in these cases.

At the high end of the scale, events with return periods in the thousands of years are sometimes studied in extreme event attribution (e.g. [13], [14]) and in the hundreds of years for designing infrastructure for flood and drought resilience (e.g. [15], [16]). In weather forecasting, the Met Office’s most severe “red” weather warning was issued once every few years per event type in the system’s first decade [17]. The return period at individual locations that were most affected by these events will have been substantially higher. Forecast reliability will also need to be assured for even more extreme events. In keeping with these examples, in the rest of this article “extreme” is used to refer to events with return periods of more than a few years.

It seems likely that for ML-based systems to be considered for use in operational weather and climate prediction systems, good performance in extreme situations needs to be shown. This should include events going beyond what is used for training systems, since it cannot be known in advance what range of input data the system will see. Operational systems need to predict events that are more severe than any in the historical record at times. It can be asked is there much value in continuing development of ML-based systems for weather and climate prediction without demonstrating at least satisfactory performance for extremes?

If an approach is taken to try to first design systems to perform well for typical weather and then improve extreme event capabilities later, this could waste a lot of time if useful methods for the former are not the same as for the latter. This is an especially large concern for ML methods with large numbers of parameters (e.g. large neural networks) that require a lot of samples for training. Particular methods may also have their own vulnerabilities. For example, generative adversarial networks are prone to “mode collapse”, where predictions seriously undersample parts of the data distribution, potentially very adversely affecting performance for extremes. Random forests cannot predict values beyond those seen in training data, so they may not be a good choice for applications where skillful prediction for beyond-sample events is important. Therefore evaluating how well such systems actually perform in extreme situations is very important for helping researchers choose the best methods to develop for their applications.

The challenge in making predictions in extreme situations comes not just from these events being rare, but also from how far they can exceed historical records. The 2021 heatwave in the Northwest USA and western Canada beat previous temperature records by 5°C in Portland, standing far above previous values, with an estimated return period in the present climate of ~1000 years [18]. Climate model simulations include events where weekly-average temperature exceeds previous records by over five standard deviations [19]. Rainfall extremes can exceed prior historical values by even greater margins. In 2018 and 2019 in Kerala, India, there were 14-day rainfall totals that exceeded 30 standard deviations, associated with strong convection [20]. Convective rainfall in the USA has led to river discharges reaching over 20 times the 10-year return level on a large number of occasions, with the most extreme recorded discharge due to rainfall being 200 times that level [21]. It therefore wouldn’t be over the top to evaluate robustness of ML-based systems to this degree of extremity for cases where convection is important, and otherwise to perhaps ~5 standard deviation perturbations above the highest values in observed or simulated training data.

2 Previous studies evaluating ML on extreme events

2.1 Recent updates

There have been several notable advances since the publication of [1]:

