

# Purely Data-Driven Approaches to Weather Prediction: Promise and Perils

**Suman Ravuri, DeepMind**

[ravuris@deepmind.com](mailto:ravuris@deepmind.com)

*On behalf of the DeepMind - Met Office Collaboration Team*





## DeepMind - Met Office Collaboration

- Two research-driven organisations working together for 3 years (since 2018), meeting weekly.
- Bringing together expertise in atmospheric science and AI.
- Wonderful learning experience, we've met 3x a week every week for these past few years.



Image credits: Met Office, Construction Specialties, Suman Ravuri, Ellen Clancy





## Case Study for this Talk

# Skilful precipitation nowcasting using deep generative models of radar

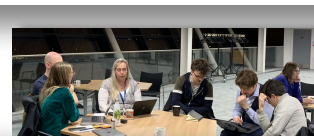
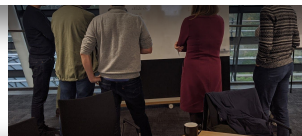
<https://doi.org/10.1038/s41586-021-03854-z>

Received: 17 February 2021

Accepted: 27 July 2021

Published online: 29 September 2021

Suman Ravuri<sup>1,5</sup>, Karel Lenc<sup>1,5</sup>, Matthew Willson<sup>1,5</sup>, Dmitry Kangin<sup>2,3</sup>, Remi Lam<sup>1</sup>, Piotr Mirowski<sup>1</sup>, Megan Fitzsimons<sup>2</sup>, Maria Athanassiadou<sup>2</sup>, Sheleem Kashem<sup>1</sup>, Sam Madge<sup>2</sup>, Rachel Prudden<sup>2,3</sup>, Amol Mandhane<sup>1</sup>, Aidan Clark<sup>1</sup>, Andrew Brock<sup>1</sup>, Karen Simonyan<sup>1</sup>, Raia Hadsell<sup>1</sup>, Niall Robinson<sup>2,3</sup>, Ellen Clancy<sup>1</sup>, Alberto Arribas<sup>2,4</sup> & Shakir Mohamed<sup>1</sup>✉



Disclaimer:

This presentation has a Machine Learning researcher's perspective

1. Problem statement and data
2. Baseline Models...
3. ... or why we need generative models
4. Deep Generative Models of Radar
5. Quantitative verification and its limitations
6. Expert evaluation
7. Conclusion



The background is a solid teal color. On the left side, there is a blue wireframe of a 3D cube. On the right side, there is a thin, light yellow curved line that forms a partial oval shape.

1

# Problem statement and data

# Precipitation Nowcasting

High-resolution (1km x 1km) rainfall estimates in the short term (5–90 minutes)

Used by expert meteorologists to:

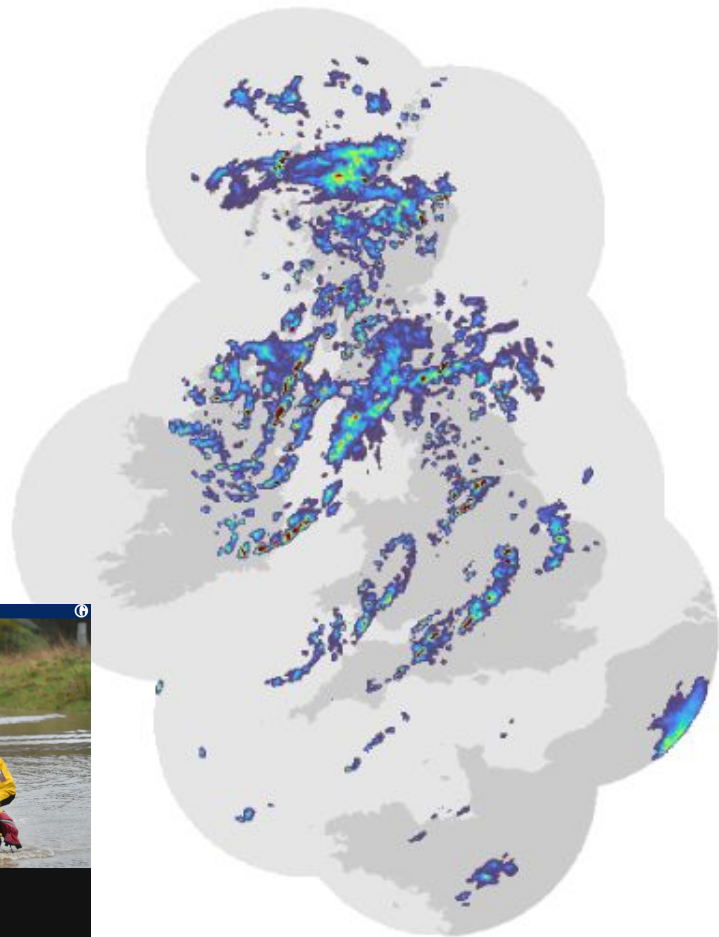
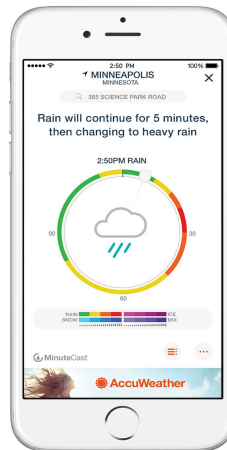
- Issue flood warnings
- Perform air traffic control
- Marine services

Important statistical questions:

- Prediction at **multiple temporal and spatial scales**
- Accounting for **uncertainty**
- Capturing **rare events**

**NWPs perform poorly here**

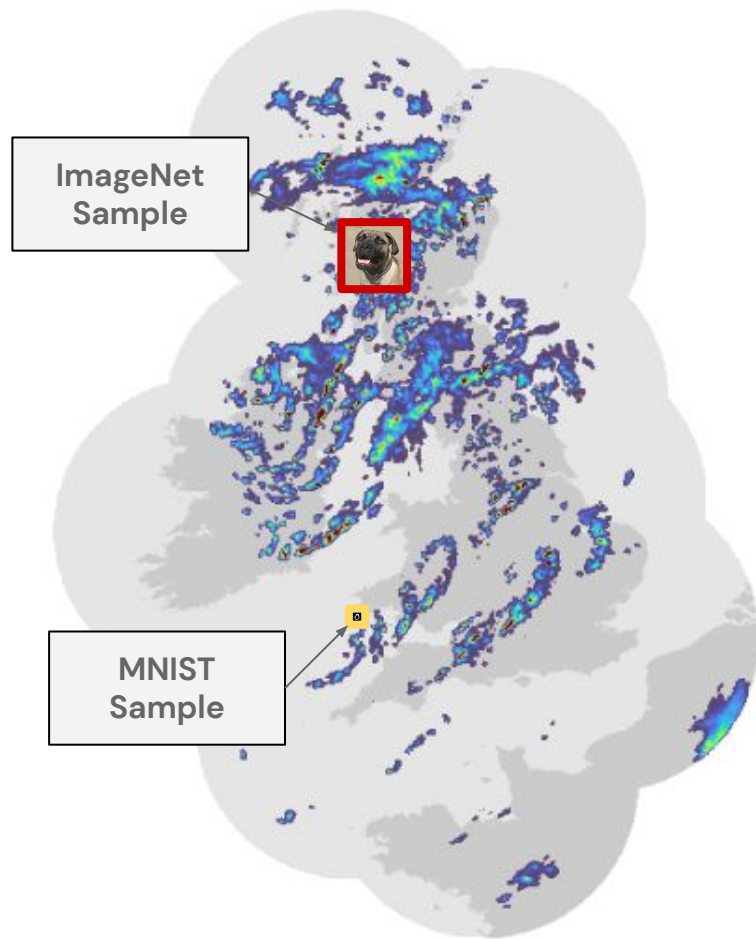
- Data-driven approaches should be strong



# Precipitation Data from the UK

Very large radar fields

- Met Office RadarNet4 Data
- Every 5 minutes, 288/day
- 1536 x 1280 pixels
- 1km x 1km grids
- Data from 2016–
- Data agreement through MO data provisioning team





## Multi-resolution Multi-scale Data (US)

- Data over continental United States from 2017–2019
- $0.01^\circ$  (lat, long) grids of  $3584 \times 7168$
- Asynchronous updating and poor radar on West Coast.



# Additional data sources?

## Numerical Weather Prediction (NWP)

- UKV model
- Work on lead time  $> 3h$
- **Rainflux** used as **baseline**



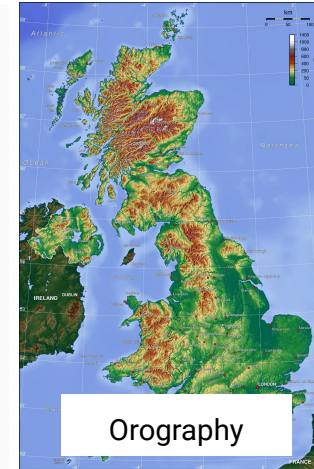
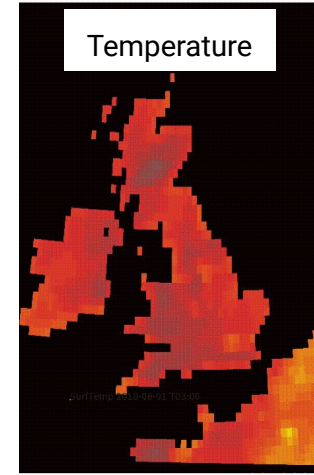
## Rain gauges

- For calibration

## Orography, land cover

- Investigated...
- ... but not used

Did not use satellite data either!





