RainBench: Towards Global Precipitation Forecasting from Satellite Imagery

Christian Schroeder de Witt*[1], Catherine Tong*[1],
Valentina Zantedeschi[23], Daniele De Martini[1],
Alfredo Kalaitzis[1], Matthew Chantry[1], Duncan Watson-Parris[1], Piotr Bilinski[41]
[1] University of Oxford, [2] INRIA, [3] UCL, [4] University of Warsaw,
{cs@robots, eu.tong@cs}.ox.ac.uk

Abstract

Extreme precipitation events, such as violent rainfall and hail storms, routinely ravage economies and livelihoods around the developing world. Climate change further aggravates this issue (Gupta et al. 2020). Data-driven deep learning approaches could widen the access to accurate multi-day forecasts, to mitigate against such events. However, there is currently no benchmark dataset dedicated to the study of global precipitation forecasts. In this paper, we introduce **RainBench**, a new multi-modal benchmark dataset for data-driven precipitation forecasting. It includes simulated satellite data, a selection of relevant meteorological data from the ERA5 reanalysis product, and IMERG precipitation data. We also release **PyRain**, a library to process large precipitation datasets efficiently. We present an extensive analysis of our novel dataset and establish baseline results for two benchmark medium-range precipitation forecasting tasks. Finally, we discuss existing data-driven weather forecasting methodologies and suggest future research avenues.

Introduction

Extreme precipitation events, such as violent rain and hail storms, can devastate crop fields and disrupt harvests (Vogel et al. 2019; Li et al. 2019). These events can be locally forecasted with sophisticated numerical weather models that rely on extensive ground and satellite observations. However, such approaches require access to compute and data resources that developing countries in need - particularly in South America and West Africa - cannot afford (Le Coz and van de Giesen 2020; Gubler et al. 2020). The lack of advance planning for precipitation events impedes socioeconomic development and ultimately affects the livelihoods of millions around the world. Given the increase in global precipitation and extreme precipitation events driven by climate change (Gupta et al. 2020), the need for accurate precipitation forecasts is ever more pressing.

Data-driven machine learning approaches circumvent the dependence on traditional resource-intensive numerical models, which typically take several hours to run (Sønderby et al. 2020), incurring a significant time lag. In contrast, deep learning models deployed on dedicated high-throughput hardware can produce inferences in a matter of seconds. However, while there have been attempts in forecasting precipitation with neural networks, they have mostly been fragmented across different local regions, which hinders a systematic comparison into their performance.

In this work, we introduce **RainBench**, a multi-modal dataset to support data-driven forecasting of global precipitation from satellite imagery. We curate three types of datasets: simulated satellite data (SimSat), numerical reanalysis data (ERA5), and global precipitation estimates (IMERG). The use of satellite images to forecast precipitation globally would circumvent the need to collect ground station data, and hence they are key to our vision for widening the access to multi-day precipitation forecasts. Reanalysis data provide estimates of complete atmospheric state, and IMERG provides rigorous estimates of global precipitation. Access to these data opens up opportunities to develop more timely and potentially physics-informed forecast models, which so far could not have been studied systematically.

Most related to our work, Rasp et al. (2020) have developed WeatherBench, a benchmark environment for global data-driven medium-range weather forecasting. This dataset forms an excellent first step in weather forecasting. However, some important features of WeatherBench limit its use for end-to-end precipitation forecasts. WeatherBench does not include any observational raw data (e.g. satellite data) and only contains ERA5 reanalysis data, which have limited resolution of extreme precipitation events. Further, WeatherBench does not include a fast dataloading pipeline to train ML models, which we found to be a significant bottleneck in our model development and testing process. This gap prompted us to also release **PyRain**, a data processing and experimentation framework with fast and configurable multi-modal dataloaders.

To summarise our contributions: (a) We introduce the multi-modal **RainBench** dataset which supports data-driven investigations for global precipitation forecasting from satellite imagery; (b) we release **PyRain**, which allows researchers to run Deep Learning (DL) experiments on RainBench efficiently, reducing time and hardware costs and thus lowering the barrier to entry into this field; (c) we introduce two benchmark precipitation forecasting tasks on RainBench and their baseline results, and present experiments

*: Equal contribution

studying class-balancing schemes. Finally, we discuss the challenges in the field and outline several fruitful avenues for future research.

Related Work

Weather forecasting systems have not fundamentally changed since they were first operationalised nearly 50 years ago. Current state-of-the-art operational weather forecasting systems rely on numerical models that forward the physical atmospheric state in time based on a system of physical equations and parameterised subgrid processes (Bauer, Thorpe, and Brunet 2015). While global simulations typically run at grid sizes of $10\,\mathrm{km}$, regional models can reach $1.5\,\mathrm{km}$ (Franch et al. 2020) . Even in the latter case, skilled forecast lengths are usually limited to a maximum of 10 days, with a conjectured hard limit of 14 to 15 days (Zhang et al. 2019). *Nowcasting*, i.e. high-resolution weather forecasting only a few hours in advance, is currently limited by the several hours that numerical forecasting models take to run (Sønderby et al. 2020).

Given the huge amounts of data currently available from both numerical models and observations, new opportunities exist to train data-driven models to produce these forecasts. The current boom in Machine Learning (ML) has inspired several other groups to approach the problem of weather forecasting. Early work by Xingjian et al. have invested using convolutional recurrent neural networks for precipitation nowcasting. More recently, Sønderby et al. from Google proposed a "(weather) model free" approach, MetNet, which seeks to forecast precipitation in continental USA using geostationary satellite images and radar measurements as inputs. This approach performs well up to 7-8 hours, but inevitably runs into a forecast horizon limit as information from global or surrounding geographic areas is not incorporated into the system. This time window has value though it would not enable substantial disaster preparedness.

