Machine learning emulation of a local-scale UK climate model

APR Report

Henry Addison

Supervised by Peter Watson, Laurence Aitchison

Interactive AI CDT
University of Bristol
UK

April 2022

Cover Letter

Dear Tilo,

The following report covers my in-progress machine learning project about emulating a climate model in order to produce rainfall projections more cheaply. A large part of my summer project was a user requirements study interviewing producers and consumers of climate projections. For brevity I have not included that in the report but have shared separately the relevant parts of my summer project write-up in case you wish to read it.

Since this is a time to reflect, I have listed below some achievements over the past 9 months and some of the problems I face ahead. I have kept this brief and informal, intended as some possible starting points for further conversation.

Publications

None so far. I am due to present my work at EGU (large Geoscience conference) in May[1] and also at an RMets Student Conference in July.

I've presented my work at CDT events and a couple of times at Met Office events:

- A University and Met Office Climate & Hydrology day in April 2022
- A meeting of the Met Office Community of Practice in October 2021

Reflections

Looking forward to both the near and long term future of my PhD I am aware of a few challenges and frustrations.

- Evaluating stochastic samples and evaluating extremes given multitude of metrics and definitions
- What to do next? There is plenty more to do in this project but I struggle to plan ahead 2 years and imagine a thread for the final dissertation
- Poor multitasking: in particular making progress on sharing my work as well as reading what others have done and gathering and analysing my data (and within that balancing data processing, model training and evaluation)
- Technical frustrations:
  - Initial dataset did not include variables at more than one height leading to re-writing a data preprocessing to use a different data source.
  - Lack of understanding of geo data science data formats such as coordinate systems for gridded data and file metadata convention
  - Mismatch between GCM and CPM data for same timestamp

I have at times spent too much time presenting my work given the limited progress I've made on the machine learning side. However, I definitely feel much more confident presenting than I did at the start of the PhD. I would like to feel the same about writing.

[1] https://doi.org/10.5194/egusphere-egu22-6150

1 Introduction

Climate change is predicted to cause intensification of heavy rainfall extremes in the UK [1], [2]. The aim of this project is to help understand impacts of climate change on UK rainfall at a local ($\sim$ 2km) scale, including extreme events, to facilitate better adaptation decisions and inform mitigation policy.

Global climate models (GCMs) simulate physical processes of the climate. They are used to experiment with and explore the climate in different conditions such as different forcings due greenhouse gas (GHG) emission scenarios or paleoclimates of prehistoric Earth. The computational resource requirements of these models restrict their resolution to grids of 25km or more. This is too coarse to provide actionable insight [3].

The Met Office's UK Climate Projections introduced local projections at a resolution of 2.2km in the UKCP18 dataset [4] using a regional climate model (RCM), which still approximates the fundamental physical equations at a higher resolution by operating on a smaller spatial domain and using a GCM to set boundary conditions. However, these projections are a "major cost in terms of computing resource"[2] and required a trade-off amongst length, domain size and ensemble size to produce. The projections are available for only 12 ensemble members over a single climate change scenario for three 20-year chunks, all driven by the same GCM [2]. There are further technical challenges as well to implement such an RCM for a different driving GCM with a different codebase and programming interface.

Machine Learning (ML) techniques could emulate the more expensive high-resolution simulations using coarse GCM data as input and generate local projections more cheaply [5]–[7]. Furthermore, by emulating an RCM rather than replicating historical observations, changing relationships in the climate might be learnable from the physics-based modelling of the GCM and RCM. This means projections from a climate model could be complemented with further samples with realistic spatial and temporal structure enabling better understanding of the uncertainty of high-resolution precipitation. This is particularly true for extreme events which are poorly sampled by their infrequent nature and the small ensemble size of the UKCP18 2.2km RCM.

Sections 2 and 3 cover the background to downscaling of GCMs and machine learning applications to related problems. The in-progress method and immediate next-steps for the ML-based emulator are described in Section 4. Section 5 lays out possible future directions of work followed by a short conclusion in Section 6.

2 Downscaling

The starting point for climate projections are GCMs, which simulate the physical processes of the atmosphere for the entire globe. They get used to generate projections spanning centuries. The computational cost of such models require them to divide the Earth into a very coarse grid (typically resolutions of 25-300km). Such resolution means they cannot capture the influence of regional or local-scale attributes nor represent local phenomenon that are important to decision makers. For example, Schaller *et al.* [8] present an example where a change in temporal and spatial resolution of modelled rainfall has a large impact on where flooding is predicted to occur. Therefore it is common to downscale these global models to smaller grids and this is often done in one of two ways: dynamical regional climate models or statistical methods.

2.1 Dynamical downscaling

Dynamical downscaling uses finer-resolution simulations, RCMs. Like GCMs, RCMs use numerical approximations to dynamically simulate physical processes affecting the climate but by using GCMs to create boundary conditions they need only cover a sub-global region so may more feasibly use a finer horizontal grid [9]. Effectively an RCM is a higher resolution simulation embedded inside a coarser, global one.

Even at these higher resolutions, it is not possible to simulate all relevant processes. These processes are instead estimated using a parameterization scheme. These are heuristics that approximate the effects of these processes on the variables of the model. Clouds are such an example. Their small size does not allow them to be resolved in a grid box of a GCM but since their effects are significant their contribution is approximated based on variables available to the model [7]. The rough approximation of the parameterization scheme is one source of error in a climate model.