2

# Baseline Models...



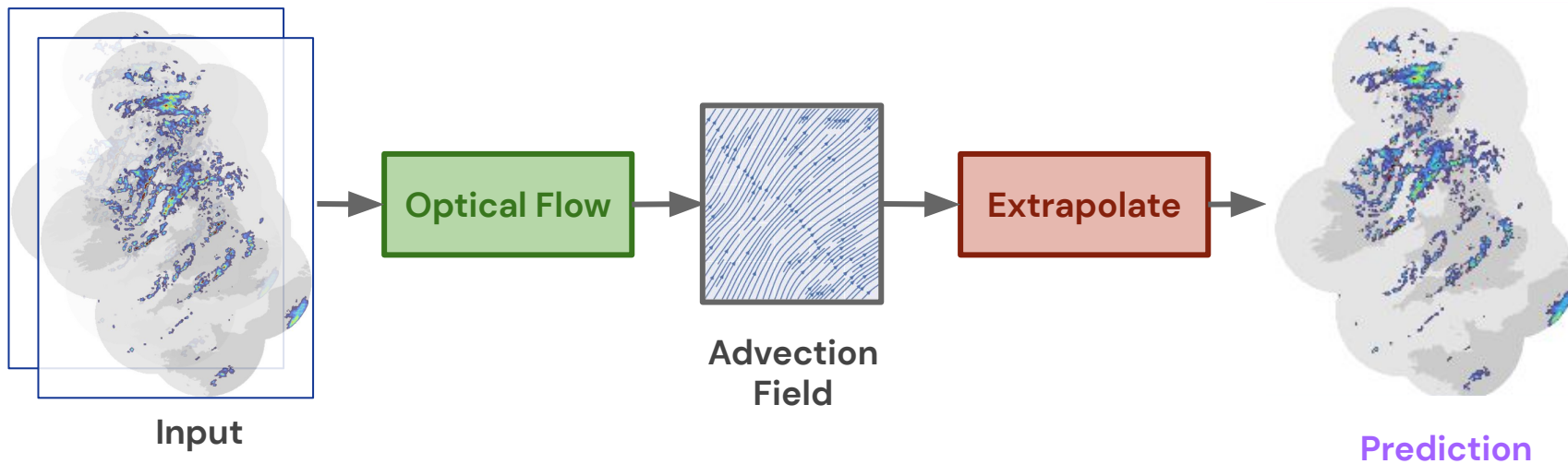


## A Quote

“All models are wrong, but  
some are useful”

- George Box (Statistician)

# Useful Physics-Inspired Baselines: Lagrangian Persistence & PySTEPS



- Stationary optical flow (e.g., Lucas Kanade) (Bowler et al, 2004)
- Future obtained with Semi-Lagrangian extrapolation (Germann et al, 2002)
- **What's wrong:** advection only, struggles with orography

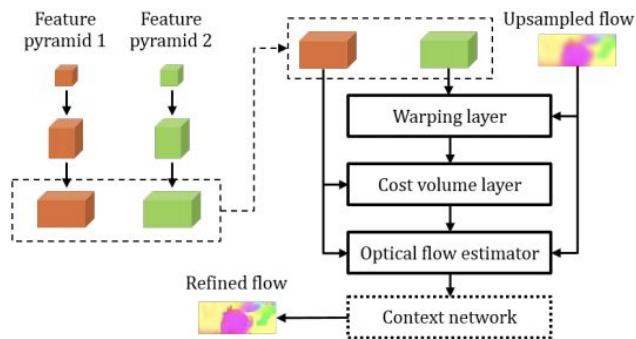
Bowler et al., 2004. Development of a precipitation nowcasting algorithm based upon optical flow techniques. Journal of Hydrology.

Germann et al., 2002. Scale-dependence of the predictability of precipitation from continental radar images. Part I: Description of the methodology. Monthly Weather Review.

# Neuralizing: Extending Optical Flow with PWC-Nets

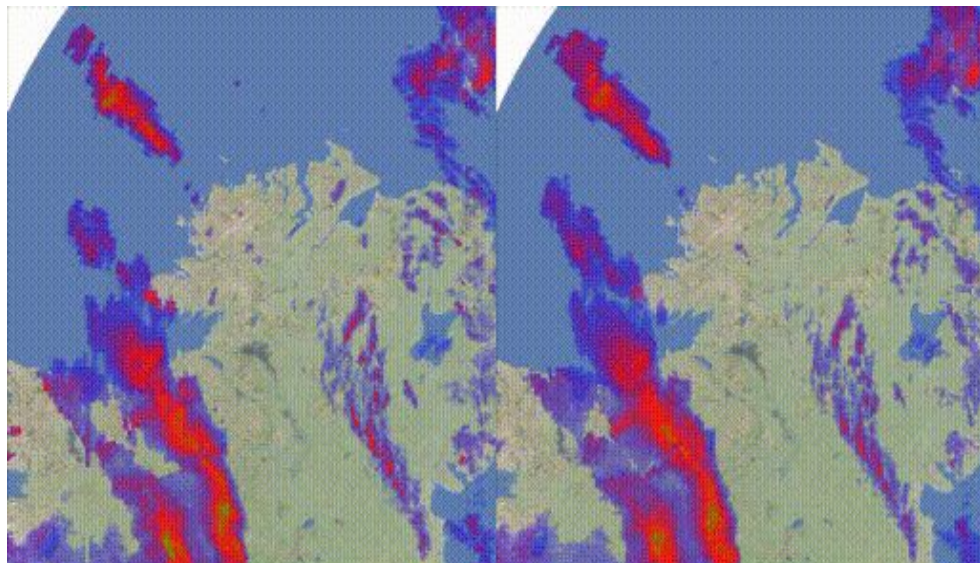
Differentiable approach  
to optical flow estimation  
(Sun et al, 2017)

$$p_{\theta}(x_t|x_{<t}) := f(x_{<t}; \theta)$$



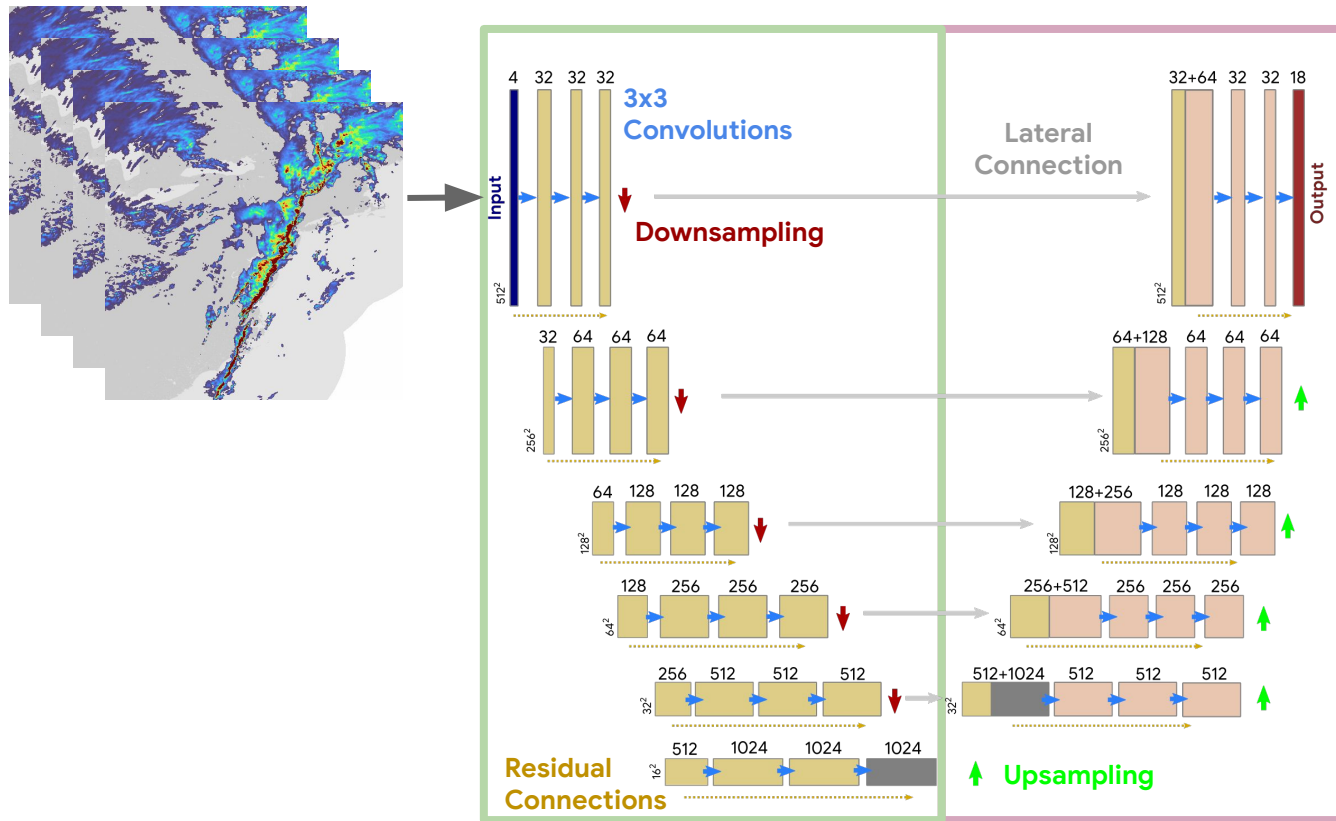
Ground truth

Prediction (PWC-Net)