The prediction of extreme precipitation (and other extreme weather events) has a long history with traditional forecasting systems (Lalaurette 2003). More recent developments in ensemble weather forecasting systems surround the introduction of novel forecasting indices (Zsótér 2006, EFI) and post-processing (Grönquist et al. 2021). There has also been other deep-learning based precipitation forecasting models as motivated by the monsoon prediction problem, for example, Saha, Mitra, and Nanjundiah (2017) and Saha et al. (2020) use a stacked autoencoder to identify climatic predictors and an ensemble regression tree model, while Praveen et al. (2020) use kriging and multi-layer perceptrons to predict monsoon rainfall from ERA5 data.

WeatherBench (Rasp et al. 2020) is a benchmark dataset for data-driven global weather forecasting, derived from data in the ERA5 archive. Its release has prompted a number of follow-up works to employ deep learning techniques for weather forecasting, although the variables considered have only been restricted to the forecasts of relatively static variables, such as 500 hPa geopotential and 850 hPa temperature (Weyn, Durran, and Caruana 2019, 2020; Rasp and Thuerey 2020; Bihlo 2020; Arcomano et al. 2020). Unlike RainBench which incorporates the element of observational input data from (simulated) satellites, WeatherBench's data comes solely from the ERA5 reanalysis archive, and thus provides no route to producing an end-to-end forecasting system.

RainBench

In this section, we introduce RainBench, which consists of data derived from three publicly-available sources: (1) European Centre for Medium-Range Weather Forecasts (ECMWF) simulated satellite data (SimSat), (2) the ERA5 reanalysis product, and (3) Integrated Multi-satellitE Retrievals (IMERG) global precipitation estimates.

SimSat We use simulated satellite data in place of real satellite imagery to minimise data processing requirements and to simplify the prediction task. SimSat data are model-simulated satellite data generated from ECMWF's high-resolution weather-forecasting model using the RTTOV radiative transfer model (Saunders et al. 2018). SimSat emulates three spectral channels from the Meteosat-10 SEVIRI satellite (Aminou 2002). SimSat provides information about global cloud cover and moisture features and has a native spatial resolution of about 0.1° – i.e. about $10\,\mathrm{km}$ – at three-hourly intervals. The product is available from April 2016 to present (with a lag time of $24\,\mathrm{h}$). Using simulated satellite data provides an intermediate step to using real satellite observations as the images are a global nadir view of Earth, avoiding issues of instrument error and large numbers of missing values. Here we aggregate the data to 0.25° – about $30\,\mathrm{km}$ – to be consistent with the ERA5 dataset.

ERA5 We use ERA5 as it is an accurate and commonly used reanalysis product familiar to the climate science community (Rasp et al. 2020). ERA5 reanalysis data provides hourly estimates of a variety of atmospheric, land and oceanic variables, such as specific humidity, temperature and geopotential height at different pressure levels (Hersbach et al. 2020). Estimates cover the full globe at a spatial resolution of 0.25° and are available from 1979 to present, with a lag time of five days.

IMERG IMERG is a global half-hourly precipitation estimation product provided by NASA (Huffman et al. 2019). We use the Final Run product which primarily uses satellite data from multiple polar-orbiting and geo-stationary satellites. This estimate is then corrected using data from reanalysis products (MERRA2, ERA5) and rain-gauge data. IMERG is produced at a spatial resolution of 0.1° – about $10\,\mathrm{km}$ – and is available from June 2000 to present, with a lag time of about three to four months.

To facilitate efficient experimentation, all data is converted from thier original resolutions to 5.625° resolutions using bilinear interpolation.

RainBench provides precipitation values from two sources, ERA5 and IMERG, as both are widely used and considered to be high-quality precipitation datasets. The ERA5 precipitation is accumulated precipitation over the last hour and is calculated as an averaged quantity over a grid-box. We aggregated IMERG precipitation into hourly

accumulated precipitation and should be considered as a point estimate of the precipitation.

Figure 1 shows the distribution of precipitation for the years 2000-2017 with both ERA5 and IMERG. IMERG is generally regarded as a more trust-worthy dataset for precipitation due to the direct inclusion of precipitation observations in the data assimilation process and the higher spatial resolution used to produce the dataset, which also result in seen difference in data distributions. IMERG has significantly larger rainfall tails than ERA5, and these tails rapidly vanish with decreasing dataset resolution. The underestimation of extreme precipitation events in ERA5 is clearly visible.

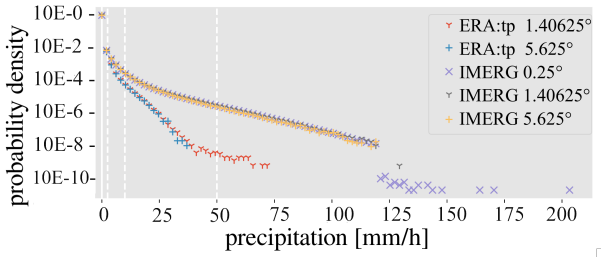

Figure 1: Precipitation histogram from 2000-2017 with ERA5 and IMERG at different resolutions. Vertical lines delineate convection rainfall types: slight (0–2 $\mathrm{mm\,h^{-1}}$), moderate (2–10 $\mathrm{mm\,h^{-1}}$), heavy (10–50 $\mathrm{mm\,h^{-1}}$), and violent (over 50 $\mathrm{mm\,h^{-1}}$) (MetOffice 2012).

PyRain

To support efficient data-handling and experimentation on Rainbench, we release PyRain, an out-of-the-box experimentation framework.

PyRain[2] introduces an efficient dataloading pipeline for complex sample access patterns that scales to the terabytes of spatial timeseries data typically encountered in the climate and weather domain. Previously identified as a decisive bottleneck by the Pangeo community[3], PyRain overcomes existing dataloading performance limitations through an efficient use of NumPy *memmap* arrays[4] in conjunction with optimised software-side access patterns.

