Machine learning applications for weather and climate need greater focus on extremes

Peter Watson

Bristol University, UK

Based on and updated from *Env. Res. Lett., 17, 111004, 2022,*

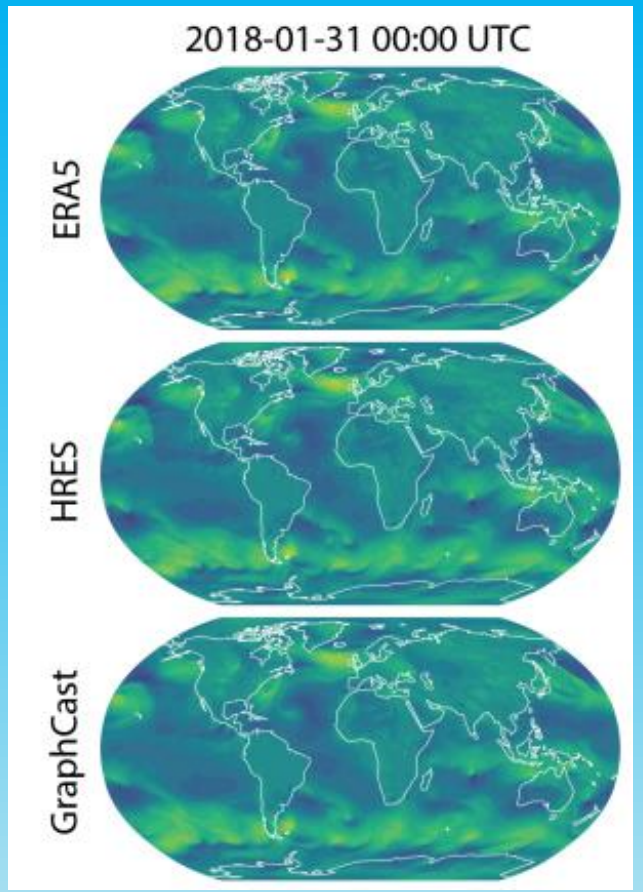
https://doi.org/10.1088/1748-9326/ac9d4e

Many studies have applied ML in impressive ways...

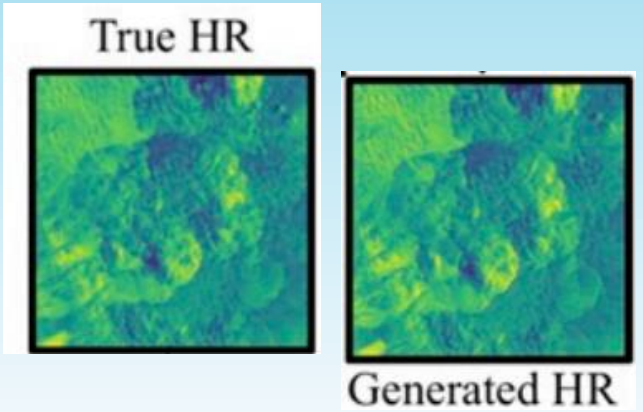
- e.g. NWP: Bi et al. (2022), Lam et al. (2023) and Chen et al. (2023) claim their models outperform state-of-art conventional models

- e.g. Downscaling, such as in Stengel et al. (2020)

But the performance for rare, extreme events of such systems is mostly unknown



Lam et al. (2022)