# A Purely Data-Driven Baseline: UNet (Regression and Classification)



Regression:

$$\sum_{t,h,w} (x_{t,h,w} - \hat{x}_{t,h,w,i})^2$$

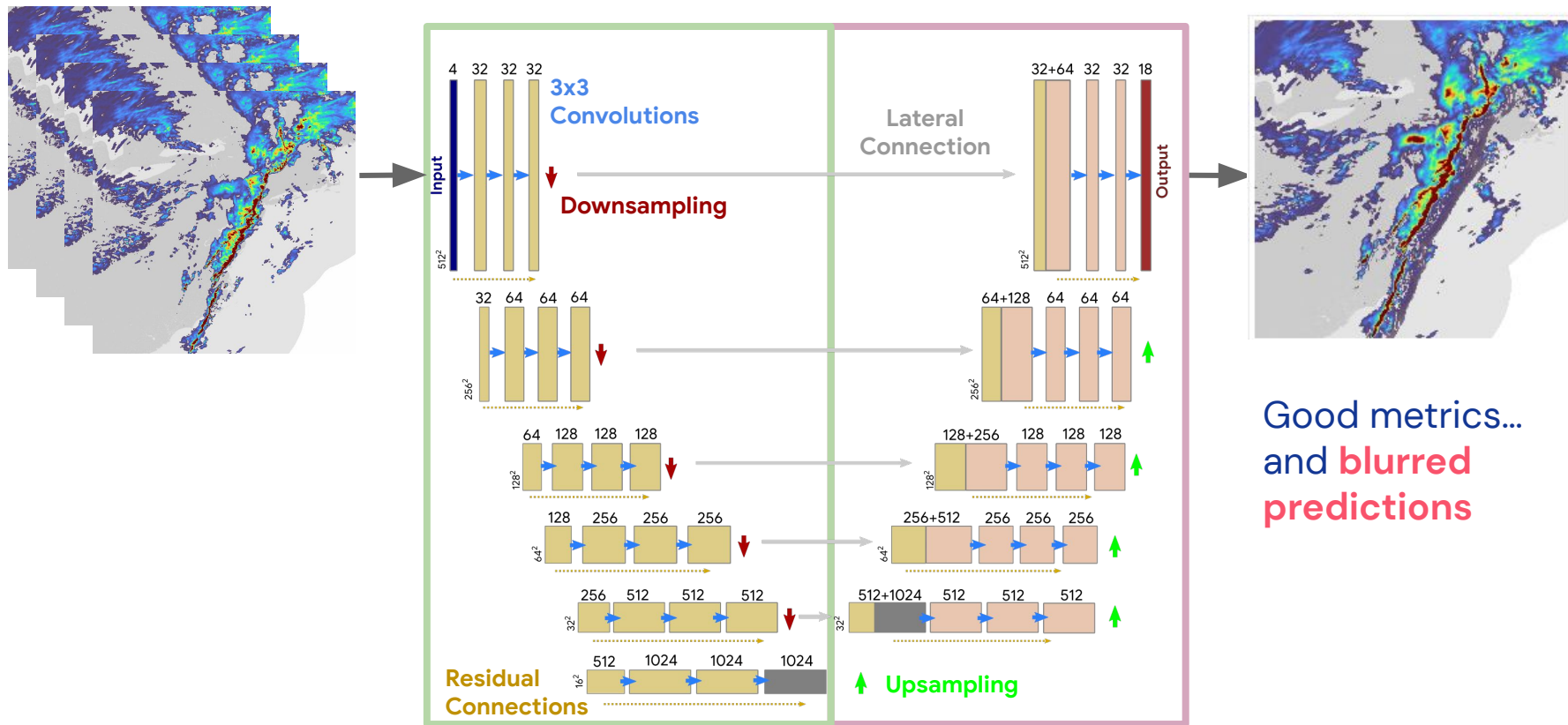
$$\hat{x}_{t,h,w} = \text{UNet}(x_{<t})_{t,h,w}$$

Classification:

$$-\sum_{t,h,w,i} \log p(x_{t,h,w} = i) \log \frac{\exp(\hat{x}_{t,h,w}^{(i)})}{\sum_j \exp(\hat{x}_{t,h,w}^{(j)})}$$

$$\hat{x}_{t,h,w}^{(i)} = \text{UNet}(x_{<t})_{t,h,w}^{(i)}$$

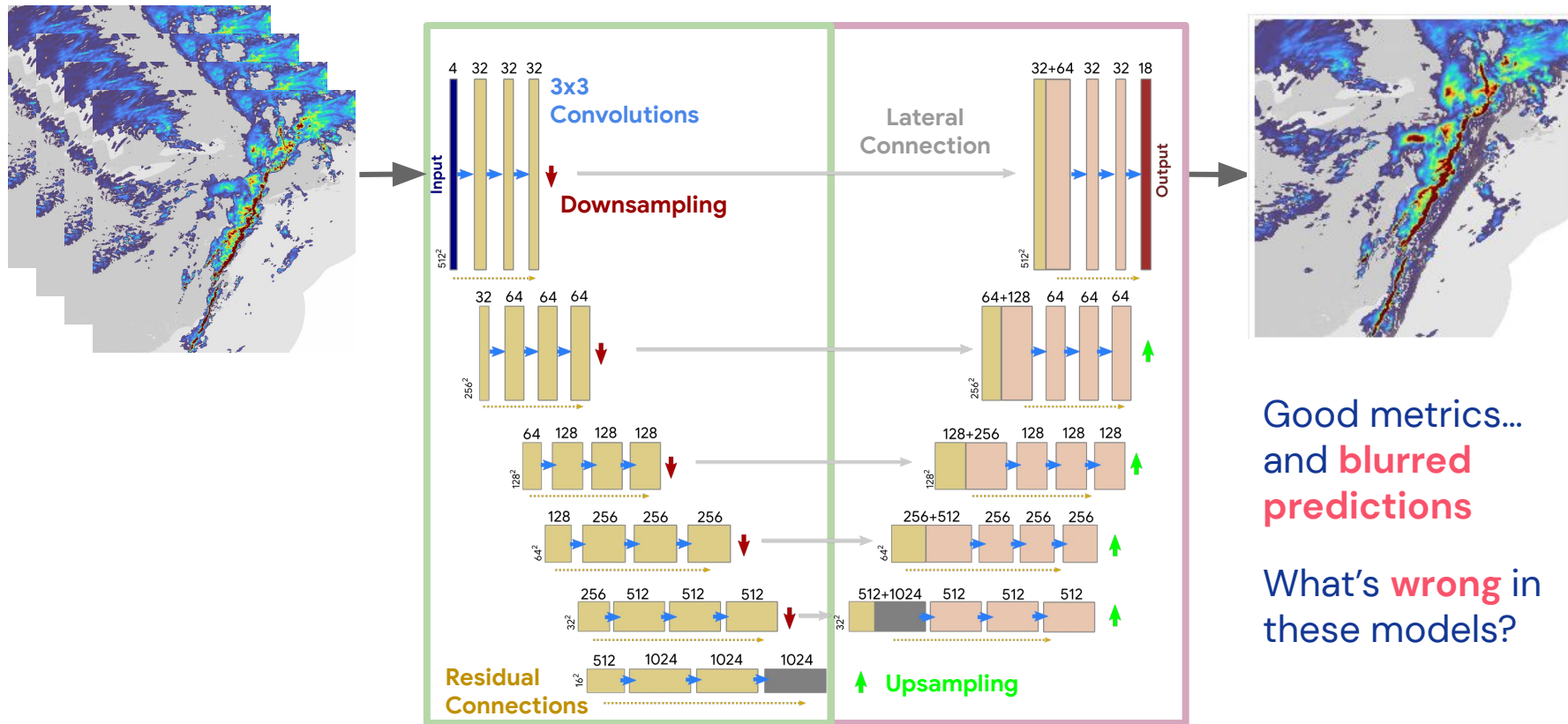
# A Purely Data-Driven Baseline: UNet (Regression and Classification)



Ayzel et al., 2020. RainNet v1.0: a convolutional neural network for radar-based precipitation nowcasting Geosci. Model Dev.

Agrawal et al., 2019, Machine Learning for Precipitation Nowcasting from Radar Images, NeurIPS Climate Change AI workshop.

# A Purely Data-Driven Baseline: UNet (Regression and Classification)



Good metrics...  
and **blurred**  
**predictions**

What's **wrong** in  
these models?