In contrast to storage formats requiring *read* system calls, including HDF5[5], Zarr[6] or xarray[7], memory-mapped files use the *mmap* system call to map physical disk space directly to virtual process memory, enabling the use of *lazy* OS demand paging and circumventing the kernel buffer. While less beneficial for chunked or sequential reads and spatial slicing, memmaps can efficiently handle the fragmented random access inherent to the randomized sliding-window access patterns along the primary axis as required in model training.

[2]https://github.com/frontierdevelopmentlab/pyrain
[3]https://pangeo.io/index.html (2021)
[4]https://docs.python.org/3/library/mmap.html (2021)
[5]https://portal.hdfgroup.org/display/HDF5/HDF5(2021)
[6]https://zarr.readthedocs.io/en/stable/ (2021)
[7]http://xarray.pydata.org/en/stable/ (2021)

	NetCDF	PyRain	Speedup
16 workers	40	2410	60.3×
64 workers	70	1930	27.6×

Table 1: Number of data samples loaded per second using PyRain versus a conventional NetCDF framework. Typical configurations assumed and performed on a NVIDIA DGX1 server with 64 CPUs.

In Table 1, we compare PyRain's memmap data reading capcity against a NetCDF+Dask [8] (Rocklin 2015) dataloader. We find empirically that PyRain's memmap dataloader offers significant speedups over other solutions, saturating even SSD I/O with few process workers when used with PyTorch's (Paszke et al. 2019) inbuilt dataloader.

Note that explicitly storing each training sample is not only slow and inflexible for research settings, but it also requires twenty to fifty times more storage and as a result comes at a higher cost than constructing samples on-the-fly. Thus, other options such as writing samples in TFRecord format (Weyn, Durran, and Caruana 2019; Abadi et al. 2016) would only be sensible for highly distributed training in production settings.

PyRain's dataloader is easily configurable and supports both complex multimodal item compositions, as well as periodic (Sønderby et al. 2020) and sequential (Weyn, Durran, and Caruana 2020) train-test set partitionings. Apart from its data-loading pipeline, PyRain also supplies flexible raw-data conversion tools, a convenient interface for data-analysis tasks, various data-normalisation methods and a number of ready-built training settings based on PyTorch Lightning[9]. While being optimised for use with RainBench, PyRain is also compatible with WeatherBench.

Evaluation Tasks

We define two benchmark tasks on RainBench for precipitation forecasting, with the ground truth precipitation values taken from either ERA5 or IMERG.

For each benchmark task, we consider three different input data settings: SimSat, reanalysis data (ERA5), or both. From the ERA5 dataset, we select a subset of variables as input to the forecast model based on our data analysis results; the inputs are geopotential (z), temperature (t), humidity (q), cloud liquid water content (clwc), cloud ice water content (ciwc), each sampled at 300 hPa, 500 hPa and 850 hPa geopotential heights; to these we add the surface pressure and the 2-meter temperature (t2m), as well as static variables that describe the location and surface of the Earth, i.e. latitude, longitude, land-sea mask, orography and soil type. From the SimSat dataset, the inputs are cloud-brightness temperature (clbt) taken at three wavelengths. We normalize each variable with its global mean and standard deviation.

Since data from each source are available at different times, we use the data subset from April 2016 to train all

[8]https://www.unidata.ucar.edu/software/netcdf/ (2021)
[9]https://pytorch-lightning.readthedocs.io/en/latest/ (2021)

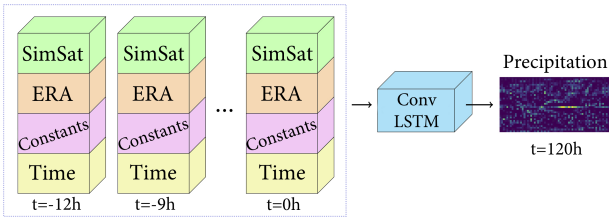

Figure 2: Model setup for the benchmark forecasting tasks.

models for the benchmark tasks, unless specified otherwise. We use data from 2018 and 2019 as validation and test sets respectively. To make sure no overlap exists between training and evaluation data, the first evaluated date is 6 January 2019 while the last training date is 31 December 2017.

We perform experiments with a neural network based on Convolutional LSTMs, which have been shown to be effective for regional precipitation nowcasting (Xingjian et al. 2015). We structure our forecasting task based on MetNet's configurations (Sønderby et al. 2020), where a single model is trained conditioned on time and is capable of forecasting at different lead times.

The network's input is composed of a time series $\{x_t\}$, where each x_t is the set of standardized features at time t, sampled in regular intervals Δt from $t = -T$ to $t = 0$; the output is a precipitation forecast y at lead time $t = \tau \leq \tau_L$. In addition to the aforementioned atmospheric features, static features (e.g. latitude) along with three time-dependant features (hour, day, month) are repeated per timestep. The input vector is then concatenated with a lead-time one-hot vector x_τ. In our experiments, we adopt $T = 12\,\mathrm{h}$, $\Delta t = 3\,\mathrm{h}$ and forecasts at 24-hour intervals up to $\tau_L = 120\,\mathrm{h}$. We note that we do not include precipitation as an input temporal feature. An overview of our setup is shown in fig:approach.

We approach the tasks as a regression problem. Following (Rasp et al. 2020), we use the mean latitude-weighted Root-Mean-Square Error (RMSE) as loss and evaluation metric. We compare the results to persistence and climatology baselines. For persistence, precipitation values at $t = 0$ are used as prediction at $t = \tau$. We compute climatology and weekly climatology baselines from the full training dataset (since 1979 for ERA5 and since 2000 for IMERG), where local climatologies are computed as a single mean over all times and per week respectively (Rasp et al. 2020).

Results

In this section, we first present our data analysis of RainBench. We then describe models' performance on the benchmark precipitation forecasting tasks, which highlights the difficulty in forecasting precipitation values on IMERG. Finally, we present an experiment on same-timestep precipitation estimation to investigate class balancing issues.