One symptom of errors in a climate model is the presence of systematic bias in the projections it generates. Bias can be observed in the output of an RCM for a period with observations. They may have, for example, temperatures that are too hot on average or predict too much rainfall on average compared to the observed records [10]. This can either be accepted as a known problem or attempts at correcting it can be made by adjusting the output to match observations (though this assumes that bias in the model remains the same of outputs outside of periods with observations) [9]. This latter approach of adjusting the output is known as bias correction.

Typically RCMs downscale to resolutions of around 10-50km but finer-grained simulations are possible and semi-tractable (the computational costs are still very large). One particular such class of finer-grained RCMs are Convection Permitting Models (CPMs).

2.1.1 Convection Permitting Models

CPMs are dynamical downscalers which use a fine enough resolution that they can also model convective processes rather than "represent[ing] the average effects of on the atmosphere of convection" [2] with a parameterization scheme.

In a first for national climate scenarios, in 2019 the UK's Met Office released 2.2km projections [2]. The UKCP18 Local Projections at 2.2km resolution use a CPM. The finer resolution of a CPM allows more detailed representation of the atmospheric effects of convection. This allows better ability to represent convective storms and more "credible estimates of changes in convectively-driven phenomena, such as hourly rainfall extremes" [2]. The trade-off for the high computational cost is a reduction in the ensemble size, time-horizon and emission scenarios explored compared to the Met Office's coarser projections.

In practice this means that only 12 for the 15 available climate models for a single emission scenario were downscaled to 2.2km using the CPM [2]. Emission scenarios, called Representative Concentration Pathways (RCPs), cover assumptions about changes to GHG emissions in the future and their corresponding increase in radiative forcing (i.e. the climate is made hotter due to reduced radiative output from the atmosphere caused by GHG retaining heat in the atmosphere). The emission scenario used by all the CPM projections is RCP8.5 (8.5 because it corresponds to a radiative forcing of $8.5\mathrm{Wm}^{-2}$ by 2100). It is the most extreme scenario that the Met Office covers in UKCP18 in which emissions are assumed to continue rising without any efforts to reduce them [11] and is considered by some to be "implausible" and "misleading" [12]. Further, rather the contiguous time period 100-year period from 1981 to 2080 used by the 12km RCM, only three 20-year chunks are covered: 1981-2000, 2021-2040, 2061-2080.

2.2 Statistical downscaling

Statistical downscaling attempts to fit a statistical relationship between a lower-resolution set of climate variables (predictors) from a simulation and higher-resolution values (predictands), often observations but for this project projections of a CPM. They are much less computationally expensive than dynamical downscaling methods but it is uncertain how well they work for future climate conditions [13].

Statistical downscaling techniques do not directly model the underlying physics. This means, unlike RCMs, they may be unable to capture emergent behaviour in the climate and extrapolate well to new conditions or to take advantage of patterns not realistically represented due to the coarse resolution of predictors [13]. This is particularly important for future projections because the climate is expected to change but the downscaled output cannot be compared and adjusted based on the observation record. However, they do offer advantages in that they may be more easily applied to different GCMs and are computationally cheaper and the problem of an assumed static climate is potentially mitigated in my project by using conditioning inputs from simulations of the future that will include climate change signals.

Many approaches have been tried based on both traditional statistics and machine learning. Gutiérrez *et al.* [3] assess the skill of 45 different methods for precipitation alone for a single European collaborative experiment, though the well-performing methods relevant to this project's problem are mainly based on generalized linear models (GLMs) or nearest neighbours (methods of analogs). I will use a GLM approach as a baseline (see Section 4.2). Some approaches fit a separate relationship for each grid box while others fit a model for the entire grid. No method has proven better than the others in general [3], [5], [14]. Vandal *et al.* [5] found no clear best option when comparing traditional approaches with some off-the-shelf machine learning approaches (Multi-task Sparse Structure Learning and Autoencoder Neural Networks) for daily and extreme precipitation over the Northeastern United States. Maraun *et al.* [13] note a lack of work on spatial dependence. The machine learning approaches tested by Vandal *et al.* [5] are deterministic in the sense that once trained, the same conditioning GCM input will produce the high-resolution output. Without a probabilistic element, these models struggle to predict the small-scale detail of precipitation [15]. I plan to curate more carefully a more recent probabilistic deep learning architecture to solve these problems with statistical downscaling.

2.2.1 Learning from projections rather than observations

Comparisons of different statistical downscaling approaches such as that carried out by VALUE have so far been mainly based on matching observations rather than emulating models to match their projections [13]. This limits the available data and makes no guarantees that such approaches will continue to work as the climate changes in the future. It also reduces the number of locations that require predictions for each time step with only 86 carefully chosen observation stations compared to nearly 300,000, closely-packed squares in the full 2.2km grid required to cover the whole of the UK and Ireland.

Walton *et al.* [16] carry out similar work using statistical methods to more quickly downscale a set of GCMs to 2km based on the dynamical downscaling of a separate set of GCMs. Their work focuses on LA and considers just the long-term mean warming based on the downscaling projections across the full set of GCMs. I plan to predict precipitation time series at daily and then sub-daily resolutions.

Works such as those by Gentine *et al.* [6] or Rasp *et al.* [7] also use projection datasets rather than observations to combine dynamical and ML-based statistical downscaling. The neural nets developed, CBRAIN [6] and NNCAM [7] respectively, are trained on high-resolution dynamical model data aimed at resolving clouds and intended to provide a replacement parameterization of subgrid process inside a GCM that is skilful enough to rival the dynamic approach they replace but fast enough to use to cover the whole globe. As with the work by Vandal *et al.* [5] these neural nets are deterministic and the intention to run them inside a GCM adds technical costs and constraints to their usage. They also only cover one part of the parameterization scheme rather than the entire model and do not produce higher resolution outputs.