3

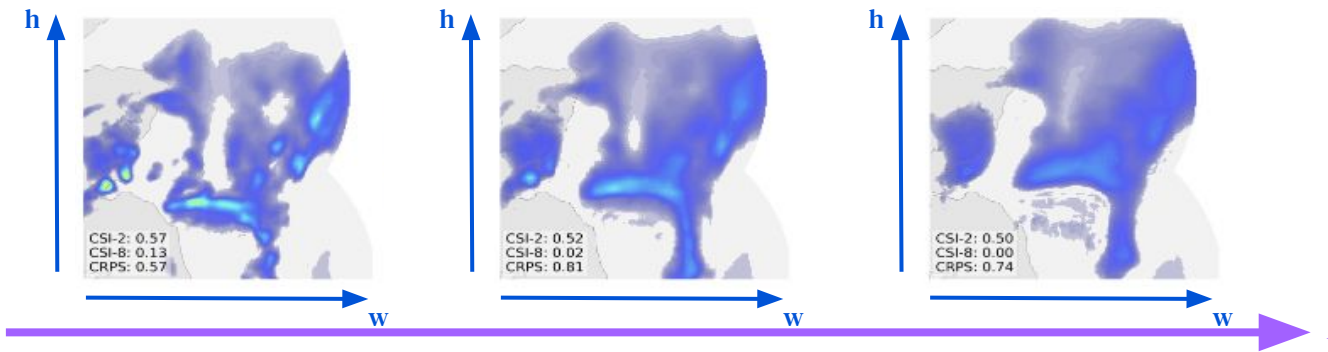
... or  
why we need  
generative  
models



# Beware of Simplistic Modeling Assumptions in the Loss!

How are we modeling  $p_\theta(x_t|x_{<t})$ ?

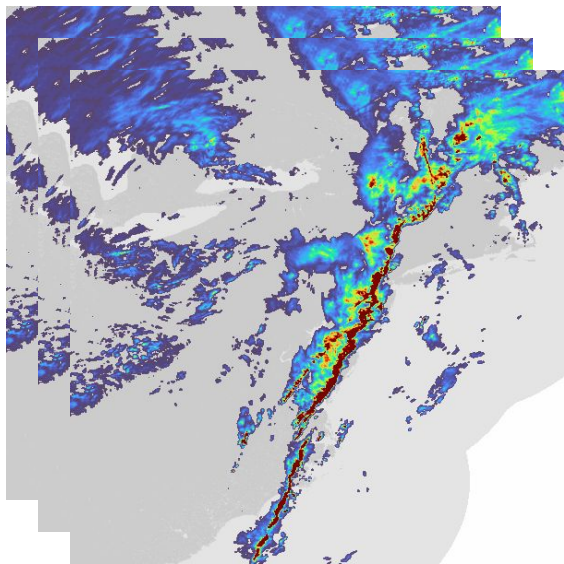
- Regression: 
$$= \prod_t \prod_h \prod_w \mathcal{N}(x_{t,h,w} | \hat{x}_{t,h,w}, \sigma^2) = \prod_t \prod_h \prod_w p_\theta(x_{t,h,w} | x_{<t})$$
- Classification: 
$$= \prod_t \prod_h \prod_w \frac{\exp(\hat{x}_{t,h,w}^{(x_{t,h,w})})}{\sum_j \exp(\hat{x}_{t,h,w}^{(j)})} = \prod_t \prod_h \prod_w p_\theta(x_{t,h,w} | x_{<t})$$
- Be careful about the objective, especially if the data has underlying stochasticity ... as there may be some poor conditional independence assumptions!



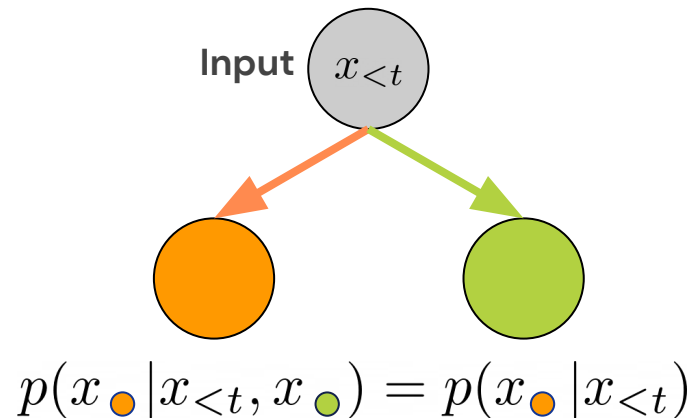
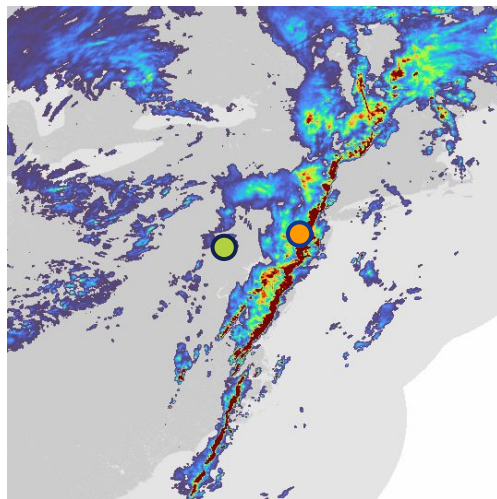
# Beware of Simplistic Modeling Assumptions in the Loss!

- Conditional independence assumption:  
Given the input, two predicted grid cells are independent

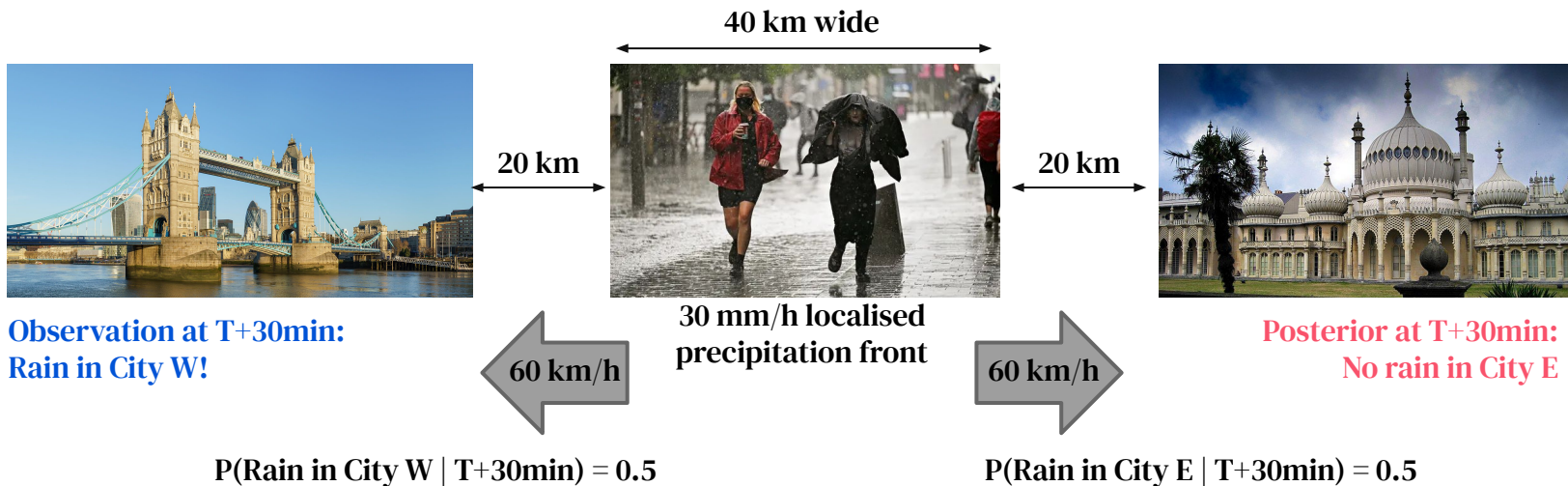
Input  $x_{<t}$



Prediction



# A Thought Experiment



$f(\text{London}, T+30\text{min}) = 15 \text{ mm/h}$

Regression

$f(\text{Brighton}, T+30\text{min}) = 15 \text{ mm/h}$

CSI = 0.5 at 15 mm/h (good metric value)

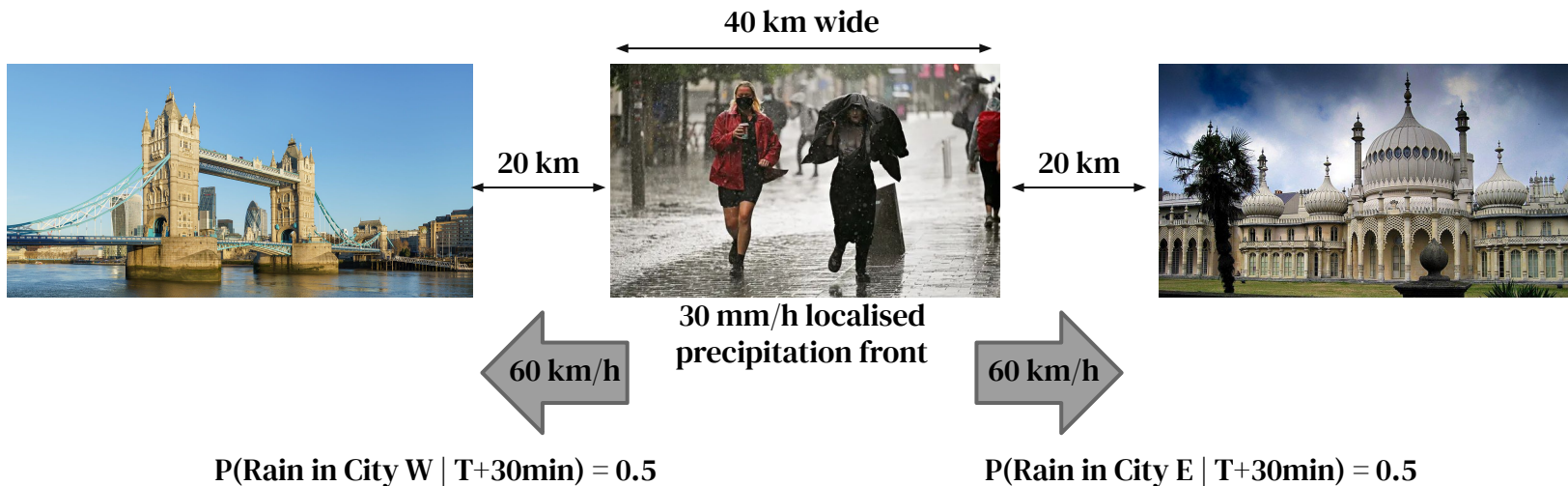
$p(\text{London}, 30 \text{ mm/h}, T+30\text{min}) = 0.5$