Data Analysis

To analyse the dependencies between all RainBench variables, we calculate pairwise Spearman's rank correlation indices over latitude band from -60 to $60°$ and date range from April 2016 to December 2019 (see Figure 3). In contrast to Pearson's correlation coefficient, Spearman's correlation coefficient is significant if there is a, potentially non-linear, monotonic relationship between variables, while Pearson's considers only linear correlations. This allows to capture relationships between variables such as between temperature and absolute latitude. Comparing correlations at altitude pressure levels $300\,\mathrm{hPa}$ (about $10\,\mathrm{km}$) and $850\,\mathrm{hPa}$ ($1.5\,\mathrm{km}$), we can see that they are almost identical, save for a few exceptions: Specific humidity, q, and geopotential height, z, correlate strongly at $300\,\mathrm{hPa}$ but not at $850\,\mathrm{hPa}$, cloud ice water content, ciwc, generally correlates more strongly at higher altitude (and cloud liquid water content, clwc, vice versa). A careful examination of the underlying physical dependencies results in the realisation that all of these asymmetries stem mostly from latitudinal correlations or effects related to cloud formation, e.g. ice and liquid form in clouds at different temperatures/altitudes.

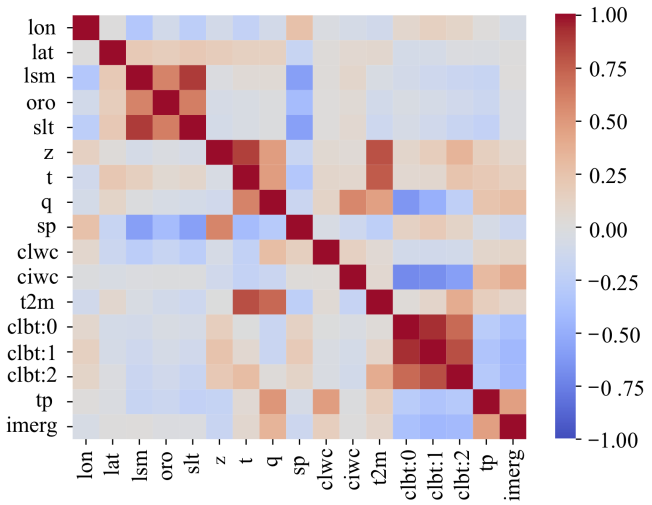

Figure 3: Spearman's correlation of RainBench variables from April 2016 to December 2019 in latitude band $[-60°, 60°]$ at pressure levels $300\,\mathrm{hPa}$ (about $10\,\mathrm{km}$) (upper triangle) and $850\,\mathrm{hPa}$ ($1.5\,\mathrm{km}$) (lower triangle). Legend: lon: longitude, lat: latitude, lsm: land-sea mask, oro: orography (topographic relief of mountains), lst: soil type, z: geopotential height, t: temperature, q: specific humidity, sp: surface pressure, clwc: cloud liquid water content, ciwc: cloud ice water content, $t2m$: temperature at 2m, clbt:i: ith SimSat channel, tp: ERA5 total precipitation, imerg: IMERG precipitation. All correlations in this plot are statistically significant ($p < 0.05$).

As we are particularly interested in variables that have predictive skill on precipitation, we note that all SimSat spectral channels moderately anti-correlate with both ERA5 and IMERG precipitation estimates. Interestingly, SimSat signals correlate much more strongly with specific humidity and cloud ice water content at higher altitude, which might be a consequence of spectral penetration depth. ERA5 state variables that correlate the most with either precipitation estimates are specific humidity and temperature. Cloud ice water content correlates moderately strongly with precipitation estimates at high altitude, but not at all at lower

Inputs	1-day	3-day	5-day
Persistence	0.6249	0.6460	0.6492
Climatology	0.4492 (1979-2017)		
Climatology (weekly)	**0.4447** (1979-2017)		
SimSat	0.4610	0.4678	0.4691
ERA	0.4562	0.4655	0.4677
SimSat + ERA	**0.4557**	**0.4655**	**0.4675**
ERA (1979-2017)	0.4485	0.4670	0.4699

Table 2: Predicting Precipitation from ERA

Inputs	1-day	3-day	5-day
Persistence	1.1321	1.1497	1.1518
Climatology	0.7696 (2000-2017)		
Climatology (weekly)	**0.7687** (2000-2017)		
SimSat	0.8166	0.8201	0.8198
ERA	0.8182	0.8224	0.8215
SimSat + ERA	**0.8134**	**0.8185**	**0.8185**
ERA (2000-2017)	0.8085	0.8194	0.8214

(a) Predicting Precipitation from IMERG

Table 3: Precipitation forecasts evaluated with Latitude-weighted RMSE (mm). All rows except where otherwise stated show models trained with data from 2016 onwards.

altitudes (where ice water content tends to be much lower). Further, a number of time-varying ERA5 state variables correlate more strongly with IMERG precipitation than ERA5 precipitation, as do SimSat signals. Conversely, a number of constant variables, such as land-sea mask, orography and soil type are significantly anti-correlated with ERA5 precipitation, but not at all correlated with IMERG. Overall, we find that all variables that are significantly correlated or anti-correlated with both ERA5 tp and IMERG are also correlated or anti-correlated with SimSat clbt:0-2, suggesting that precipitation prediction from simulated satellite data alone may be feasible.

Precipitation Forecasting

Table 3 compares the neural model forecasts in different data settings when predicting precipitation from ERA5 and IMERG. Using the ERA5 precipitation as target, Table 2 shows that training from SimSat alone gives the worst results across the data settings. This confirms the difficulty in precipitation forecast from satellite data alone, which does not contain as much information about the atmospheric state as sophisticated reanalysis data such as ERA5. Importantly, the complementary benefits of utilizing data from both sources is already visible despite our simple concatenation setup, as training from both SimSat and ERA5 achieves the best results across all lead times (when holding the number of training instances constant).

Figure 4 shows example forecasts from one random input sequence across the different data settings for predicting ERA5 precipitation. We observe that the forecasts can capture the general precipitation distribution across the globe, but there is various degrees of blurriness in the outputs. As we shall discuss later in the paper, considering probabilistic forecasts would be a promising solution to blurriness, which might have arisen as the mean predicted outcome.