The scope of this project is on precipitation in the UK and Ireland. In particular, I will emulate the CPM used by Met Office to make projections at 2.2km resolution. I used the Met Office's UKCP18 datasets for both the GCM global 60km projections [17] and the CPM local 2.2km projections [4]. The Met Office's CPM 2.2km projections are the first of their kind at a national level [2] so while other coarse global projections exist using different GCMs, they have no large, local-scale CPM output associated with them. The next section explores more advanced machine learning approaches that have worked for a similar problem, natural image super-resolution.

3 Machine learning and natural image super-resolution

Statistical downscaling of the full grid (rather than just a selection of observation stations) can be considered as similar to the super-resolution problem of natural images. The grid squares of the GCM and CPM outputs are like pixels with their intensity corresponding to climate variables such as hourly rainfall (albeit not just integer values between 0 and 255). Therefore I am interested in the extensive application of machine learning and deep generative models in particular to natural image generation and super-resolution. I intend to show that such an approach can be used to generalize beyond the MNIST, CIFAR-10 and ImageNet datasets commonly used to experiment with these models.

3.1 U-Net

U-Net [18] has been adapted from its roots in medical image segmentation for rainfall predictions, both with categorical outputs (through binning) [19] and continuous outputs [15], [20]. Ravuri *et al.* [15] find that it performs well on Critical Score Index (for a given threshold, what proportion of predictions from the model under test are on the same side of the threshold as the original UKCP18 dataset) but at the expense of producing blurry predictions which do not match the fine detail of rainfall events. U-Net is not a probabilistic model and unlikely to show great improvement over other off-the-shelf deep learning approaches as Vandal *et al.* [5] found. However, it should serve well as a baseline for comparison with a probabilistic approach.

3.2 Deep generative models

Common approaches for state-of-the-art sample generation, probability density estimation and super-resolution of natural images rely on generative models. These include Generative Adversarial Networks (GANs) [21], Variational Autoencoders (VAEs) [22], and autoregressive (AR) models like PixelRNN [23]. More recently Score-Based Generative Models (SCGMs, aka Noise-Conditioned Score Networks and Diffusion Models) [24]–[27] have shown competitive results in the natural image domain and offer desirable trade-offs compared to other approaches. As was discussed in Section 2.2, overall probabilistic models have been under-explored for use in downscaling.

3.2.1 Generative Adversarial Networks

GANs show excellent sampling performance in generating sharp, realistic images (both qualitatively and computationally) [21]. GANs [15], [28] have been successfully applied in similar problem domains: short-term forecasting (nowcasting) of sequences of rainfall radar fields and satellite cloud thickness fields.

GANs consist of two neural networks: a generator and a discriminator. The generator learns to create new samples, ideally ones that could come from some true distribution (usually provided in the form of a training dataset) to which the generator does not have direct access. The discriminator learns to tell apart samples from the true distribution from made up ones. By putting them in competition the discriminator gets better at spotting fake samples from the generator compared to real samples from the true distribution and using this signal the generator gets better at making samples which could be from the true distribution [29].

However, GANs do not model the underlying distribution so density estimation is impossible and GANs can suffer from mode collapse [30] meaning samples from the generator do not cover the full range of the distribution. It can also be difficult to find an optimal solution when training a GAN [29].

3.2.2 Variational Autoencoders

VAEs are a form of autoencoder with an encoding step and a decoding step. This encoding and decoding is probabilistic and rather than encoding into a latent space they encode into a latent distribution then decode samples from it [22], [31].

VAEs have been successfully used for image-based problems including generation [32], super-resolution [33] and labelling and segmentation [34]. An advantage of the autoencoder approach is it allows for simple manipulation in the latent space which lead to complex, meaningful changes in the pixel-space [35], [36]. The latent representation can also be used to feed other models thus allowing the use of VAEs with other approaches like GANs [37] or autoregressive models (e.g. PixelVAE[38]).

3.2.3 Autoregressive models

An AR approach like PixelRNN [23] can provide exact evaluation of the probability density of new points (unlike GANs and VAEs) and probabilistic sampling (unlike GANs) [30], [31], [39]. They show competitive density estimation performance in the domain of natural images [40].

AR models use recurrent networks to learn a joint distribution by using the probability chain rule and conditional distributions over the previous pixels (adapted for the particular context of this project to include conditioning on the inputs from a GCM for each pixel):

$$p(\mathbf{x_{1:N}}|\mathbf{g_{1:M}}) = \prod_{i=1}^{N} p(x_i|\mathbf{x_{1:i-1}}, \mathbf{g_{1:M}}) \tag{1}$$

where $\mathbf{x_{1:n}}$ denotes the first n pixels (going from top-left to bottom-right row-by-by) and $\mathbf{g_{1:M}}$ the whole corresponding GCM input. The special case $p(x_1|\mathbf{x_{1:0}}, \mathbf{g_{1:M}}$ is interpreted as simply $p(x_1|\mathbf{g_{1:M}})$.

The downside of this autoregressive approach is that sampling is a sequential process making it slow (O(N)) compared to VAEs. It has been shown for PixelCNN (an approach similar to and competitive with PixelRNN but computationally cheaper [41]) that this can be improved to O(logN) [42]. Sampling from an AR model will be much faster than running an RCM but won't match the sampling performance of a VAE approach even with this improvement.