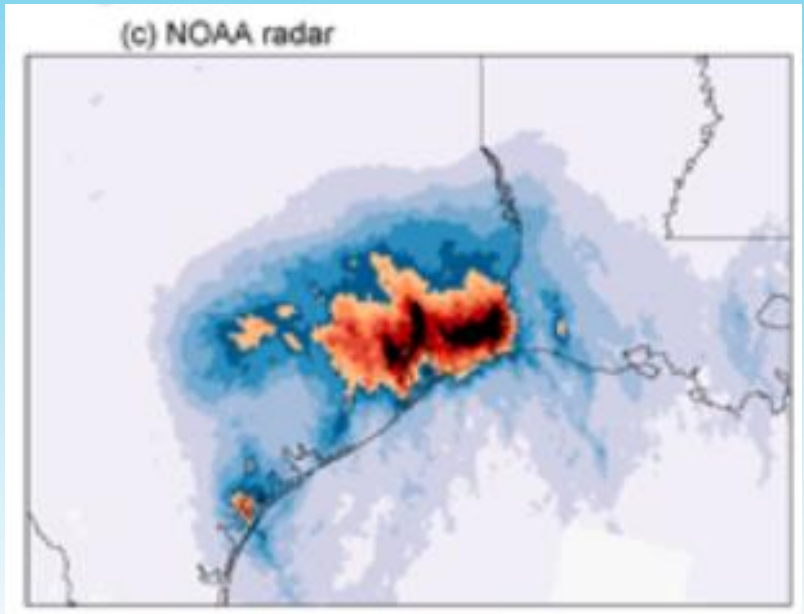


Stengel et al. (2020)

Importance of predictions for extremes

- Critical for
  - any general purpose system that will face events of unprecedented severity e.g. NWP, climate simulation postprocessing.
  - any specific use for simulating highly damaging weather and climate events e.g. extreme weather attribution
- (Though note not all extreme *impact* events are so meteorologically extreme)

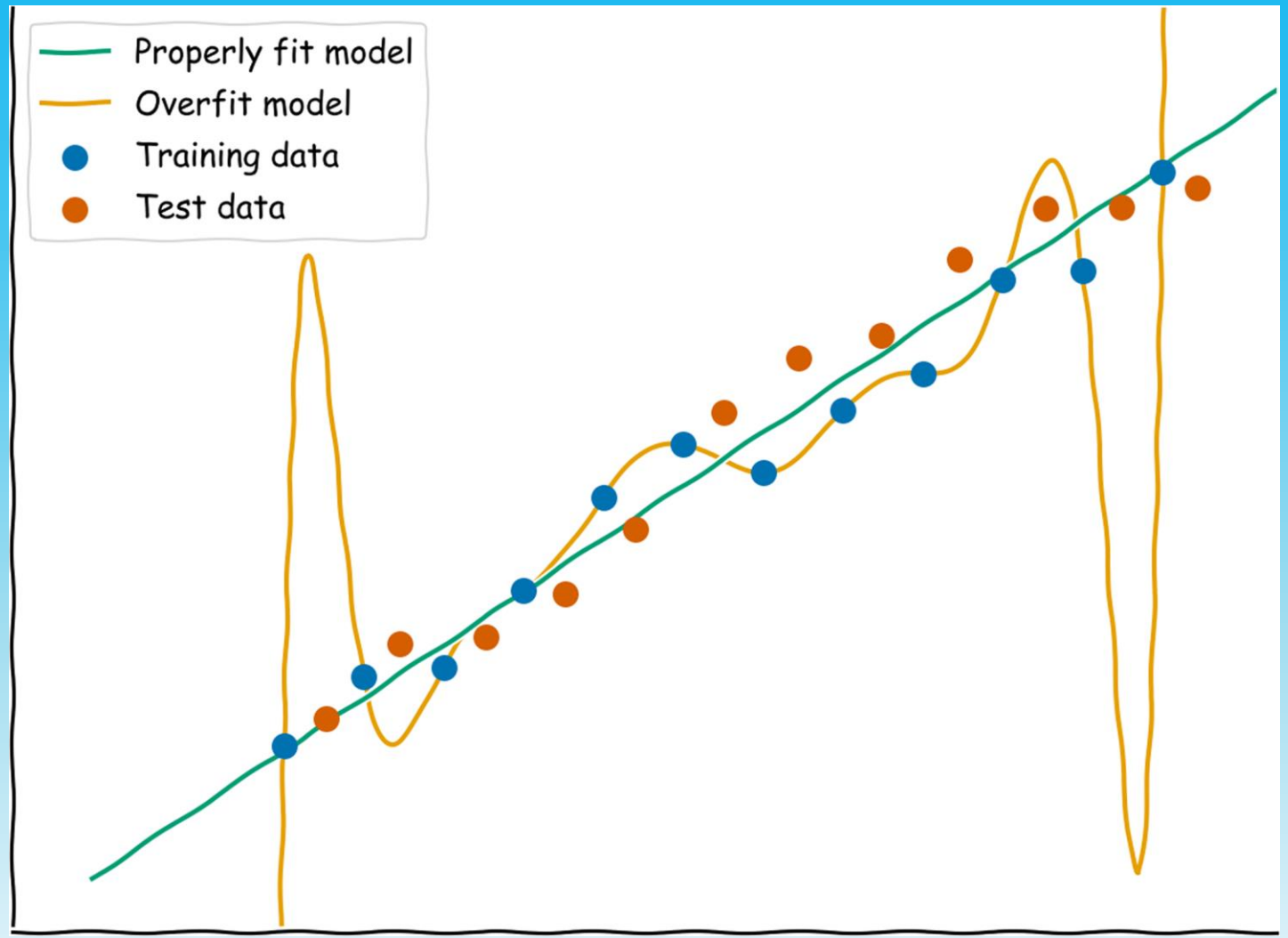
UK 2022 Storm Eunice gust return period up to ~200 years