- Lam et al. (2023) [4] presented a medium-range weather prediction model based on a graph neural network. Amongst the results shown were precision and recall for hot and cold temperatures in the most intense 0.5% of events for given months of the year (in the supplementary information), corresponding to events with approximately at least 7 year return periods. Skill was competitive with that of a leading conventional weather forecast model.
- Vosper et al. (2023) [9] tested two different variants of generative adversarial network (GAN) on the problem of downscaling coarse (1.0°) tropical cyclone rainfall data to high resolution (0.1°). The variants were a Wasserstein GAN (WGAN) and variational autoencoder GAN (VAEGAN). A key part of this study was holding back the 100 storms with the highest coarse resolution rainfall values in the 44 year dataset in a separate "extreme test" dataset, with all storms used in training having lower peak rainfall rates. Both GANs performed well on a test dataset that was drawn from the same population as the training dataset. However, performance on the extreme test dataset was very different: the WGAN performed well, whereas the VAEGAN produced large errors. This indicates that ML methods can in some situations extrapolate to extremes well. But there is also a real risk that a method can perform well within its training envelope yet fail badly when seeing a more intense event, as in the VAEGAN case. This underlines the message of this review.
- Magnusson et al. (2023) [22] evaluated global weather forecasts from the deterministic Pangu-Weather system [5] on two extreme UK events in 2022. The forecast for storm Eunice of mean sea level pressure and 10m wind speed at 48 hour lead time was reasonable. For the July heatwave, in which observed temperatures exceeded 40°C, the forecasts were about 5°C too low, even at 12 hour lead time. In both cases, the ML system was somewhat less accurate than the existing conventional forecast, but it is notable that it achieved predictions nearly as realistic in these extreme cases.
- Addison et al. ("Machine learning emulation of a km-scale UK climate model", in prep.), advancing from [10], trained a diffusion model to predict 8.8km resolution rainfall given variables at 60km resolution from a global climate model. They used about 500 years of high-resolution simulation data for training and the model performed well, even on the days with the highest rainfall in the 108 year test dataset.
- Antonio et al. ("Post-processing East African rainfall forecasts using a generative machine learning model", in prep.) trained a GAN to postprocess short-term weather forecasts of rainfall in East Africa. This included using the extreme 2018 March-May season in Kenya [23] as one of the test datasets. The model succeeded in improving aspects of the forecasts such as the diurnal cycle of rainfall and fractions skill score up to the 99.9th percentile. Performance was similar to that on the primary test dataset, a year with fairly typical weather.

2.2 Complete set of published studies

There are ten published studies in total that I have been able to find in the literature that indicate that ML-based systems can have reasonable skill in extreme situations with return periods of more than a few years. These are the six in [1], the three published studies discussed in sec. 2.1, and reference [24], which was overlooked in [1]. These are summarised in table 1.

Table 1: Summary of studies that found that ML-based systems can perform reasonably at predicting extreme events that have return periods of more than a few years. TDL = training dataset length. MaxRP = maximum return period evaluated.

Study	Summary information	Notes
Adewoyin et al. (2021) [25]	• Convolutional recurrent neural network • TDL: 10 years • MaxRP: ~6 years	• Downscaled daily-mean precipitation at 16 UK locations.

Boulaguiem et al. (2022) [26]	• Generative adversarial network (GAN) • TDL: 50 years • MaxRP: ~2000 years	• Produced samples of maps of annual summer maximum temperature and winter maximum precipitation over Europe. • The density in the tails of the predicted distribution appeared reasonable, though errors were not precisely quantified. • The structure of their GAN was adapted to work better for extremes.
Frame et al. (2022) [27]	• Long short-term memory neural network • TDL: Up to 34 years per river catchment • MaxRP: >100 years	• Simulation of river flow time series in the USA. • In one test they removed events in the training dataset with return periods greater than 5 years and found that prediction scores were still good for events with return periods exceeding 100 years (estimated using a fitted distribution).
Grönquist et al. (2021) [28]	• Convolutional neural network • TDL: 15 years • MaxRP: Unquantified, but record-breaking	• Postprocessed global weather forecasts at 48 hour lead time. • Improved forecast skill scores on extreme events including Hurricane Winston (the most intense southern hemisphere hurricane on record) and an unprecedented cold wave in southeast Asia.
Herman and Schumacher (2018) [24]	• Random forests • TDL: 10 years 8 months • MaxRP: ~10 years	• USA daily precipitation forecasting at lead times 2-3 days. • Achieved probabilistic scores for predicting extreme events comparable to conventional weather prediction models in some regions.
Lam et al. (2023) [4]	• Graph neural network • TDL: 43 years 9 months • MaxRP: ~7 years	• Medium-range weather forecasting, including results for moderate temperature extremes. • Also see sec. 2.1.
Lopez-Gomez et al. (2022) [29]	• Convolutional neural network • TDL: 24 years • MaxRP: ~1000 years	• Global weather forecasts of daily temperature, up to lead times of 4 weeks. • Produced sensible forecasts for record-breaking events: the 2017 European heatwave and the 2021 Northwest USA heatwave. • They used a modified loss function that put greater weight on extreme events.
Magnusson et al. (2023) [22]	• Transformer-based, deterministic • TDL: 39 years • MaxRP: Unquantified, but record-breaking	• Tests of medium-range weather forecasts by Pangu-Weather on two extreme UK weather events. • Also see sec. 2.1.
Nevo et al. (2022) [30]	• Bespoke combination of ML models • TDL: 5 years • MaxRP: ~5 years	• Used a combination of ML models for flood-prediction, evaluated in India and Bangladesh. • Median performance on 5-year return period events, using only less severe events in training, was similar to that for typical events overall, though poor in some cases.
Vosper et al. (2023) [9]	• Wasserstein GAN and variational autoencoder GAN • TDL: 44 years • MaxRP: ~44 years	• Predicting high-resolution tropical cyclone rainfall given low-resolution rainfall. • WGAN performed well at extrapolation, whilst VAEGAN did not. • Also see sec. 2.1.