3.2.4 Score-Based Generative Model

Probabilistic models assume that observed data, such as high-resolution rainfall over the UK, is drawn from an unknown distribution $p^*(\mathbf{x})$. A conditional model such as high-resolution rainfall conditioned on coarse GCM inputs $p^*(\mathbf{x}|\mathbf{g})$ can also be considered but for simplicity this section will stick with the context free version.

Song *et al.* [26] combine earlier approaches [24], [25] into a single framework called Score-Based Generative Models with Stochastic Differential Equations (SDE). The idea is to imagine a diffusion process $\{\mathbf{x}(t)_{t=0}^{T}\}$ modelled by an SDE:

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)d\mathbf{w} \tag{2}$$

When run forward a sample, $\mathbf{x}(0)$, from the data distribution, p_0, is gradually perturbed over time into a sample from a final noise distribution, p_T, somewhat like how a structured gas sample will gradually diffuse randomly across a room. The final distribution is chosen as something tractable for sampling, usually a Gaussian.

More interesting for us is running the diffusion process backwards:

$$\mathrm{d}\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(x)]dt + g(t)d\bar{\mathbf{w}} \tag{3}$$

By solving this, samples from p_T (which is easy by design) can be converted into samples from the original data distribution. This requires two steps: calculating the score, $\nabla_{\mathbf{x}} \log p_t(x)$, and then applying numerical approaches to solve Equation 3.

The score is estimated as a neural net $s_\theta(\mathbf{x}, t)$ where θ are determined by minimizing:

$$\mathbb{E}_t\{\lambda(t)\mathbb{E}_{\mathbf{x}(0)}\mathbb{E}_{\mathbf{x}(0)|\mathbf{x}(t)}\left[||s_\theta(\mathbf{x}(t), t) - \nabla_{\mathbf{x}(t)} \log p_{0t}(\mathbf{x}(t)|\mathbf{x}(0))||_2^2\right\} \tag{4}$$

where λ is a positive weighting function that is chosen along with f and g.

Song *et al.* [26] summarize three approaches for solving the reverse SDE. General-purpose numerical methods can be used to find approximate solutions to the SDE. Predictor-Corrector sampling takes this a step further by using making use of estimated score at each timestep to apply a correction to the sample estimated at that timestep by the general purpose solver. Alternatively the problem can be reformulated as a deterministic process without affecting the trajectory probabilities and in turn solved using an ODE solver.

Diffusion models show excellent performance in problems in the natural image domain [26], [27]. Diffusion models offer a good trade-off in terms of sample sharpness (VAEs tend to produce blurry samples), sample diversity (no mode collapse like GANs) and sampling cost (more efficient than AR). They have not been applied in the domain of climate model downscaling.

4 In-progress method of emulator study

In this section, the in-progress method is described including details of the datasets, design of the emulator and baselines, preliminary samples and plans for evaluation and short-term next steps.

4.1 Dataset

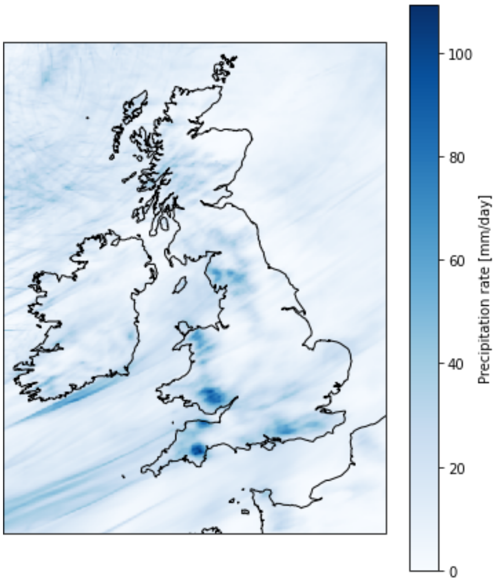
I am interested in emulating the precipitation output from the Met Office's UK CPM. I am using the Met Office's UKCP18 datasets for both the GCM global 60km projections [17] and the CPM local 2.2km projections [4]. For the whole UK and Ireland domain, the high-resolution grid is 484 x 606 while the coarse grid is 17 x 23. For practical purposes I limit the geographical domain to the area covered by a 64x64 high-resolution grid centred on London. London's population density makes this an area of high impact and the grid size matches to similar super-resolution problems of 64x64 pixel images. The temporal frequency is daily. Examples from the CPM and GCM can be seen in Figures 1a and 1b for the whole UK and Ireland domain, and examples from the London domain can be seen in Figures 1c and 1d

In the standard super-resolution problem, the mapping of low-resolution to high-resolution operates on the same channels (RGB or grey-scale) and as a starting point I similarly attempt to downscale coarse precipitation to high-resolution precipitation. However, it is not possible to train a model using GCM data as conditioning input because the weather conditions in GCM and CPM are different for the same timestamp. This is due to how the simulations work and as far are the CPM is concerned, the GCM stops at the edge of Europe. This allows scope of the GCM and CPM to diverge (e.g. weather fronts rotate or change speed). Therefore, the dataset used in the training phase will consist of coarsened data from the CPM (scale factors 2x up to the ideal 27x are used) to use as the conditioning input and the full, high-resolution precipitation used as output. It will be better to use other variables as input as the improved modelling possible in the CPM means coarsened rainfall from the CPM will be different to the inherently coarse rainfall from the GCM. Variables such as wind, temperature and humidity are well-represented in GCM and so the next step will be to use inputs based on these. Based on work by Chan *et al.* [43], the next inputs to consider are relative vorticity (which uses wind components to measure circulation in the atmosphere and correlated with how stormy or fair conditions are), vertical stability (which uses a range of vertical temperatures and indicates how favourable conditions are for convection) and mean sea-level pressure. Once fitted, samples can then be drawn from the model either using coarsened CPM data or data from a GCM.