We also see the importance in using a large training dataset, since extending the considered training instances to the full ERA5 dataset outperforms the baselines further in the 1-day forecasting regime (shown in the last rows).

Table 3a shows the forecast results when predicting IMERG precipitation. As before, the neural model's forecasting skill based on both SimSat and ERA input outperforms the other input settings. The higher observed RMSEs suggest that this is a considerably more difficult task, which we believe to be closely tied to IMERG featuring more extreme precipitation events (Figure 1). In the next section, we investigate this issue further by considering a same-timestep precipitation estimation task.

A key limitation in our current experimental setup is that it requires all of ERA5, IMERG and SimSat channels to be available at each time step, limiting the range of our training data to April 2016 and onward. Nevertheless, our neural models significantly outperform persistence baselines. The fact that local climatology trained over longer time periods significantly outperforms our network model baselines suggests the development of alternative modelling setups that can make use of the full available datasets from each source.

Same-Timestep Precipitation Estimation

We now describe a set of experiments for same-timestep precipitation estimation on IMERG. This analysis is done independently from the precipitation forecasting benchmark tasks, in order to provide an in-depth understanding of the challenges in modelling extreme precipitation events.

We use a gradient boosting decision tree learning algorithm (Ke et al. 2017, LightGBM) in order to estimate same-timestep IMERG precipitation directly from ERA5 and SimSat. Our training set consists of 1 million randomly sampled grid points/pixels within the time interval April 2016 to December 2019. We compare the (not latitude-adjusted) RMSE for two pixel sampling variants: A) unbalanced sampling, meaning grid points are chosen randomly from the raw data distribution and B) balanced sampling, in which we bin IMERG precipitation into the four classes defined in Figure 1 and sample grid points such that we end up with an equal amount of pixels per bin.

In Figure 4, we find that taking a balanced sampling approach reduces the per-class validation RMSE of moderate, heavy and violent precipitation. This balanced sampling approach also has detrimental effects on the mean forecasting performance but not the macro-mean performance, as the 'slight' class dominates the dataset and is misclassified more often. However, balancing the training set does result in a lower macro RMSE.

Designing an appropriate class-balanced sampling may play a crucial role toward improving predictions of extreme

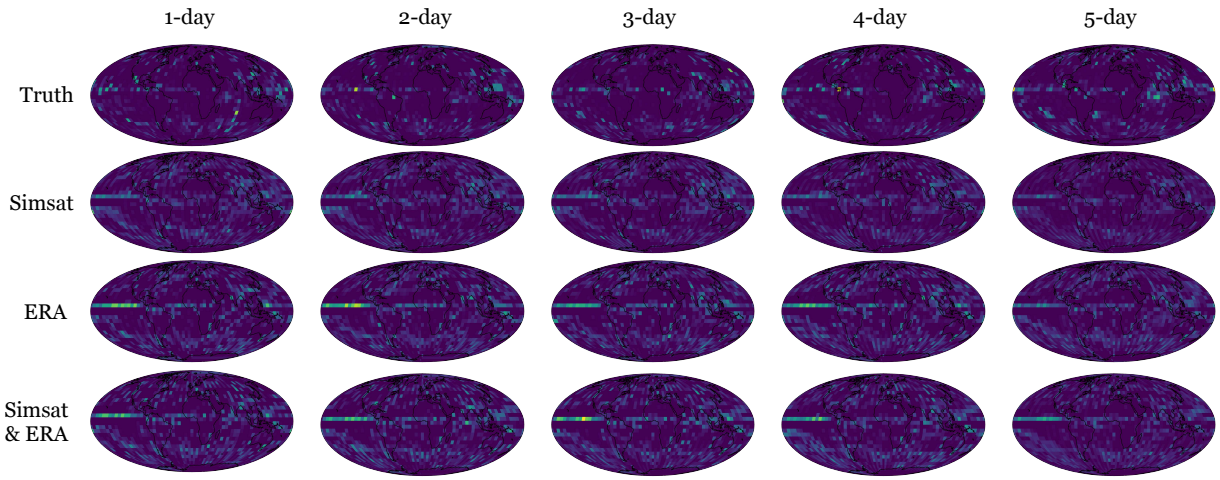

Figure 4: ERA5 Precipitation forecasts on one random sample.

	L	M	H	V	Mean	Macro
Unbalanced						
ERA	**0.20**	4.08	16.2	63.1	0.65	20.9
SimSat	**0.20**	4.38	16.8	54.1	0.65	18.9
SimSat + ERA	**0.20**	4.03	16.5	53.0	**0.65**	18.4
Balanced						
ERA	1.05	**2.75**	12.4	58.0	1.40	18.6
SimSat	1.17	3.10	13.3	50.1	1.26	16.9
SimSat + ERA	1.30	3.15	**11.8**	**44.3**	1.38	**15.1**

Table 4: Comparing RMSE Results with and without a class-balanced training dataset. The modelling task is same-timestep estimation of IMERG precipitation.

precipitation events. It is not quite clear how a per-pixel sampling scheme may be translated into a global output context approach such as in MetNet (Sønderby et al. 2020) where each individual pixel's input distribution should be kept balanced, while training as many pixels per input data sample as possible for efficiency. A possible way of navigating this challenge would be to sample greedily, i.e. based on the currently most imbalanced pixel and combine this with learning rate adjustments for other pixels trained on the same frame based on how imbalanced these pixels are at that timestep.

Discussion

We outline the key challenges in global precipitation forecasting, our proposed solutions, we also discuss promising research avenues that can build on our work.

Challenges

From our experiments, we identified a number of challenges inherent to data-driven extreme precipitation forecasting.