Hurricane Harvey rainfall: return periods ~1000s of years
e.g. van Oldenborgh et al. (2017)

Reasons for scepticism

- Will there be big errors near the edge of the training domain?
- Can't easily tell from errors in the middle.
- Valuable to know what can work well e.g. big vs small models, pure ML vs hybrids
- Not as an afterthought – risk of wasting lots of time if the wrong options are pursued!



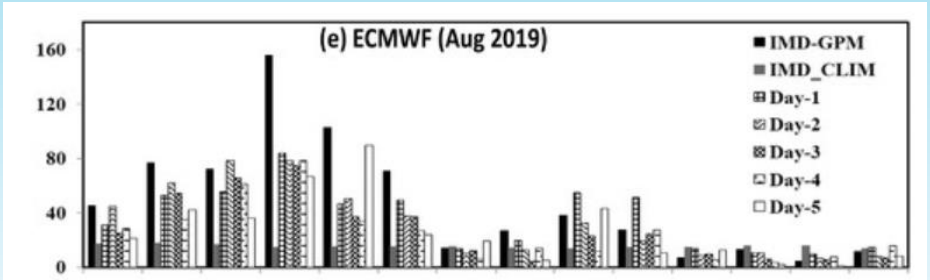
Lee et al. (2022)

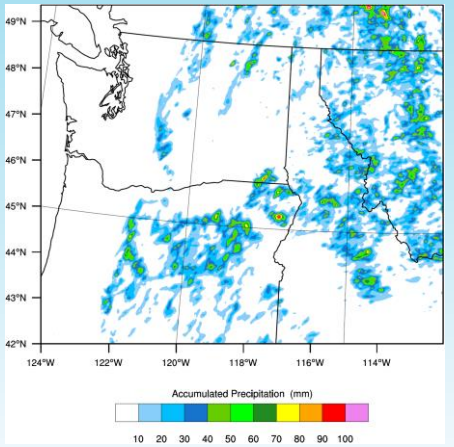
How extreme is “extreme”?

Percentile	Return period (daily data)	Intensity			
		Normal		Exponential	
		/σ	/p99	/σ	/p99
99th	100 days	2.3	1	4.6	1
99.997th	100 years	4.0	1.7	10	2.3
99.9997th	1000 years	4.5	1.9	13	2.8

But even greater extremes do occur, e.g.:

30σ 14-day rainfall in Kerala, India, 2018 and 2019 (Mukhopadhyay et al., 2021)