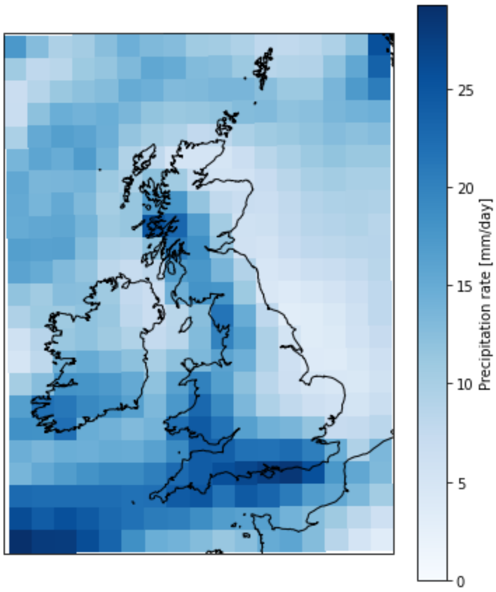
Extremes are of particular interest. Data are available hourly for three 20-year periods (1981-2000, 2021-2040, 2061-2080) for 12 ensemble members in the UKCP18 dataset. I will use a hold-out test set of 10% of the data to test the model's ability to extrapolate to the extremes. This will consist of the wettest days (determined by the sum of the CPM precipitation over the entire London grid) and 5 days either side to account for the temporal dependence of precipitation. A further 20% is used as a set of extremes to use for validation purposes. The remaining 70% of the data is split randomly into standard training, validation and test sets (70%, 20%, 10%).

4.2 Planned baselines

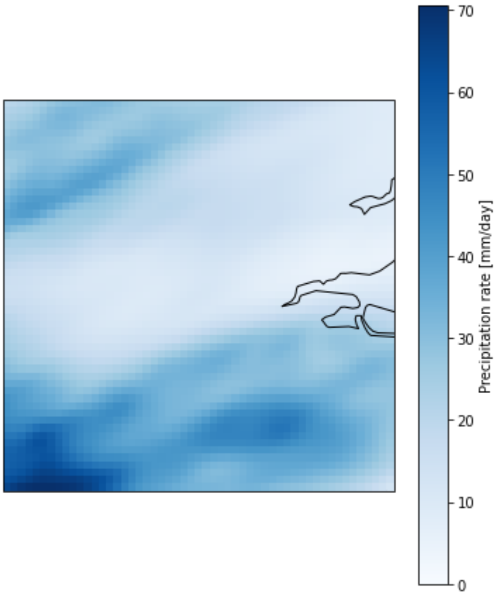
As baselines to compare the new model, I will use a two-step Generalized Linear Model (GLM) approach that performed well in the VALUE experiment [3] (first step determines if a wet day or not, second step models rainfall intensity on wet days) and a deep learning regressor based on U-Net and Rainnet [15], [20]. Due to a change in data source, results from these are not available.



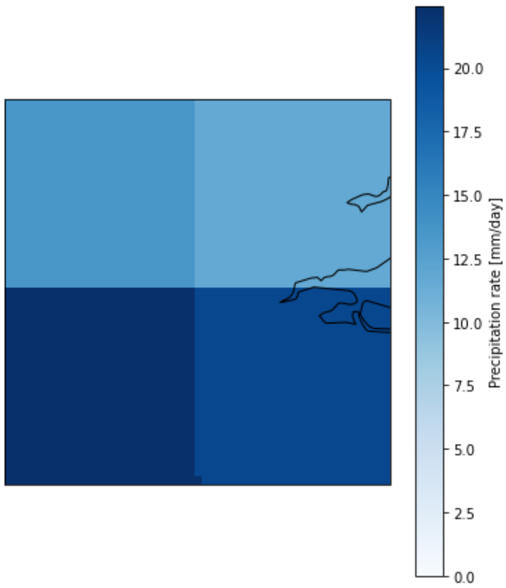
(a) High-resolution CPM rainfall intensity



(b) Low-resolution GCM rainfall intensity



(c) High-resolution CPM rainfall intensity over London area



(d) Low-resolution GCM rainfall intensity over London area

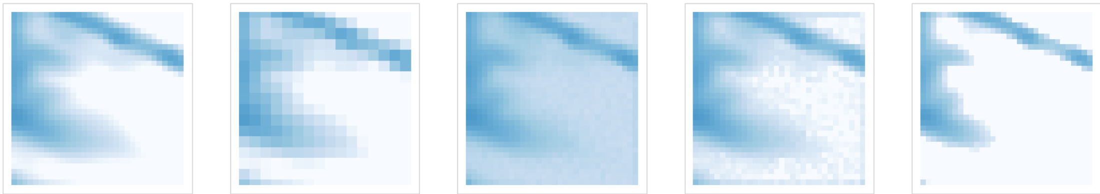
Figure 1: Example rainfall intensity (mm/day) map plots for same day, 1981-01-26, from different datasets

4.3 Emulator design

The emulator is a score-based generative model based on NCSN++[26], adapted to allow conditional training and sampling. Such models offer a nice trade-off in terms of sample diversity (e.g. mode collapse in GANs), sample sharpness (e.g. blurriness of VAEs) and sampling cost (e.g. slow sampling of AR models).