Class imbalance Extreme precipitation events, by their nature, rarely occur (see Figure 1). In the context of supervised learning, this manifests as a class imbalance problem, in which a model might rarely predict extreme values. Designing an appropriate class sampling strategy (e.g. inverse frequency sampling) can mitigate this imbalance, as shown in our same-timestep prediction experiments. Further, we believe that a mixture of pixelwise-weighting and balanced sampling could be a potential solution.

Probabilistic forecasts. The current machine learning setup produces deterministic predictions, which may lead to an averaging of possible futures into a single blurry prediction. This limitation may be overcome with probabilistic modelling, which may take different forms. For instance, Sønderby et al. made use of a cross-entropy loss over a categorical distribution to handle probabilistic forecasts. Stochastic video prediction techniques (Babaeizadeh et al. 2018) and conditional generative adversarial learning (Mirza and Osindero 2014) have also been shown to produce realistic predictions in other fields. Other relevant techniques that predict distribution parameters are Variational Auto-Encoders (Kingma and Welling 2014) and normalizing flows (Rezende and Mohamed 2015).

Data normalisation. Feature scaling is a common data-processing step for training machine learning models and well-understood to be advantageous (Bhanja and Das 2019). Our current approach normalizes each variable using its global mean and standard deviation; This disregards any local spatial differences, which is important for modelling local weather patterns (Weyn, Durran, and Caruana 2019). Previous work suggested that patch-wise normalisation may be appropriate (Grönquist et al. 2021, Local Area-wise Standardization (LAS)). We suggest studying a refinement to LAS, which adjusts the kernel size with latitude such that the spatial normalisation context remains constant (*Latitude-Adjusted LAS*) per-channel image-size normalisation.

Data topology. Lastly, the spherical input and output data topology of global forecasting contexts poses interesting questions to neural network architecture. While a multitude of approaches to handle spherical input topologies have been suggested, see (Llorens Jover 2020) for an overview, it seems yet unclear which approach works best. Our dataset might constitute a valuable benchmark for such research.

Future Research Avenues

Apart from overcoming the challenges outlined above, we have identified a variety of opportunities for further research.

Physics-informed multi-task learning. Apart from using reanalysis data for model training, we do not currently exploit the fact that many aspects of weather forecasting are well-understood from a physical perspective. One way of informing model training of physical constraints would be to train precipitation forecasting concurrently with prediction of physical state variables, including temperature and specific humidity, in a multi-task setting, e.g. through using separate decoder heads for different variables (similarly to Caruana (1997)). This approach promises to combine the advantages of data-driven learning with low-level feature regularisation through a physics-informed inductive bias. Multi-task learning can also be regarded as a form of data augmentation (Shorten and Khoshgoftaar 2019), promising to further increase forecasting performance using real or simulated satellite data without requiring access to reanalysis data at inference time.

Increasing spatial resolution. Data at higher spatial resolution tends to capture heavy and extreme precipitation events better but poses a number of challenges. Large sample batch sizes may lead to network activation storage that exceeds GPU global memory capacity even for distributed training. Apart from exploring TPU or nvlink-based solutions, another way would be to switch to mixed-precision or half-precision or employ techniques that trade-off memory for compute such as gradient checkpointing (Pinckaers, van Ginneken, and Litjens 2019). PyRain's dataloader efficiently maximises total disk throughput, which may itself become a bottleneck at very high resolutions. Storing all or part of the training data memmaps on one or several high-speed local SSDs may increase disk throughput a few-fold. Apart from memory and disk throughput, there is also a lack of suitably highly resolved historical climate data for pre-training (Rasp et al. 2020). One possible way of overcoming this would be to integrate high-resolution local forecasting model or sensor data into the training process (Franch et al. 2020), another exciting approach spearheaded in computational fluid dynamics (Jabarullah Khan and Elsheikh 2019) is to employ a multi-fidelity approach, where hierarchical variance-reduction techniques are employed to enable training to be performed at lower-resolution data as often as possible, thus minimising the need for training on high-resolution data.

Reducing IMERG Early Run lag time. While the final IMERG product becomes available at a time lag of ca. 3-4 months, a preliminary, Early Run, product based on raw satellite data becomes available after ca. 4 hours. We postulate that this lag could be further reduced if, instead of high-dimensional observational data, forecasting agencies were exchanging their locally processed low-dimensional embeddings derived from local encoder networks. Embeddings could then be feed into a late fusion network architecture similar to Rudner et al. (2019, Multi^{3}Net).

Multi-time-step loss function. Numerical forecasting systems forward the physical state in time by following an iterative setting, where the output of the previous step is fed as input to the next step. As the update rules are identical for each step, it in principle suffices for neural networks to learn a single such update step and apply it multiple times during inference depending on the prediction lead time, thus reducing the number of trainable weights and potentially increase generalisation performance. To avoid instability issues inherent to iterative approaches (Rasp et al. 2020), model rollouts can be trained end-to-end (McGibbon and Bretherton 2019; Brenowitz and Bretherton 2018). Weyn, Durran, and Caruana (2020) pioneer this approach but limit themselves to just two time steps. To overcome device memory constraints in such a setting and to scale to a large number of time steps tollouts, iteration layers could be chosen to be reversible (Gomez et al. 2017) such that activations can be computed on-the-fly during backpropagation and do not need to be stored in device memory.

Conclusion

We presented **RainBench**, a novel benchmark suite for data-driven extreme precipitation forecasting, and **PyRain**, an associated rapid experimentation framework with a fast dataloader. Both RainBench and PyRain are open source and well-documented. We furthermore present neural baselines for multi-day precipitation forecasting from both reanalysis and simulated satellite data. Despite our simple approach, we find that our neural baselines beat climatology and persistence baselines for up to 5 day forecasts. In addition, we use a gradient boosting decision tree algorithm to study the impact of precipitation class balancing on regression in a precipitation estimation setting and present various forms of data exploration, including a correlation study.