These results show that there are good prospects that ML-based systems could have skill for extreme events with multi-year return periods and beyond, but there are not enough studies to know whether this is true in most cases. Eight studies evaluated neural network-based models, indicating that neural networks can be successful for this task. Three studies obtained reasonable evaluation results for extreme events with estimated return periods much longer than the training dataset, indicating that generalisation to more extreme events is possible ([26], [27], [29]). Eight studies did not change their model architecture or training procedure to particularly target achieving good performance on extremes, indicating that existing methods are often capable of generalising to extreme events.

Future research recommendations

- Filling particularly important research gaps:
    - simulating weather events with multi-decadal return periods
    - evaluating the performance of stochastic generative models (e.g. GANs, diffusion models) for extremes, which only appears to have been examined so far in two published studies ([9], [26]).
    - examining models' extrapolation behaviour, such as by scaling up anomalies in input variables, as illustrated in [31].
- Routinely showing simple but useful diagnostics such as scatter plots and qq plots, including for the most extreme data values, and samples of predictions of the most extreme cases.
- Testing extrapolation to extreme events by withholding the most extreme data in training and using these in a separate test set (as in [9], [27], [30]). (However, care needs to be taken to avoid the "forecaster's dilemma", where skill scores are distorted they are only evaluated on events with extreme outcomes when there is substantial noise in the target variables e.g. [32]. For example, it may be better to select extremes based on the predictors rather than the outcome, as in [9].)
- Testing how ML models perform as anomalies in input variables are magnified to correspond to events much more severe than any in the source data.
- Applying interpretability methods to assess trustworthiness for predicting extremes.

If existing machine learning approaches turn out not perform well enough at predicting a given extreme event, investigate more robust approaches, such as incorporating physical principles and building hybrids of conventional and ML-based models.

5 Conclusions

In order for ML to be applied broadly in weather and climate prediction and simulation systems, it needs to be shown that it can perform at least reasonably well for extreme events. ML models with high numbers of parameters, such as neural networks, may be expected to struggle in these cases as they typically need large samples of events to be trained to make skillful predictions. However, the studies reviewed here that do evaluate ML model skill on extremes actually indicate that ML-based systems can still perform well on out-of-sample extreme events, even for those with return periods of hundreds or thousands of years. This sample of studies is not enough to draw general conclusions from, though, and there are important questions that have not been addressed by any study that I could find. The situation could be greatly improved if study authors added certain simple diagnostics, and also if studies were designed to show the performance for extremes, as described above. This would be highly valuable for the rest of the community who would learn what ML methods are best to use to predict and simulate extreme events successfully.

Acknowledgments

I thank Peter Dueben in particular for discussions in preparing [1]. This work was supported by a NERC Independent Research Fellowship (Grant No. NE/S014713/1).

References

[1] P. A. G. Watson, “Machine learning applications for weather and climate need greater focus on extremes,” *Environ. Res. Lett.*, vol. 17, p. 111004, 2022, doi: 10.1088/1748-9326/ac9d4e.

[2] S. Ravuri *et al.*, “Skilful precipitation nowcasting using deep generative models of radar,” *Nature*, vol. 597, no. 7878, pp. 672–677, 2021.