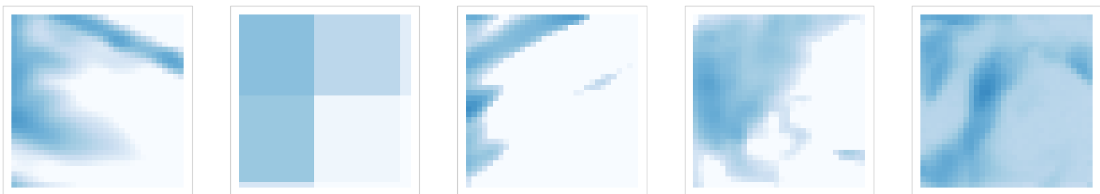
4.4 Preliminary Samples

Figure 2 shows high resolution samples from an untuned model conditioned on precipitation coarsened between 2x, 8x and 16x.



(a) 2x downscaling

(b) 8x downscaling



(c) 16x downscaling

Figure 2: Example samples for different scale factors: CPM rainfall (column 1), conditioning coarsened CPM precipitation (column 2), conditional samples (columns 3 to 5)

4.5 Proposed Evaluation

Evaluation will take two forms: comparison of climatological statistics of samples from emulator (to those from baselines and the CPM itself) and examination of samples for individual events. The former will compare metrics based on output aggregated across time and space such as the count of very wet days and ensuring preservation of seasonal variability and intensity-duration frequency curves from the CPM. This is to ensure the output is plausible from the long-term climate point of view. Individual event evaluation will use metrics from weather-forecasting to ensure that samples produced by the emulator realistically re-produce the fine spatial structure of individual precipitation events. These are metrics such as critical success index, fractional skill score and continuous ranked probability score. In both cases, since there will be stochastic samples, I am interested in the spread in the metrics as well as the central tendency.

4.6 Next steps

I am currently in the process of fitting an emulator model (and U-Net and GLM baselines) conditioned on coarsened relative vorticity (a derivative of wind) and evaluate it based on the coarsened CPM data both in terms of the climatological statistics and individual events.

The next steps after that will be to evaluate samples conditioned on GCM data (which will only allow for evaluation in terms of climatological statistics) and try other conditioning variables based on temperature and humidity.

5 Future Work

There is much more work to be carried out. The most important are assessing the performance of the emulator on much more extreme events and generalizing to other climate models beyond those used in UKCP18. There is also work to be done extending the emulator to higher temporal frequencies, larger geographical domains (to explore longer range spatial correlations) and incorporating expert knowledge of climate scientists.

5.1 More extreme extremes

The extreme test set is roughly the 90th percentile and above (so 1 in 10 day return period). As extremes go, projection consumers are interested in much more extreme events. It will be of interest to study how well the emulator characterizes events with a 1-in-100 year return period.

5.2 Generalizing to other climate model ensembles

The method above describes an emulator trained and tested data from ensemble members of the Met Office's GCM and CPM used for the UKCP18 project. This is one of a number of different set of climate models. Further work will explore how well the emulation process will generalize to downscaling large ensembles of other climate models. As well as producing more samples for climate projections with existing high resolution variables, it would also be of value to downscale climate model projections for which there are not yet dynamically downscaled equivalents. The emulator could thus be a tool to allow climate modellers to decide which temporal and spatial domains and greenhouse gas emission scenarios to target the limited resources available for producing the more expensive dynamically downscaled projections.

5.3 Higher frequencies and sequences

The work so far as focussed on single, daily snapshots of variables. Data of higher temporal frequency is available (down to hourly for precipitation). This higher frequency and the temporal evolution of events is of interest to consumers. There may also be value in working on sequences of inputs and outputs. This may be done by either by considering large blocks at a time or using an autoregressive approach.

5.4 Incorporating human knowledge

The emulator does not use any knowledge of the underlying physical aspects of each pixel. Surface topography is known to have an effect on rainfall. High-resolution digital topography could be added as an input. Alternatively similar information could be included with a simple mean rain field by temporal aggregation of the training set.

Experts could also be involved in the evaluation process. Their knowledge would be helpful to test the realism and plausibility of samples.

6 Conclusion

High-resolution projections of precipitation are needed to more effectively adapt to future changes in precipitation. However, to make predictions solely with numerical climate simulations is very expensive. A solution is to use machine learning to fit a mapping from cheaper, low-resolution climate projections to high-resolution precipitation. Generative models allow for both fast sample generation which can contain detail (compared to more blurry deterministic approaches like U-Net) and to generate more samples without needing new inputs. The samples from such a model will complement the simulation-based projections.

Score-based generative models offer a good trade-off over sampling cost, sample diversity and sample quality. The next steps are to continue training and evaluating a score-based generative model that can efficiently produce realistic, stochastic samples of high-resolution rainfall based on coarsened inputs. Once this is complete there are numerous other opportunities for improving the skill and utility of the emulator so that it can be put to use by organizations such as the Met Office.

7 Acknowledgements

As well as my supervisors at University of Bristol, Peter Watson and Laurence Aitchison, I have also received valuable help and guidance from Elizabeth Kendon of the Met Office and Univerity of Bristol and Suman Ravuri of DeepMind. Compute and storage resources for this work was carried out using Blue Pebble, part of the computational facilities of the Advanced Computing Research Centre, University of Bristol[2], and JASMIN, run by STFC's Scientific Computing Department and CEDA[3].

References

[1] M. G. Donat, A. L. Lowry, L. V. Alexander, P. A. O'Gorman and N. Maher, "More extreme precipitation in the world's dry and wet regions," *Nature Climate Change*, vol. 6, pp. 508–513, 2016. DOI: 10.1038/nclimate2941.