In the near future, we will augment RainBench with real satellite data. We plan on also including historical climate data for pre-training. Concurrently, we will explore various directions for future research, as discussed above. In particular, we believe increasing the spatial resolution of our input data is crucial to closing the gap to operational forecasting models. Ultimately, we hope that our benchmark and framework will lower the barrier of entry for the global research community such that our work contributes to rapid progress in data-driven weather prediction, democratisation of access to adequate weather forecasts and, ultimately, help protect and improve livelihoods in a warming world.

Acknowledgements

This research was conducted at the Frontier Development Lab (FDL), Europe. The authors gratefully acknowledge support from the European Space Agency ESRIN Phi Lab, Trillium Technologies, NVIDIA Corporation, Google Cloud, and SCAN. The authors are thankful to Peter Dueben, Stephan Rasp, Julien Brajard and Bertrand Le Saux for useful suggestions.

References

Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mane, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viegas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; and Zheng, X. 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv:1603.04467 [cs]* .

Aminou, D. 2002. MSG's SEVIRI instrument. *ESA Bulletin(0376-4265)* (111): 15–17.

Arcomano, T.; Szunyogh, I.; Pathak, J.; Wikner, A.; Hunt, B. R.; and Ott, E. 2020. A Machine Learning-Based Global Atmospheric Forecast Model. *Geophysical Research Letters* 47(9): e2020GL087776.

Babaeizadeh, M.; Finn, C.; Erhan, D.; Campbell, R. H.; and Levine, S. 2018. Stochastic Variational Video Prediction. *arXiv:1710.11252 [cs]* URL http://arxiv.org/abs/1710.11252. ArXiv: 1710.11252.

Bauer, P.; Thorpe, A.; and Brunet, G. 2015. The quiet revolution of numerical weather prediction. *Nature* 525(7567): 47–55.

Bhanja, S.; and Das, A. 2019. Impact of Data Normalization on Deep Neural Network for Time Series Forecasting. *arXiv:1812.05519 [cs, stat]* URL http://arxiv.org/abs/1812.05519. ArXiv: 1812.05519.

Bihlo, A. 2020. A generative adversarial network approach to (ensemble) weather prediction. *arXiv preprint arXiv:2006.07718* .

Brenowitz, N. D.; and Bretherton, C. S. 2018. Prognostic Validation of a Neural Network Unified Physics Parameterization. *Geophysical Research Letters* 45(12): 6289–6298. ISSN 1944-8007. doi:10.1029/2018GL078510.

Caruana, R. 1997. Multitask Learning. *Machine Learning* 28(1): 41–75. ISSN 1573-0565. doi:10.1023/A:1007379606734. URL https://doi.org/10.1023/A:1007379606734.

Franch, G.; Maggio, V.; Coviello, L.; Pendesini, M.; Jurman, G.; and Furlanello, C. 2020. TAASRAD19, a high-resolution weather radar reflectivity dataset for precipitation nowcasting. *Scientific Data* 7(1): 234. ISSN 2052-4463. doi:10.1038/s41597-020-0574-8. URL https://www.nature.com/articles/s41597-020-0574-8. Number: 1 Publisher: Nature Publishing Group.

Gomez, A. N.; Ren, M.; Urtasun, R.; and Grosse, R. B. 2017. The Reversible Residual Network: Backpropagation Without Storing Activations. *arXiv:1707.04585 [cs]* ArXiv: 1707.04585.

Grönquist, P.; Yao, C.; Ben-Nun, T.; Dryden, N.; Dueben, P.; Li, S.; and Hoefler, T. 2021. Deep learning for post-processing ensemble weather forecasts. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379(2194): 20200092. doi:10.1098/rsta.2020.0092. Publisher: Royal Society.

Gubler, S.; Sedlmeier, K.; Bhend, J.; Avalos, G.; Coelho, C.; Escajadillo, Y.; Jacques-Coper, M.; Martinez, R.; Schwierz, C.; de Skansi, M.; et al. 2020. Assessment of ECMWF SEAS5 seasonal forecast performance over South America. *Weather and Forecasting* 35(2): 561–584.

Gupta, A. K.; Yadav, D.; Gupta, P.; Ranjan, S.; Gupta, V.; and Badhai, S. 2020. Effects of climate change on Agriculture. *Food and Agriculture Spectrum Journal* 1(3).

Hersbach, H.; Bell, B.; Berrisford, P.; Hirahara, S.; Horányi, A.; Muñoz-Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Schepers, D.; et al. 2020. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society* 146(730): 1999–2049.

Huffman, G.; Stocker, E.; Bolvin, D.; Nelkin, E.; and Tan, J. 2019. GPM IMERG Final Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V06. Technical report. doi:10.5067/GPM/IMERG/3B-HH/06.

Jabarullah Khan, N. K.; and Elsheikh, A. H. 2019. A Machine Learning Based Hybrid Multi-Fidelity Multi-Level Monte Carlo Method for Uncertainty Quantification. *Frontiers in Environmental Science* 7. ISSN 2296-665X. Publisher: Frontiers.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, 3146–3154.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]* URL http://arxiv.org/abs/1312.6114. ArXiv: 1312.6114.

Lalaurette, F. 2003. Early detection of abnormal weather conditions using a probabilistic extreme forecast index. *Quarterly Journal of the Royal Meteorological Society* 129(594): 3037–3057. ISSN 1477-870X. doi:10.1256/qj.02.152.

Le Coz, C.; and van de Giesen, N. 2020. Comparison of Rainfall Products over Sub-Saharan Africa. *Journal of Hydrometeorology* 21(4): 553–596.

Li, Y.; Guan, K.; Schnitkey, G. D.; DeLucia, E.; and Peng, B. 2019. Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the United States. *Global Change Biology* 25(7): 2325–2337. ISSN 1365-2486.

Llorens Jover, I. 2020. Geometric deep learning for medium-range weather prediction. URL https://infoscience.

epfl.ch/record/278138. Master's Thesis [Accessed Dec. 22, 2020].