Classification

$p(\text{Brighton}, 30 \text{ mm/h}, T+30\text{min}) = 0.5$

CSI = 0.5 at 30 mm/h (good metric value)

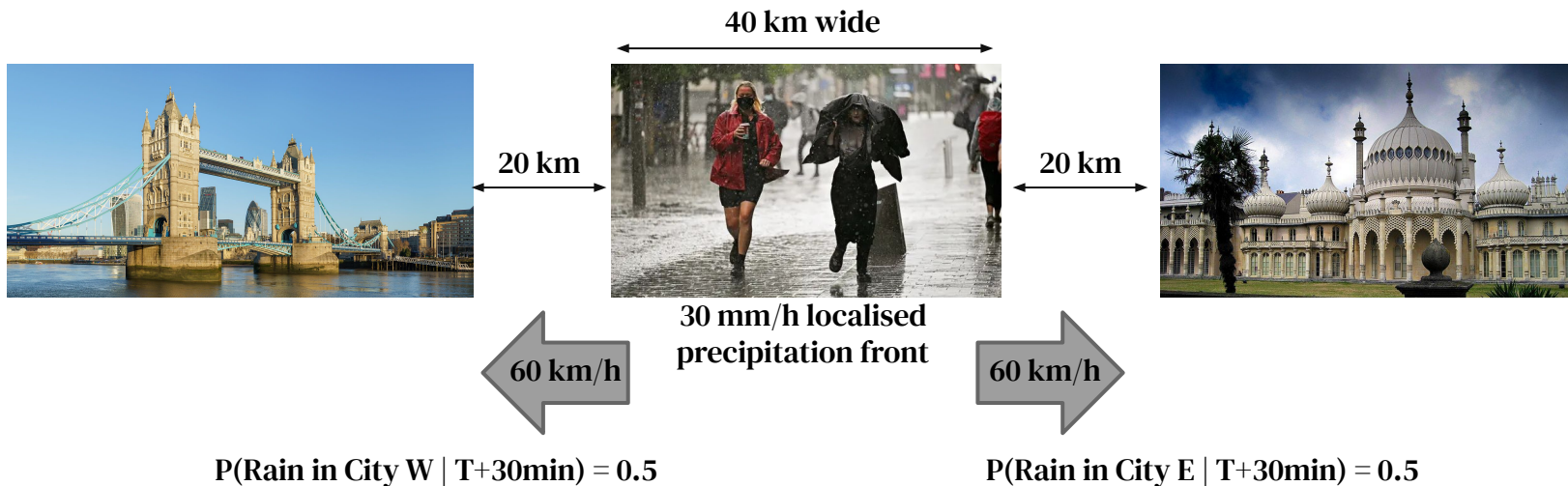
# A Thought Experiment



But what about questions we care about?



# A Thought Experiment



What is the probability of rain at both cities simultaneously at T+30min?

**Truth**

$P(\text{Rain in City E and City W}) = 0.0$

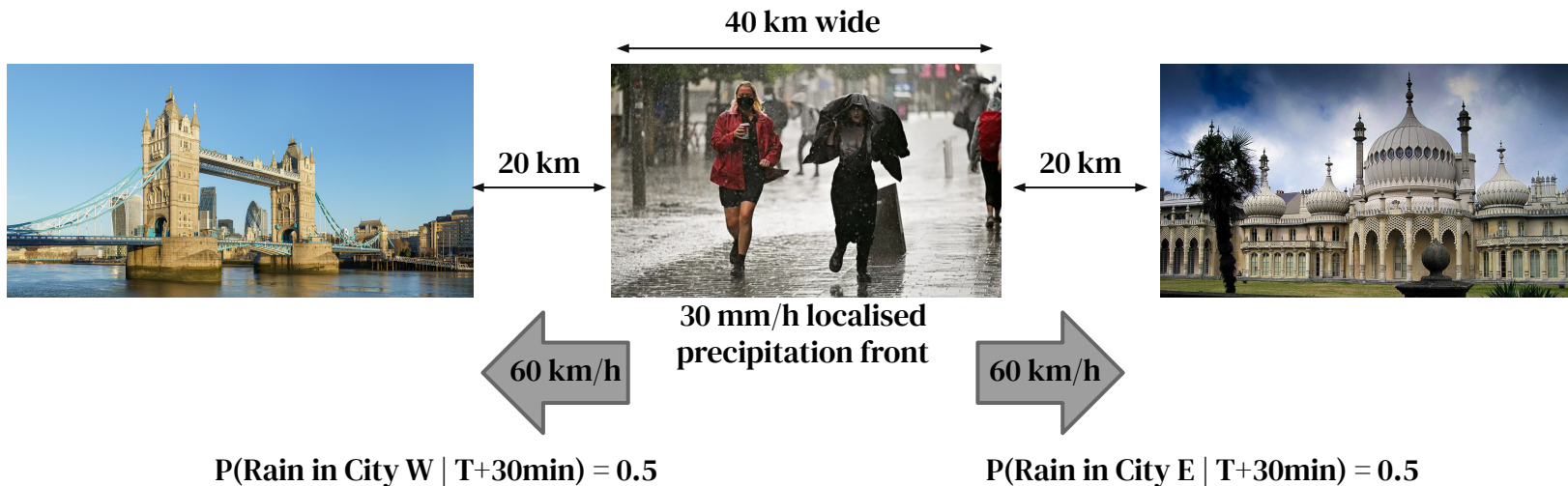
**Regression**

$P(\text{Rain in City E and City W}) = 1.0$

**Classification**

$P(\text{Rain in City E and City W}) = 0.25$

# A Thought Experiment



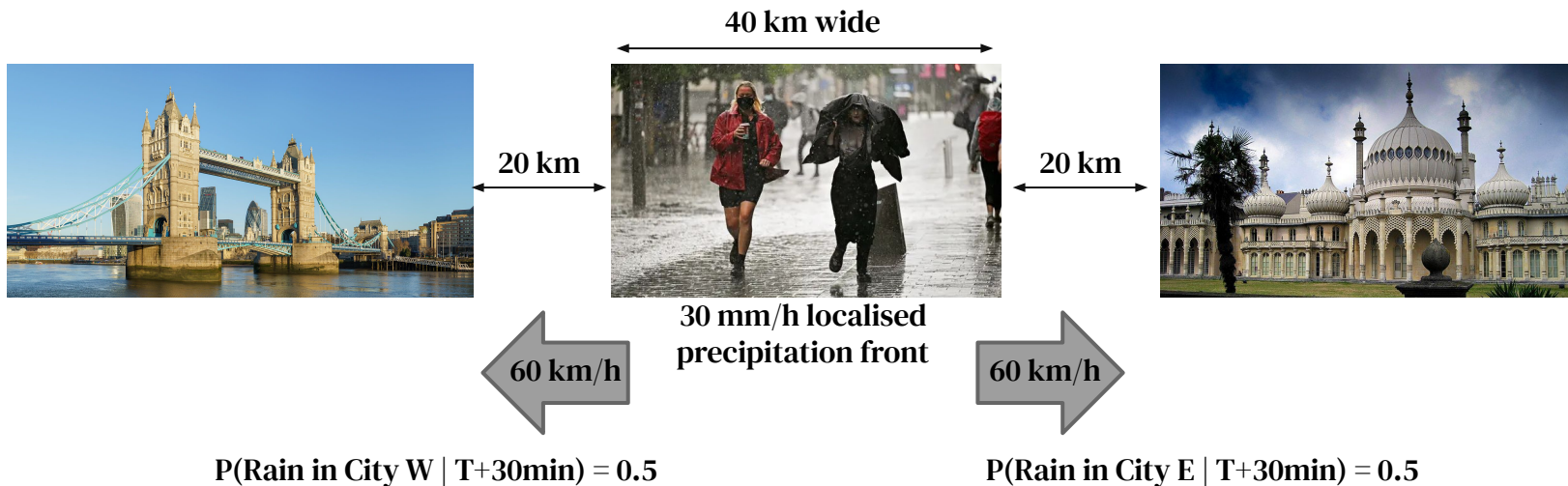
What is the probability of no rain at City E over the next hour?