[2] E. J. Kendon, G. Fosser, J. Murphy *et al.*, "Ukcp convection-permitting model projections: Science report," 2019. [Online]. Available: https://www.metoffice.gov.uk/pub/data/weather/uk/ukcp18/science-reports/UKCP-Convection-permitting-model-projections-report.pdf.

[3] J. M. Gutiérrez, D. Maraun, M. Widmann *et al.*, "An intercomparison of a large ensemble of statistical downscaling methods over europe: Results from the value perfect predictor cross-validation experiment," *International Journal of Climatology*, vol. 39, no. 9, pp. 3750–3785, 2019. DOI: 10.1002/joc.5462.

[4] Met Office Hadley Centre, *UKCP18 Local Projections at 2.2km Resolution for 1980-2080*, Centre for Environmental Data Analysis, 2019. [Online]. Available: https://catalogue.ceda.ac.uk/uuid/d5822183143c4011a2bb304ee7c0baf7 (visited on 09/06/2021).

[5] T. Vandal, E. Kodra and A. R. Ganguly, "Intercomparison of machine learning methods for statistical downscaling: The case of daily and extreme precipitation," *Theoretical and Applied Climatology*, vol. 137, no. 1-2, pp. 557–570, 2018. DOI: 10.1007/s00704-018-2613-3.

[6] P. Gentine, M. Pritchard, S. Rasp, G. Reinaudi and G. Yacalis, "Could machine learning break the convection parameterization deadlock?" *Geophysical Research Letters*, vol. 45, no. 11, pp. 5742–5751, 2018, ISSN: 0094-8276 1944-8007. DOI: 10.1029/2018gl078202.

[7] S. Rasp, M. S. Pritchard and P. Gentine, "Deep learning to represent subgrid processes in climate models," *Proc Natl Acad Sci U S A*, vol. 115, no. 39, pp. 9684–9689, 2018, ISSN: 1091-6490 (Electronic) 0027-8424 (Linking). DOI: 10.1073/pnas.1810286115.

[8] N. Schaller, J. Sillmann, M. Muller *et al.*, "The role of spatial and temporal model resolution in a flood event storyline approach in western norway," *Weather and Climate Extremes*, vol. 29, p. 100 259, 2020, ISSN: 2212-0947. DOI: ARTN10025910.1016/j.wace.2020.100259.

[2] http://www.bristol.ac.uk/acrc/
[3] https://jasmin.ac.uk/

[9] F. J. Tapiador, A. Navarro, R. Moreno, J. L. Sánchez and E. García-Ortega, "Regional climate models: 30 years of dynamical downscaling," *Atmospheric Research*, vol. 235, p. 104 785, 2020, ISSN: 01698095. DOI: 10.1016/j.atmosres.2019.104785.

[10] J. H. Christensen, F. Boberg, O. B. Christensen and P. Lucas-Picher, "On the need for bias correction of regional climate change projections of temperature and precipitation," *Geophysical Research Letters*, vol. 35, no. 20, 2008, ISSN: 0094-8276. DOI: ArtnL2070910.1029/2008gl035694.

[11] Met Office Hadley Centre, "UKCP18 Guidance: Representative Concentration Pathways," Tech. Rep., 2018. [Online]. Available: https://www.metoffice.gov.uk/binaries/content/assets/metofficegovuk/pdf/research/ukcp/ukcp18-guidance---representative-concentration-pathways.pdf.

[12] Z. Hausfather and G. P. Peters, "Emissions - the 'business as usual' story is misleading," *Nature*, vol. 577, no. 7792, pp. 618–620, 2020, ISSN: 1476-4687 (Electronic) 0028-0836 (Linking). DOI: 10.1038/d41586-020-00177-3.

[13] D. Maraun, M. Widmann and J. M. Gutierrez, "Statistical downscaling skill under present climate conditions: A synthesis of the value perfect predictor experiment," *International Journal of Climatology*, vol. 39, no. 9, pp. 3692–3703, 2019. DOI: 10.1002/joc.5877.

[14] H. J. Fowler, S. Blenkinsop and C. Tebaldi, "Linking climate change modelling to impacts studies: Recent advances in downscaling techniques for hydrological modelling," *International Journal of Climatology*, vol. 27, no. 12, pp. 1547–1578, 2007. DOI: 10.1002/joc.1556.

[15] S. Ravuri, K. Lenc, M. Willson *et al.*, "Skillful precipitation nowcasting using deep generative models of radar," 2021. DOI: arxiv:2104.00954.

[16] D. B. Walton, F. P. Sun, A. Hall and S. Capps, "A hybrid dynamical-statistical downscaling technique. part i: Development and validation of the technique," *Journal of Climate*, vol. 28, no. 12, pp. 4597–4617, 2015. DOI: 10.1175/Jcli-D-14-00196.1.

[17] Met Office Hadley Centre, *UKCP18 Global Projections at 60km Resolution for 1900-2100*, Centre for Environmental Data Analysis, 2018. [Online]. Available: https://catalogue.ceda.ac.uk/uuid/97bc0c622a24489aa105f5b8a8efa3f0 (visited on 09/06/2021).

[18] O. Ronneberger, P. Fischer and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention, Pt Iii*, vol. 9351, pp. 234–241, 2015, ISSN: 0302-9743. DOI: 10.1007/978-3-319-24574-4_28.

[19] S. Agrawal, L. Barrington, C. Bromberg, J. Burge, C. Gazen and J. Hickey, "Machine learning for precipitation nowcasting from radar images," 2019. DOI: arxiv:1912.12132.