McGibbon, J.; and Bretherton, C. S. 2019. Single-Column Emulation of Reanalysis of the Northeast Pacific Marine Boundary Layer. *Geophysical Research Letters* 46(16): 10053–10060. ISSN 1944-8007. doi:10.1029/2019GL083646.

MetOffice. 2012. Fact sheet 3 — Water in the atmosphere. Technical report, MetOffice UK. URL https://www.metoffice.gov.uk/binaries/content/assets/metofficegovuk/pdf/research/library-and-archive/library/publications/factsheets/factsheet_3-water-in-the-atmosphere.pdf. [Accessed Dec. 22, 2020].

Mirza, M.; and Osindero, S. 2014. Conditional Generative Adversarial Nets. *arXiv:1411.1784 [cs, stat]* URL http://arxiv.org/abs/1411.1784. ArXiv: 1411.1784.

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. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d. Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*, 8024–8035. Curran Associates, Inc.

Pinckaers, H.; van Ginneken, B.; and Litjens, G. 2019. Streaming convolutional neural networks for end-to-end learning with multi-megapixel images. *arXiv:1911.04432 [cs]* URL http://arxiv.org/abs/1911.04432. ArXiv: 1911.04432.

Praveen, B.; Talukdar, S.; Shahfahad; Mahato, S.; Mondal, J.; Sharma, P.; Islam, A. R. M. T.; and Rahman, A. 2020. Analyzing trend and forecasting of rainfall changes in India using non-parametrical and machine learning approaches. *Scientific Reports* 10(1): 10342. ISSN 2045-2322. doi:10.1038/s41598-020-67228-7. Number: 1 Publisher: Nature Publishing Group.

Rasp, S.; Dueben, P. D.; Scher, S.; Weyn, J. A.; Mouatadid, S.; and Thuerey, N. 2020. WeatherBench: A benchmark dataset for data-driven weather forecasting. *arXiv:2002.00469 [physics, stat]* ArXiv: 2002.00469.

Rasp, S.; and Thuerey, N. 2020. Purely data-driven medium-range weather forecasting achieves comparable skill to physical models at similar resolution .

Rezende, D.; and Mohamed, S. 2015. Variational Inference with Normalizing Flows. In *International Conference on Machine Learning*, 1530–1538. PMLR. URL http://proceedings.mlr.press/v37/rezende15.html. ISSN: 1938-7228.

Rocklin, M. 2015. Dask: Parallel Computation with Blocked algorithms and Task Scheduling. In Huff, K.; and Bergstra, J., eds., *Proceedings of the 14th Python in Science Conference*, 130 – 136.

Rudner, T. G. J.; Rußwurm, M.; Fil, J.; Pelich, R.; Bischke, B.; Kopačková, V.; and Biliński, P. 2019. Multi3Net: Segmenting Flooded Buildings via Fusion of Multiresolution, Multisensor, and Multitemporal Satellite Imagery. *Proceedings of the AAAI Conference on Artificial Intelligence* 33(01): 702–709. ISSN 2374-3468. doi:10.1609/aaai.v33i01.3301702. Number: 01.

Saha, M.; Mitra, P.; and Nanjundiah, R. S. 2017. Deep learning for predicting the monsoon over the homogeneous regions of India. *Journal of Earth System Science* 126(4): 54. ISSN 0973-774X. doi:10.1007/s12040-017-0838-7.

Saha, M.; Santara, A.; Mitra, P.; Chakraborty, A.; and Nanjundiah, R. S. 2020. Prediction of the Indian summer monsoon using a stacked autoencoder and ensemble regression model. *International Journal of Forecasting* ISSN 0169-2070. doi:10.1016/j.ijforecast.2020.03.001.

Saunders, R.; Hocking, J.; Turner, E.; Rayer, P.; Rundle, D.; Brunel, P.; Vidot, J.; Roquet, P.; Matricardi, M.; Geer, A.; et al. 2018. An update on the RTTOV fast radiative transfer model (currently at version 12). *Geoscientific Model Development* 11(7).

Shorten, C.; and Khoshgoftaar, T. M. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data* 6(1): 60. ISSN 2196-1115. doi:10.1186/s40537-019-0197-0. URL https://doi.org/10.1186/s40537-019-0197-0.

Sønderby, C. K.; Espeholt, L.; Heek, J.; Dehghani, M.; Oliver, A.; Salimans, T.; Hickey, J.; Agrawal, S.; and Kalchbrenner, N. 2020. MetNet: A Neural Weather Model for Precipitation Forecasting. *Submission to journal* URL https://arxiv.org/abs/2003.12140.

Vogel, E.; Donat, M. G.; Alexander, L. V.; Meinshausen, M.; Ray, D. K.; Karoly, D.; Meinshausen, N.; and Frieler, K. 2019. The effects of climate extremes on global agricultural yields. *Environmental Research Letters* 14(5): 054010.

Weyn, J. A.; Durran, D. R.; and Caruana, R. 2019. Can Machines Learn to Predict Weather? Using Deep Learning to Predict Gridded 500-hPa Geopotential Height From Historical Weather Data. *Journal of Advances in Modeling Earth Systems* 11(8): 2680–2693. ISSN 1942-2466.

Weyn, J. A.; Durran, D. R.; and Caruana, R. 2020. Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *arXiv:2003.11927 [physics, stat]* URL http://arxiv.org/abs/2003.11927. ArXiv: 2003.11927.

Xingjian, S.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; and Woo, W.-c. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, 802–810.

Zhang, F.; Sun, Y. Q.; Magnusson, L.; Buizza, R.; Lin, S.-J.; Chen, J.-H.; and Emanuel, K. 2019. What Is the Predictability Limit of Midlatitude Weather? *Journal of the Atmospheric Sciences* 76(4): 1077–1091. ISSN 0022-4928. doi:10.1175/JAS-D-18-0269.1. Publisher: American Meteorological Society.

Zsótér, E. 2006. Recent developments in extreme weather forecasting. doi:10.21957/kl9821hnc7. Issue: 107 Pages: 8-17 Publisher: ECMWF.