**Truth**  
 $P(\text{No Rain in City E}) = 0.5$

**Regression**  
 $P(\text{No Rain in City E}) = 0.0$

**Classification**  
 $P(\text{No Rain in City E}) = 0.5^8$

# A Thought Experiment



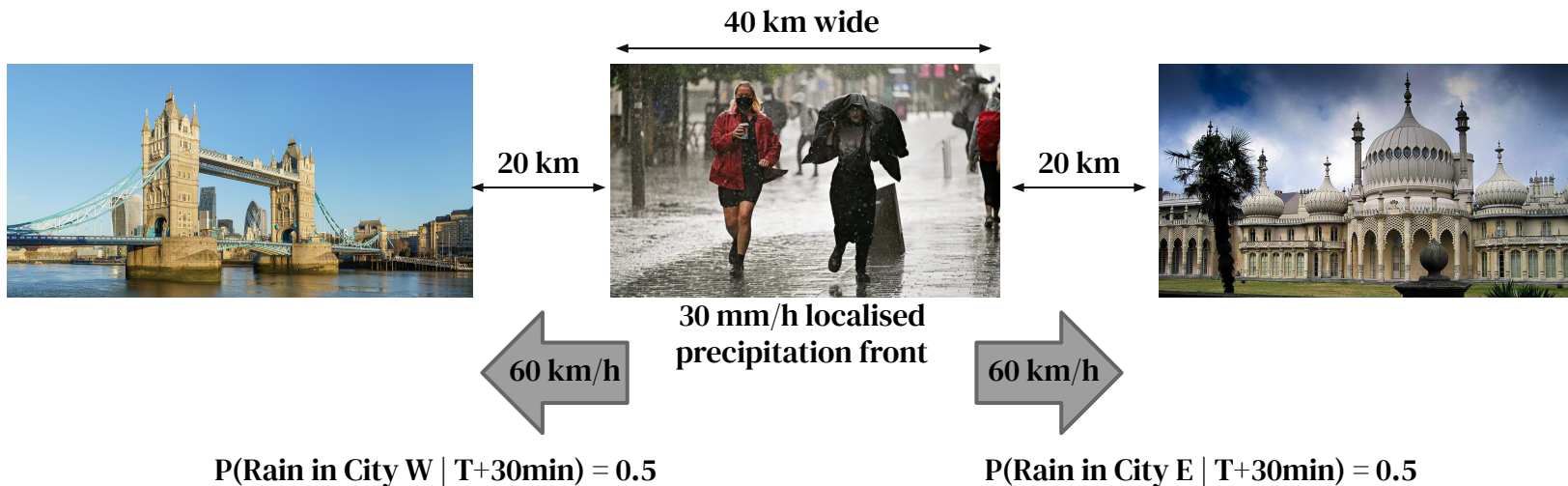
What is the probability of 20mm of rain at City E over the next hour?

**Truth**  
 $P(20\text{mm in City E}) = 0.5$

**Regression**  
 $P(20\text{mm in City E}) = 0.0$

**Classification**  
 $P(20\text{mm in City E}) = 0.5^8$

# A Thought Experiment

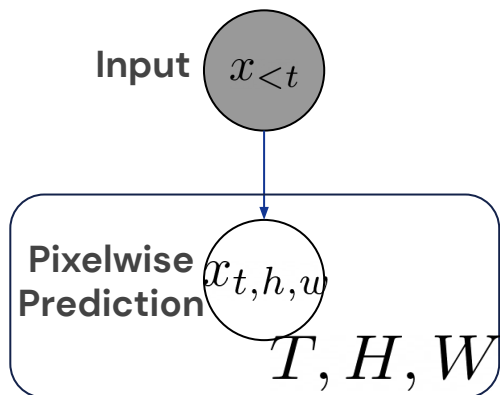


**Forecast quality** (scores on metrics) is strong,  
**but forecast value** (ability to make better decisions) is poor!

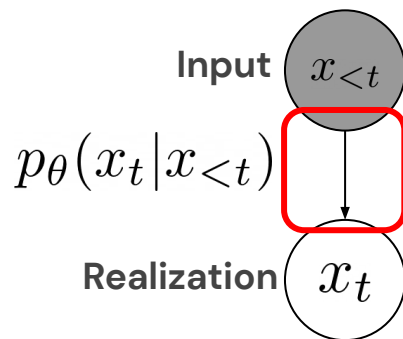
# From Discriminative to Generative Modeling

- What went wrong?
  - Baseline models directly predict grid cell-wise rain-rate probability
  - These are **discriminative models**
- We want **generative models**

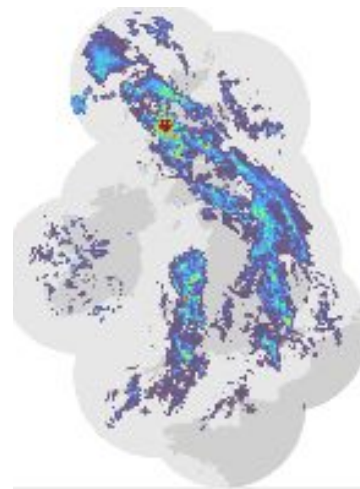
**Discriminative Model**



**Generative Model**



**Realization**



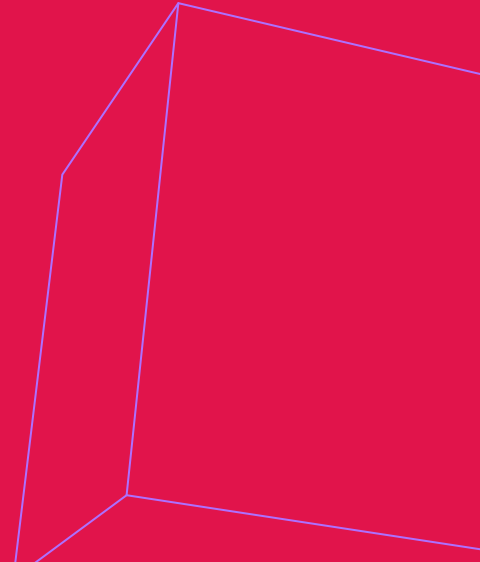
- But how to model  $p_{\theta}(x_t | x_{<t})$ ?