[3] J. Pathak *et al.*, “FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators,” *arXiv:2202.11214*, 2022.

[4] R. Lam *et al.*, “Learning skillful medium-range global weather forecasting,” *Science*, eadi2336, 2023, doi: 10.1126/science.adi2336.

[5] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, “Accurate medium-range global weather forecasting with 3D neural networks,” *Nature*, vol. 619, no. 7970, pp. 533–538, 2023, doi: 10.1038/s41586-023-06185-3.

[6] L. Chen *et al.*, “FuXi: A cascade machine learning forecasting system for 15-day global weather forecast,” *arXiv:2306.12873*, 2023, [Online]. Available: https://arxiv.org/abs/2306.12873.

[7] K. Stengel, A. Glaws, D. Hettinger, and R. N. King, “Adversarial super-resolution of climatological wind and solar data,” *Proc. Natl. Acad. Sci.*, vol. 117, no. 29, pp. 16805–16815, 2020, doi: 10.1073/pnas.1918964117.

[8] L. Harris, A. T. T. McRae, M. Chantry, P. D. Dueben, and T. N. Palmer, “A generative deep learning approach to stochastic downscaling of precipitation forecasts,” *J. Adv. Model. Earth Syst.*, vol. 14, e2022MS003120, 2022, [Online]. Available: https://doi.org/10.1029/2022MS003120.

[9] E. Vosper *et al.*, “Deep Learning for Downscaling Tropical Cyclone Rainfall to Hazard-Relevant Spatial Scales,” *J. Geophys. Res.*, vol. 128, e2022JD038163, 2023, doi: 10.1029/2022JD038163.

[10] H. Addison, E. Kendon, S. Ravuri, L. Aitchison, and P. A. Watson, “Machine learning emulation of a local-scale UK climate model,” NeurIPS 2022 Workshop: Tackling Climate Change with Machine Learning. 2022. Available: https://www.climatechange.ai/papers/neurips2022/21.

[11] S. Rasp, M. S. Pritchard, and P. Gentine, “Deep learning to represent subgrid processes in climate models,” *Proc. Natl. Acad. Sci.*, vol. 115, no. 39, pp. 9684–9689, 2018, doi: 10.1073/pnas.1810286115.

[12] A. Gettelman *et al.*, “Machine learning the warm rain process,” *J. Adv. Model. Earth Syst.*, vol. 13, no. 2, e2020MS002268, 2021.

[13] M. D. Risser and M. F. Wehner, “Attributable human-induced changes in the likelihood and magnitude of the observed extreme precipitation during Hurricane Harvey,” *Geophys. Res. Lett.*, vol. 44, no. 24, pp. 12–457, 2017.

[14] G. J. Van Oldenborgh *et al.*, “Attribution of extreme rainfall from Hurricane Harvey, August 2017,” *Environ. Res. Lett.*, vol. 12, 124009, 2017.

[15] Environment Agency, “Flood and coastal erosion risk management: Long-term investment scenarios (LTIS) 2014,” 2014. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/381939/FCRM_Long_term_investment_scenarios.pdf.

[16] Environment Agency, “Meeting our future water needs: a national framework for water resources,” 2020. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_d

ata/file/872759/National_Framework_for_water_resources_main_report.pdf.

[17] D. Suri and P. A. Davies, “A Decade of Impact-Based NSWWS Warnings at the Met Office,” *Eur. Forecast.*, vol. 26, pp. 30–36, 2021, [Online]. Available: http://www.euroforecaster.org/latenews/suri.pdf.

[18] S. Y. Philip *et al.*, “Rapid attribution analysis of the extraordinary heat wave on the Pacific coast of the US and Canada in June 2021,” *Earth Syst. Dynam.*, vol. 13, no. 4, pp. 1689–1713, 2022, doi: 10.5194/esd-13-1689-2022.

[19] E. M. Fischer, S. Sippel, and R. Knutti, “Increasing probability of record-shattering climate extremes,” *Nat. Clim. Chang.*, vol. 11, no. 8, pp. 689–695, 2021, doi: 10.1038/s41558-021-01092-9.