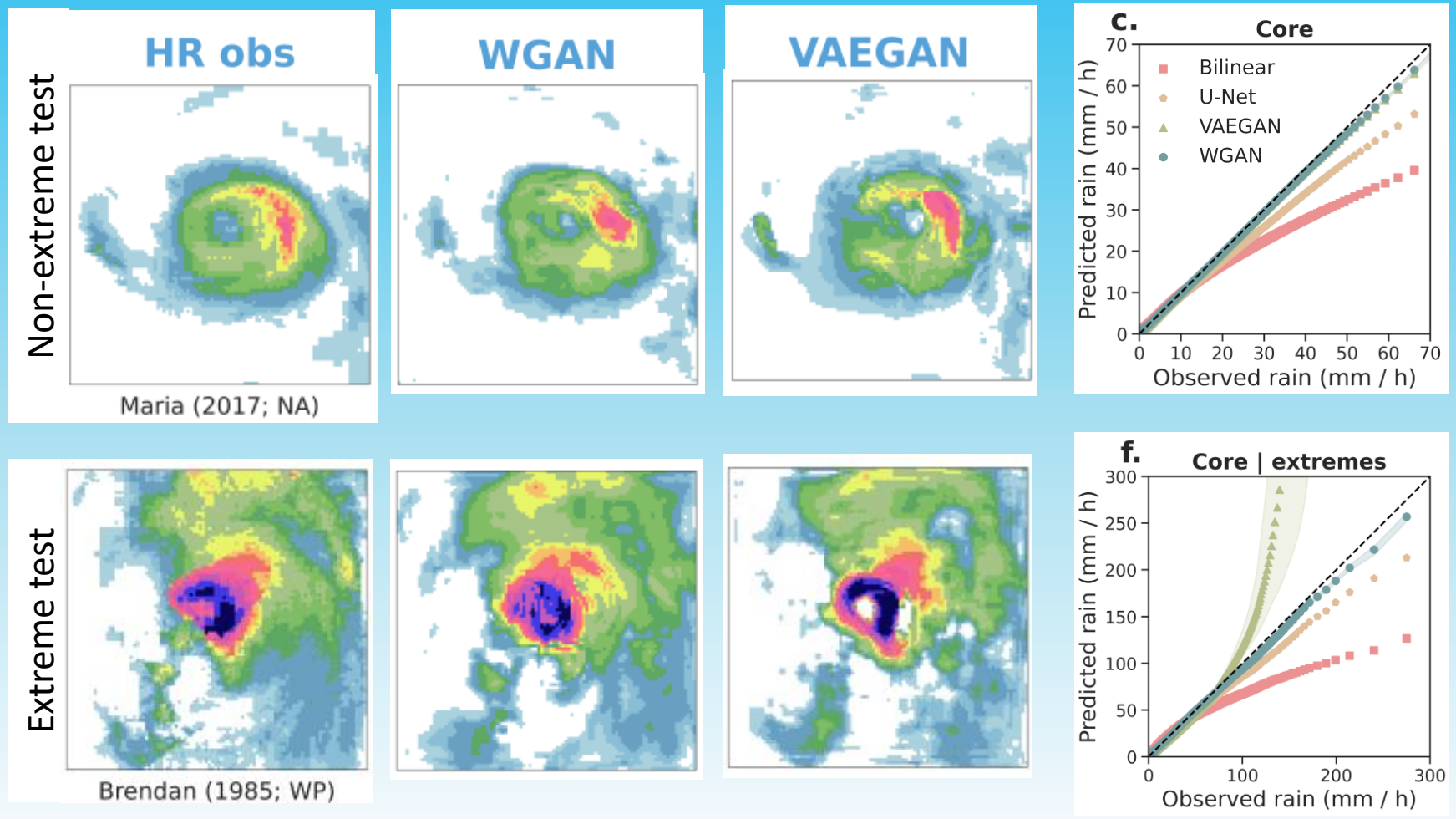




River discharge up to 200x the 10-year return level from rainstorms (Smith et al., 2018)