[20] G. Ayzel, T. Scheffer and M. Heistermann, "Rainnet v1.0: A convolutional neural network for radar-based precipitation nowcasting," *Geoscientific Model Development*, vol. 13, no. 6, pp. 2631–2644, 2020, ISSN: 1991-9603. DOI: 10.5194/gmd-13-2631-2020.

[21] T. Karras, T. Aila, S. Laine and J. Lehtinen, *Progressive Growing of GANs for Improved Quality, Stability, and Variation*, 2018. arXiv: 1710.10196 [cs.NE].

[22] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2014.

[23] A. Van Den Oord, N. Kalchbrenner and K. Kavukcuoglu, "Pixel recurrent neural networks," in *33rd International Conference on Machine Learning, ICML 2016*, vol. 4, 2016, pp. 2611–2620.

[24] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Advances in Neural Information Processing Systems*, vol. 32. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/3001ef257407d5a371a96dcd947c7d93-Abstract.html.

[25] J. Ho, A. Jain and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

[26] Y. Song, J. Sohl-Dickstein, Diederik, A. Kumar, S. Ermon and B. Poole, “Score-based generative modeling through stochastic differential equations,” in *ICLR*, 2021. DOI: arxiv:2011.13456.

[27] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” in *Advances in Neural Information Processing Systems*, vol. 34, 2021. [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

[28] J. Leinonen, D. Nerini and A. Berne, “Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network,” *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–13, 2020.

[29] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta and A. A. Bharath, “Generative adversarial networks an overview,” *Ieee Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018, ISSN: 1053-5888. DOI: 10.1109/Msp.2017.2765202.

[30] A. Grover, M. Dhar and S. Ermon, “Flow-gan: Combining maximum likelihood and adversarial learning in generative models,” *Thirty-Second Aaai Conference on Artificial Intelligence / Thirtieth Innovative Applications of Artificial Intelligence Conference / Eighth Aaai Symposium on Educational Advances in Artificial Intelligence*, vol. 32, no. 1, pp. 3069–3076, 2018.

[31] D. P. Kingma and M. Welling, “An introduction to variational autoencoders,” *Foundations and Trends in Machine Learning*, vol. 12, no. 4, pp. 4–89, 2019, ISSN: 1935-8237. DOI: 10.1561/2200000056.

[32] K. Gregor, F. Besse, D. Jimenez Rezende, I. Danihelka and D. Wierstra, “Towards conceptual compression,” in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon and R. Garnett, Eds., vol. 29, Curran Associates, Inc., 2016.

[33] Z.-S. Liu, W.-C. Siu and Y.-L. Chan, “Photo-realistic image super-resolution via variational autoencoders,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 4, pp. 1351–1365, 2021, ISSN: 1051-8215 1558-2205. DOI: 10.1109/tcsvt.2020.3003832.

[34] K. Sohn, H. Lee and X. Yan, “Learning structured output representation using deep conditional generative models,” in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama and R. Garnett, Eds., vol. 28, Curran Associates, Inc., 2015.

[35] T. D. Kulkarni, W. F. Whitney, P. Kohli and J. Tenenbaum, “Deep convolutional inverse graphics network,” in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama and R. Garnett, Eds., vol. 28, Curran Associates, Inc., 2015.

[36] T. White, “Sampling generative networks,” *arXiv preprint*, 2016. DOI: arXiv:1609.04468.

[37] A. B. L. Larsen, S. K. Sønderby, H. Larochelle and O. Winther, “Autoencoding beyond pixels using a learned similarity metric,” in *Proceedings of The 33rd International Conference on Machine Learning*, B. Maria Florina and Q. W. Kilian, Eds., vol. 48, PMLR, 2016, pp. 1558–1566.

[38] I. Gulrajani, K. Kumar, F. Ahmed *et al.*, “Pixelvae: A latent variable model for natural images,” in *International Conference on Learning Representation*, 2017.

[39] I. Kobyzev, S. Prince and M. Brubaker, “Normalizing flows: An introduction and review of current methods,” *IEEE Trans Pattern Anal Mach Intell*, vol. PP, 2020, ISSN: 1939-3539 (Electronic) 0098-5589 (Linking). DOI: 10.1109/TPAMI.2020.2992934.

[40] J. Ho, X. Chen, A. Srinivas, Y. Duan and P. Abbeel, “Flow++: Improving flow-based generative models with variational dequantization and architecture design,” in *Proceedings of the 36th International Conference on Machine Learning*, C. Kamalika and S. Ruslan, Eds., vol. 97, PMLR, 2019, pp. 2722–2730.

[41] A. van den Oord, N. Kalchbrenner, L. Espeholt, k. kavukcuoglu koray, O. Vinyals and A. Graves, "Conditional image generation with pixelcnn decoders," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon and R. Garnett, Eds., vol. 29, Curran Associates, Inc.

[42] S. Reed, A. Oord, N. Kalchbrenner *et al.*, "Parallel multiscale autoregressive density estimation," in *Proceedings of the 34th International Conference on Machine Learning*, P. Doina and T. Yee Whye, Eds., vol. 70, PMLR, 2017, pp. 2912–2921.

[43] S. C. Chan, E. J. Kendon, N. Roberts, S. Blenkinsop and H. J. Fowler, "Large-scale predictors for extreme hourly precipitation events in convection-permitting climate simulations," *Journal of Climate*, vol. 31, no. 6, pp. 2115–2131, 2018, ISSN: 0894-8755. DOI: 10.1175/Jcli-D-17-0404.1.