[20] P. Mukhopadhyay *et al.*, “Unraveling the mechanism of extreme (More than 30 sigma) precipitation during august 2018 and 2019 over Kerala, India,” *Weather Forecast.*, vol. 36, no. 4, pp. 1253–1273, 2021, doi: 10.1175/WAF-D-20-0162.1.

[21] J. A. Smith, A. A. Cox, M. L. Baeck, L. Yang, and P. Bates, “Strange Floods: The Upper Tail of Flood Peaks in the United States,” *Water Resour. Res.*, vol. 54, no. 9, pp. 6510–6542, 2018, doi: 10.1029/2018WR022539.

[22] L. Magnusson, “Exploring machine-learning forecasts of extreme weather,” *ECMWF Newsletter*, vol. 176, pp. 8–9, 2023, [Online]. Available: https://www.ecmwf.int/en/newsletter/176/news/exploring-machine-learning-forecasts-extreme-weather.

[23] M. Kilavi *et al.*, “Extreme Rainfall and Flooding over Central Kenya Including Nairobi City during the Long-Rains Season 2018: Causes, Predictability, and Potential for Early Warning and Actions,” *Atmosphere*, vol. 9, no. 12. 2018, doi: 10.3390/atmos9120472.

[24] G. R. Herman and R. S. Schumacher, “Money Doesn’t Grow on Trees, but Forecasts Do: Forecasting Extreme Precipitation with Random Forests,” *Mon. Weather Rev.*, vol. 146, no. 5, pp. 1571–1600, 2018, doi: 10.1175/MWR-D-17-0250.1.

[25] R. A. Adewoyin, P. Dueben, P. Watson, Y. He, and R. Dutta, “TRU-NET: a deep learning approach to high resolution prediction of rainfall,” *Mach. Learn.*, vol. 110, no. 8, pp. 2035–2062, 2021, doi: 10.1007/s10994-021-06022-6.

[26] Y. Boulaguiem, J. Zscheischler, E. Vignotto, K. van der Wiel, and S. Engelke, “Modeling and simulating spatial extremes by combining extreme value theory with generative adversarial networks,” *Environ. Data Sci.*, vol. 1, e5, 2022, doi: 10.1017/eds.2022.4.

[27] J. M. Frame *et al.*, “Deep learning rainfall–runoff predictions of extreme events,” *Hydrol. Earth Syst. Sci.*, vol. 26, no. 13, pp. 3377–3392, 2022, doi: 10.5194/hess-26-3377-2022.

[28] P. Grönquist *et al.*, “Deep learning for post-processing ensemble weather forecasts,” *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 379, no. 2194, p. 20200092, 2021, doi: 10.1098/rsta.2020.0092.

[29] I. Lopez-Gomez, A. McGovern, S. Agrawal, and J. Hickey, “Global Extreme Heat Forecasting Using Neural Weather Models,” *arXiv:2205.10972*, 2022, doi: 10.48550/arxiv.2205.10972.

[30] S. Nevo *et al.*, “Flood forecasting with machine learning models in an operational framework,” *Hydrol. Earth Syst. Sci.*, vol. 26, no. 15, pp. 4013–4032, 2022, doi: 10.5194/hess-26-4013-2022.

[31] A. Hernanz, J. A. García-Valero, M. Domínguez, and E. Rodríguez-Camino, “A critical view on the suitability of machine learning techniques to downscale climate change projections: Illustration for temperature with a toy experiment,” *Atmos. Sci. Lett.*, vol. 23, no. 6, p. e1087, 2022, doi: https://doi.org/10.1002/asl.1087.

[32] S. Lerch, T. L. Thorarinsdottir, F. Ravazzolo, and T. Gneiting, “Forecaster’s Dilemma:

Extreme Events and Forecast Evaluation," *Stat. Sci.*, vol. 32, no. 1, pp. 106–127, 2017, [Online]. Available: http://www.jstor.org/stable/26408123.