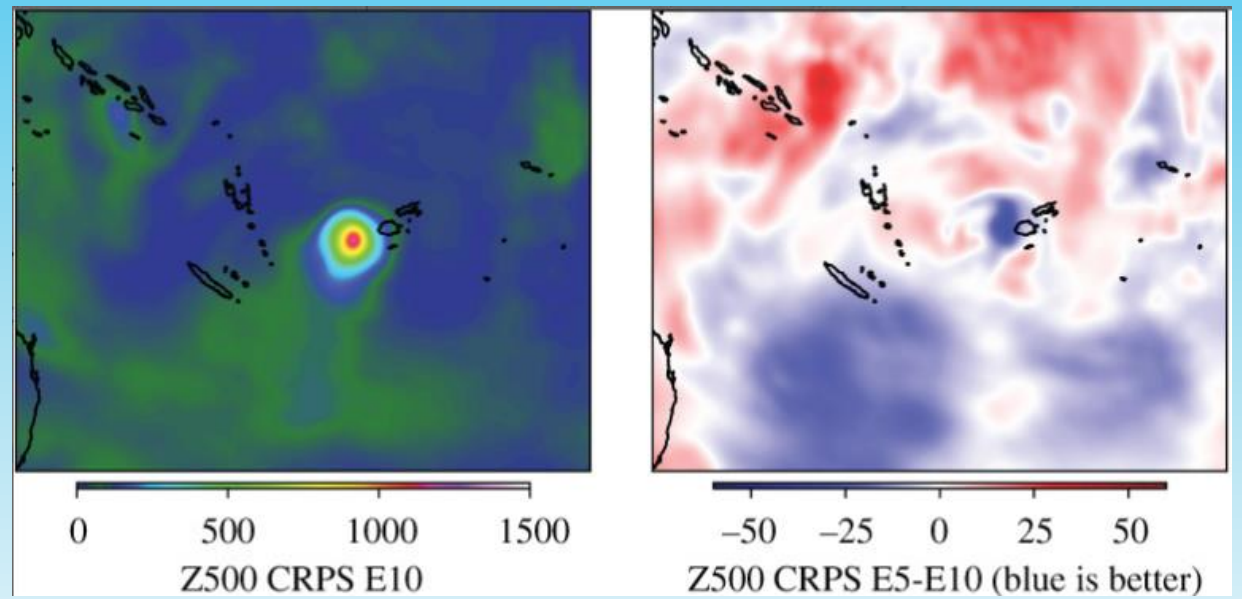
Example – hurricane rainfall downscaling

- Vosper et al. (2023): two GANs tested – both do well on typical events.
- Most 100 intense hurricanes held out as extreme test set.
  - One GAN does reasonably, another poorly.

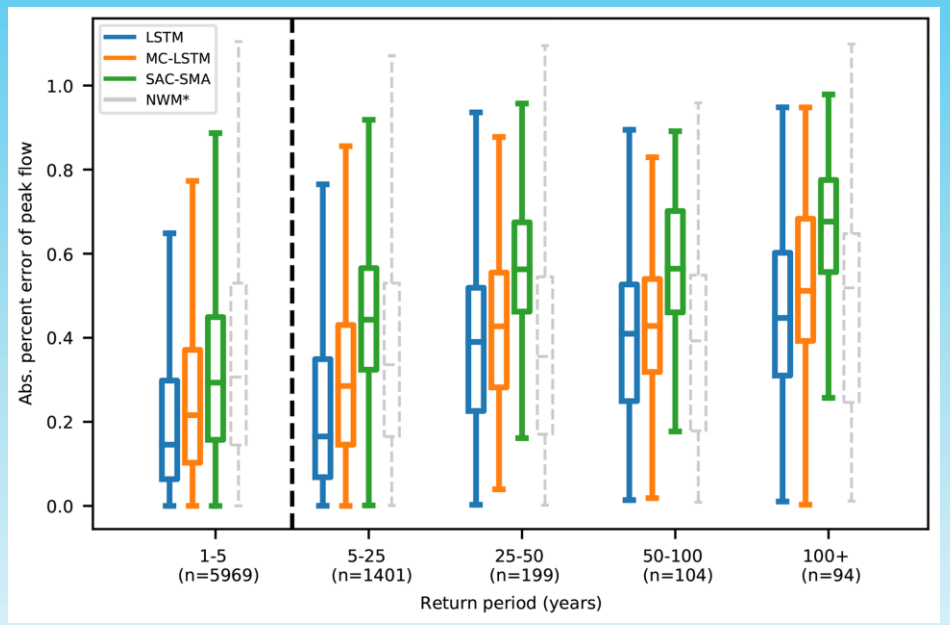


More success examples

- Other studies also show “big” ML models can perform at least OK for extremes.
- But I can only find ten studies that examine events with return periods of at least several years; 4 for ≳100 years.
- None show clear failure.