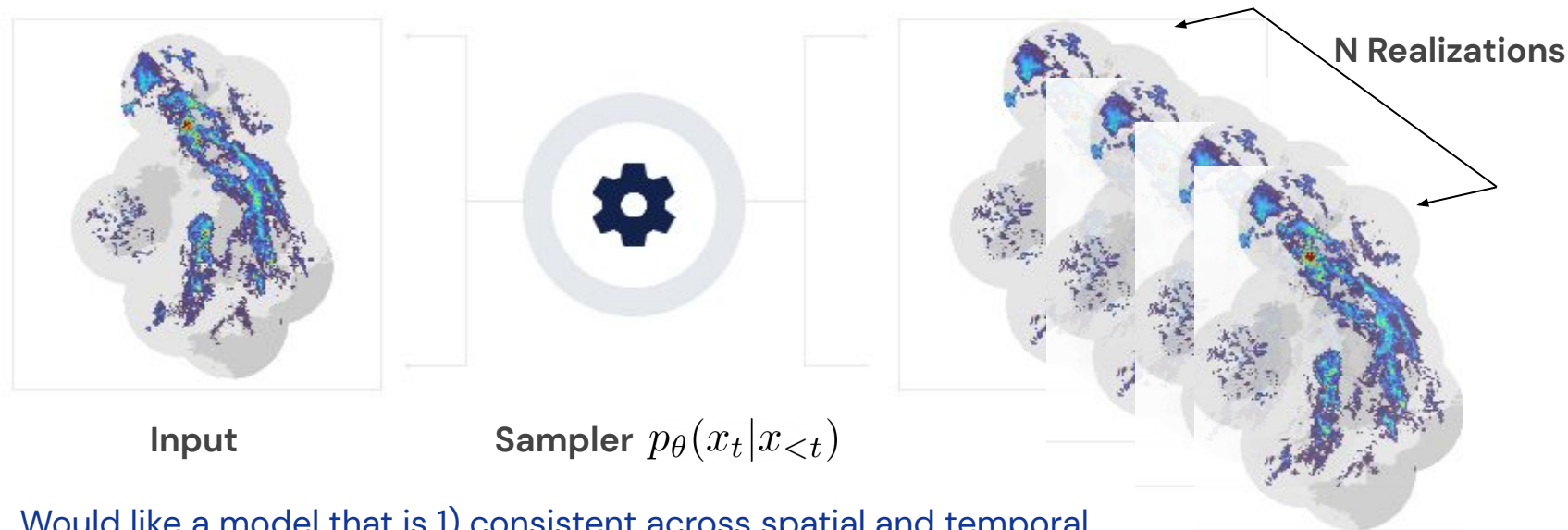


4

# Deep Generative Models of Radar

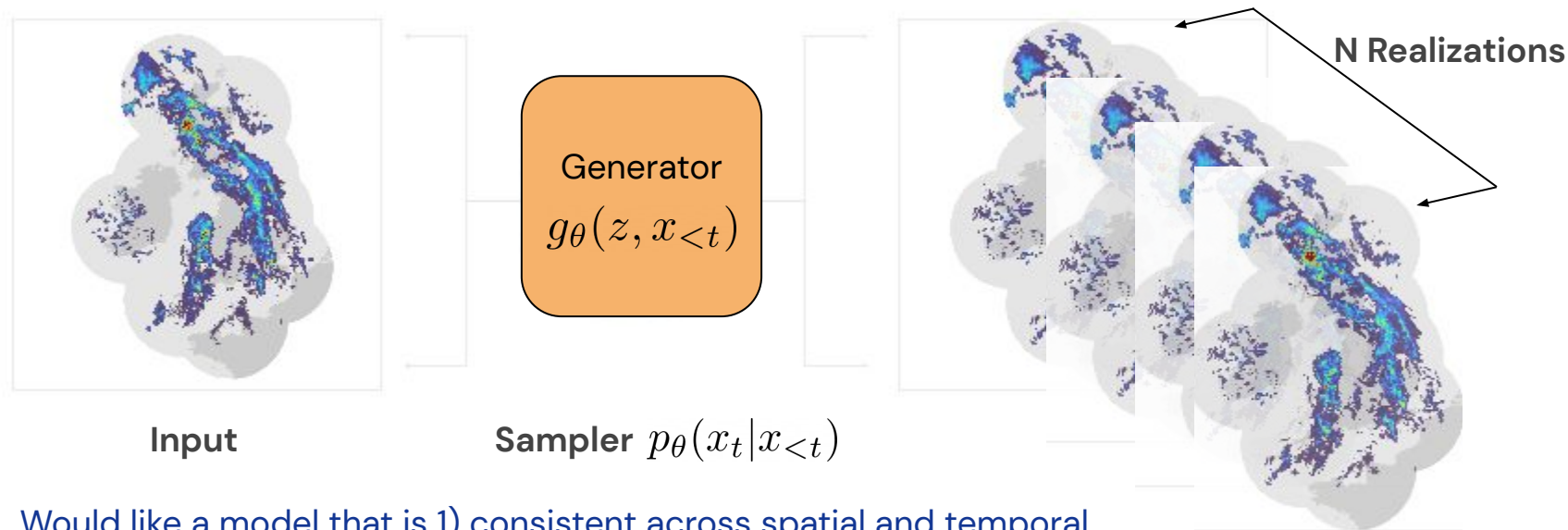


# A Neural Sampler: Take Inspiration from Ensemble NWP



- Would like a model that is 1) consistent across spatial and temporal scales, 2) captures rare events, and 3) properly accounts for uncertainty
- From realizations, calculate relevant probabilities (such as accumulation rain over catchment area)
- **Generative Adversarial Network seems like a good fit**

# A Neural Sampler: Take Inspiration from Ensemble NWP



- Would like a model that is 1) consistent across spatial and temporal scales, 2) captures rare events, and 3) properly accounts for uncertainty
- From realizations, calculate relevant probabilities (such as accumulation rain over catchment area)
- **Use models inspired by BigGAN and DVD-GAN**

[5] Brock, Donahue, and Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. ICLR 2019

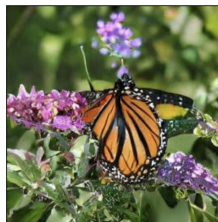
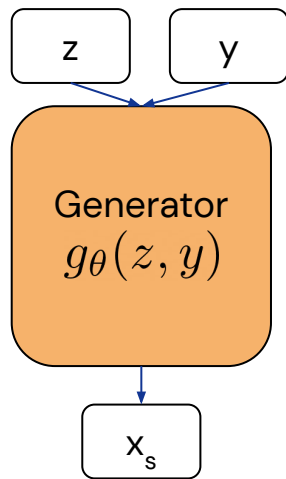
[6] Clark, Donahue, and Simonyan. Adversarial Video Generation on Complex Datasets. Arxiv 2019

[7] Luc et al., Transformation-based Adversarial Video Prediction on Large-Scale Data. Arxiv 2020.

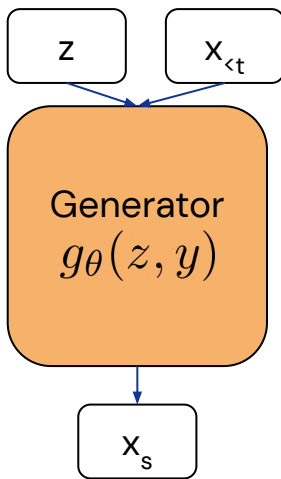
# BigGAN and DVD-GAN

BigGAN

$y \sim \text{Cat}(\text{Butterfly}, \dots, \text{Dog})$

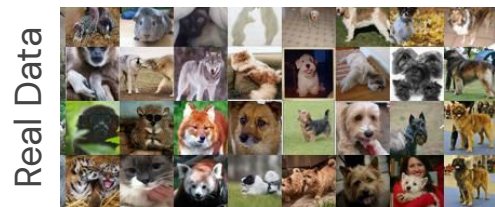
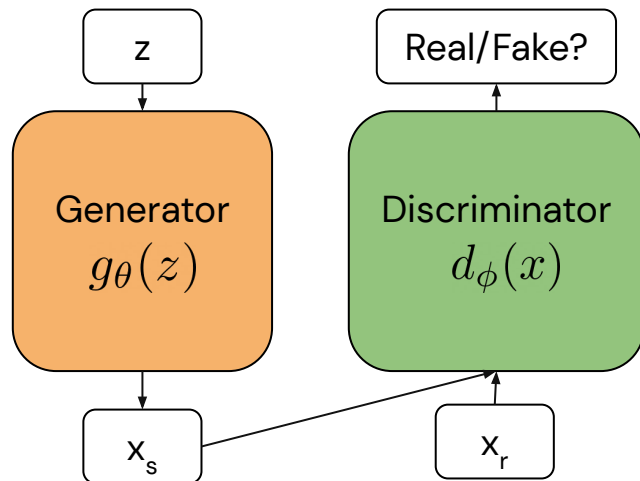


DVD-GAN



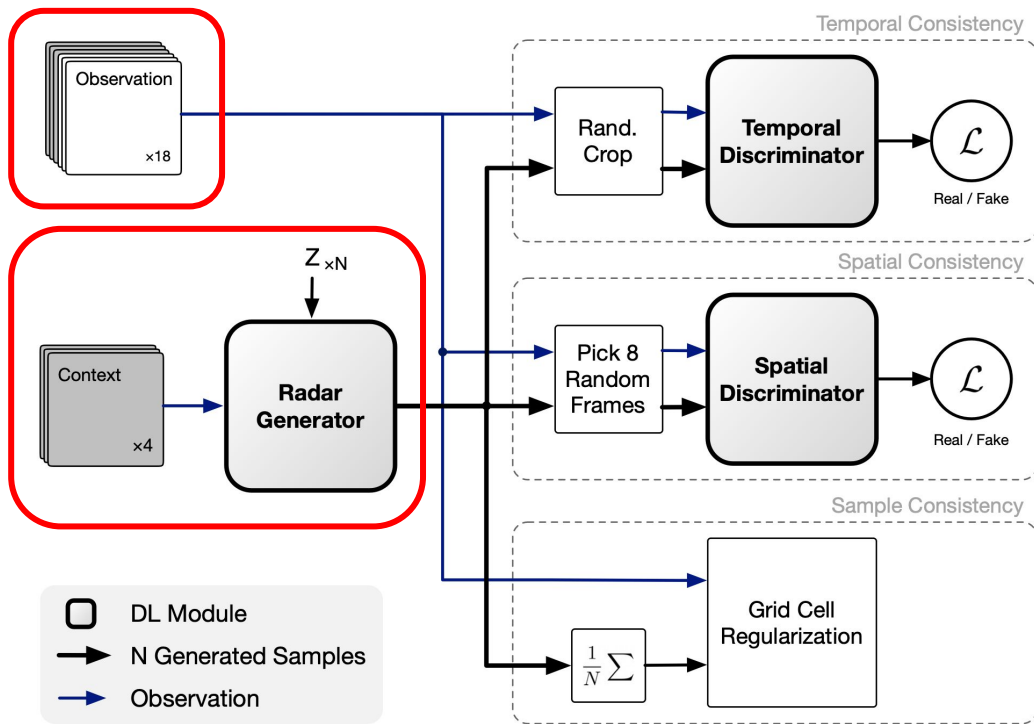
Adversarial Training

$z \sim \mathcal{N}(0, I)$



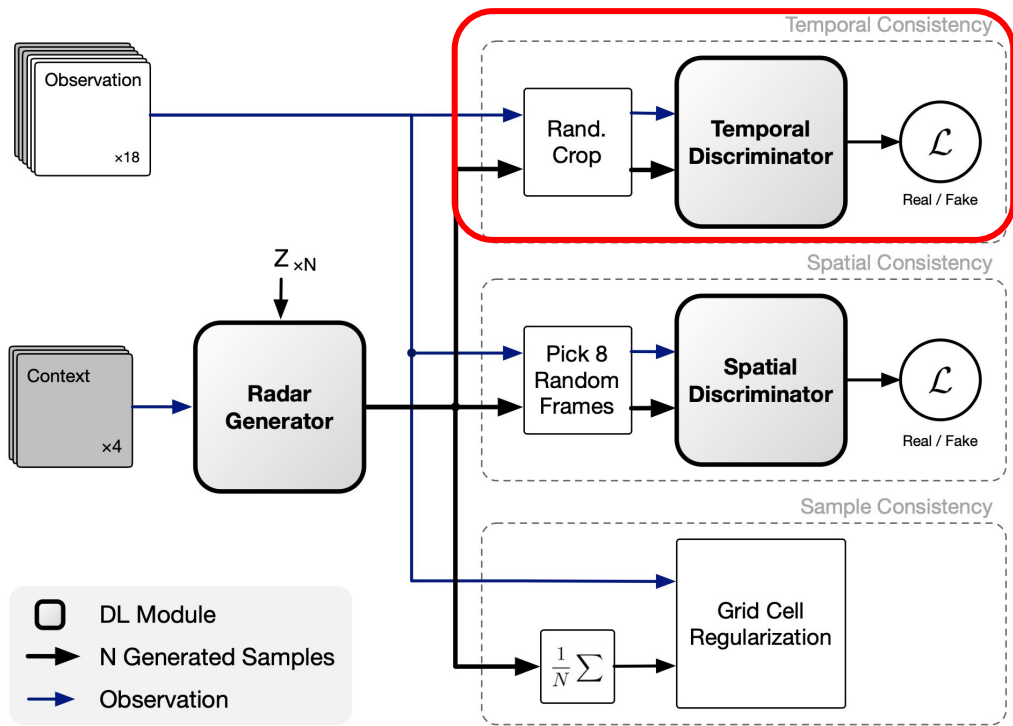
Source: <https://cs.stanford.edu/people/karpathy/cnnembed/>