E.g. NWP post-processing for TC Winston (Gronquist et al., 2020)

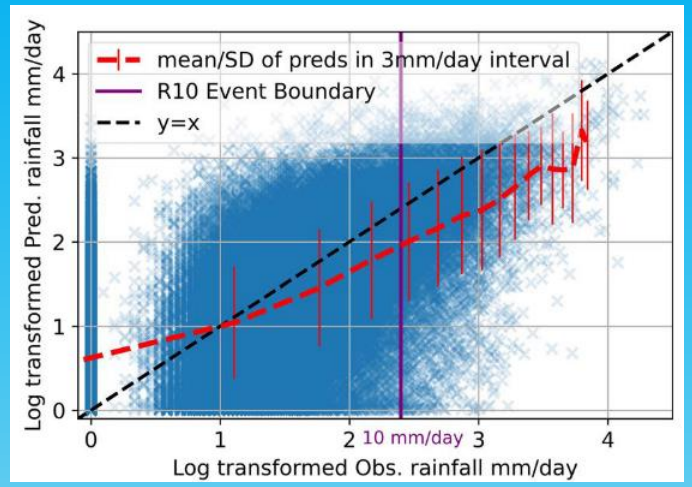


E.g. Peak river flow estimates (Frame et al., 2021)

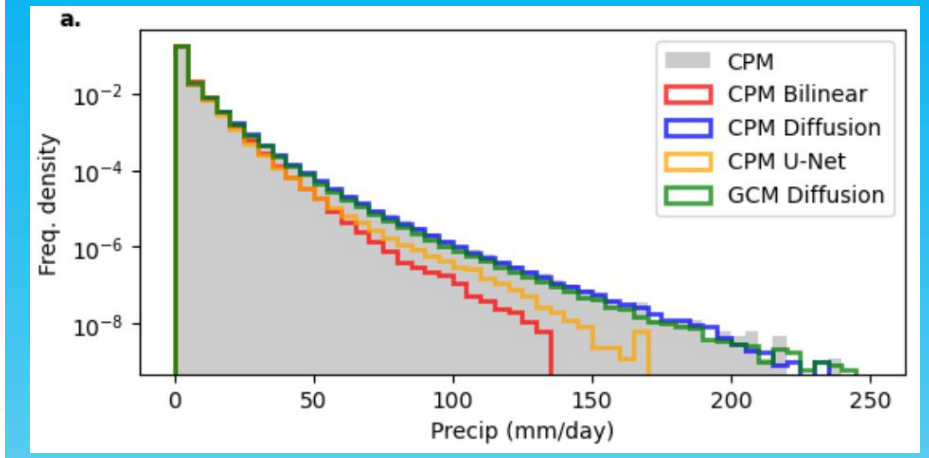
More useful diagnostics

E.g.

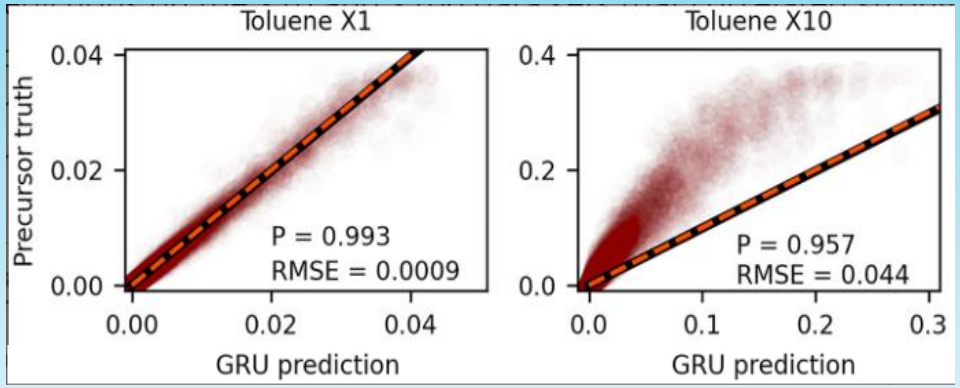
- Scatter plots, including most extreme points
- QQ plots/histograms, including the most extreme values
- Extrapolation tests, with scaled up inputs



Adewoyin et al. (2021) – UK precip



Addison et al., in prep – UK precip



Schreck et al. (2022) – aerosol chemistry

Summary

- ML does work at least sometimes for extremes, but evaluated in only a small number of studies.
- Critical to evaluate in many cases to show ML is useful, and not as an afterthought.
- More research indicating which methods have the best chance of working for extremes would be very valuable – good way to argue novelty.