# Nowcasting model: conceptual diagram

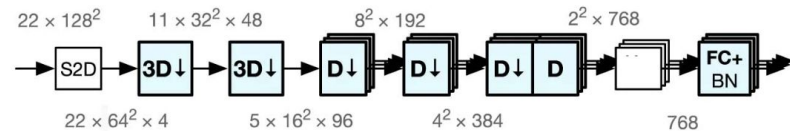




# Nowcasting model: conceptual diagram

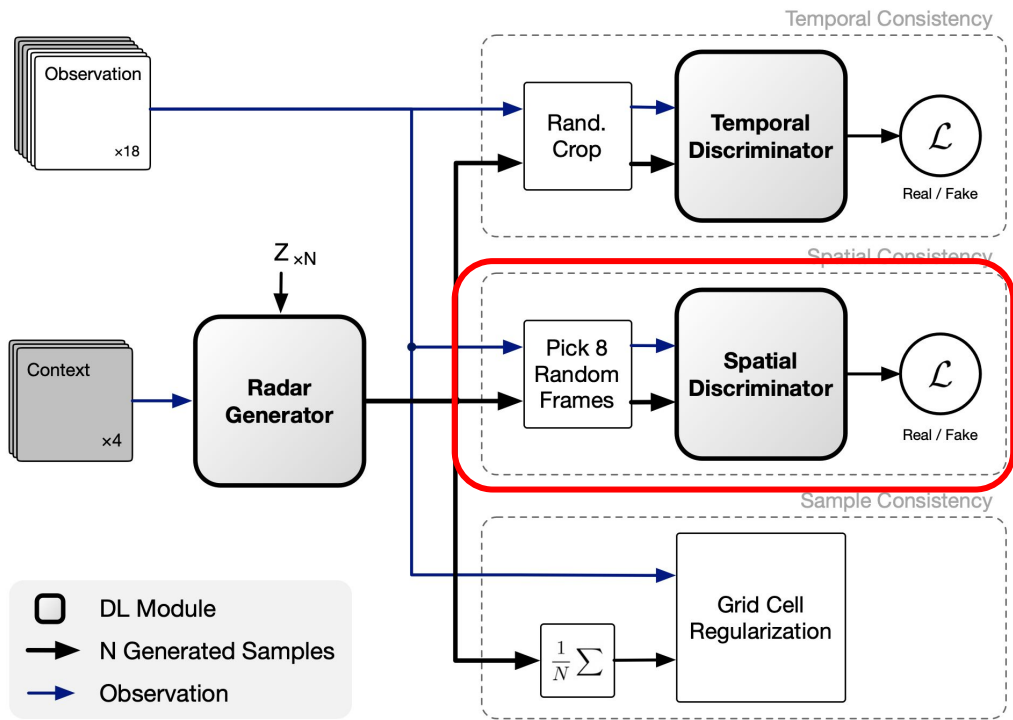


Temporal Discriminator



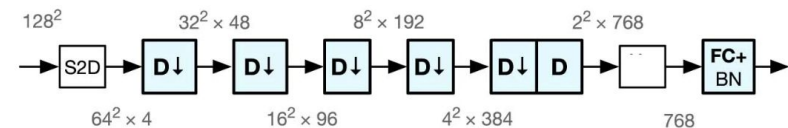
Spatio-temporal consistency of realisations.

# Nowcasting model: conceptual diagram

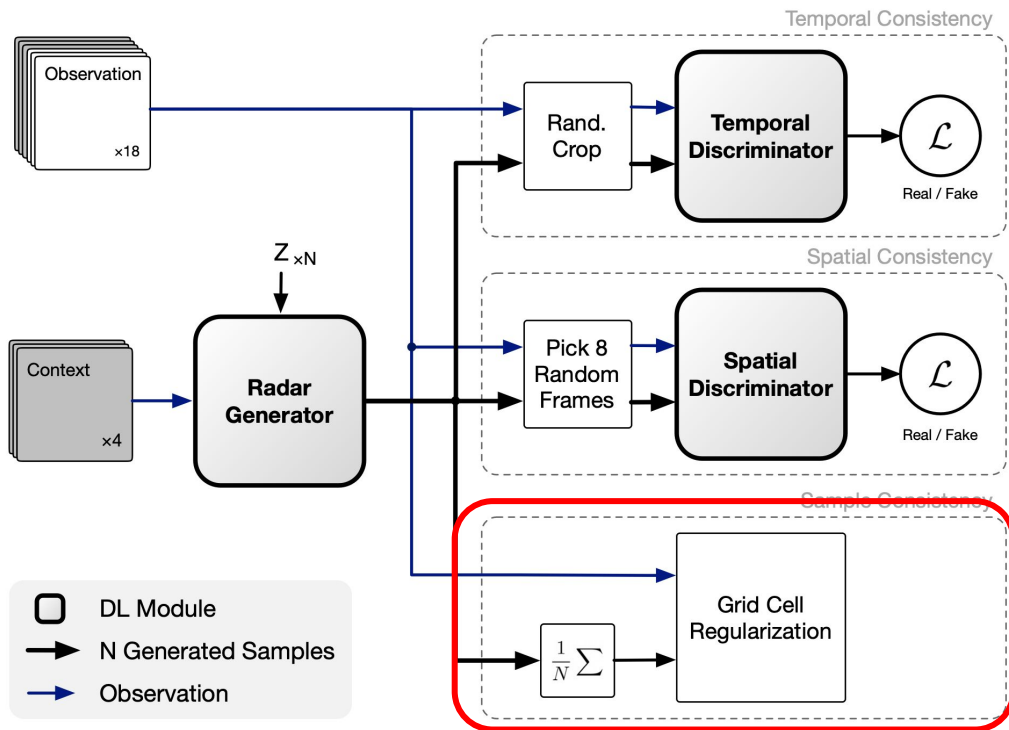


Ensures that each frame is spatially consistent

Spatial Discriminator



# Nowcasting model: conceptual diagram

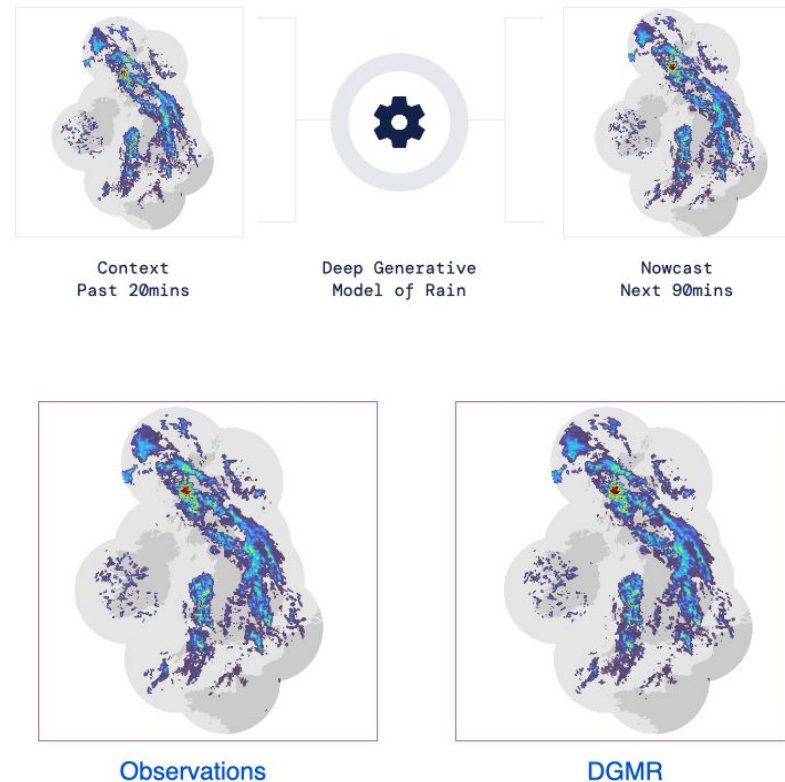
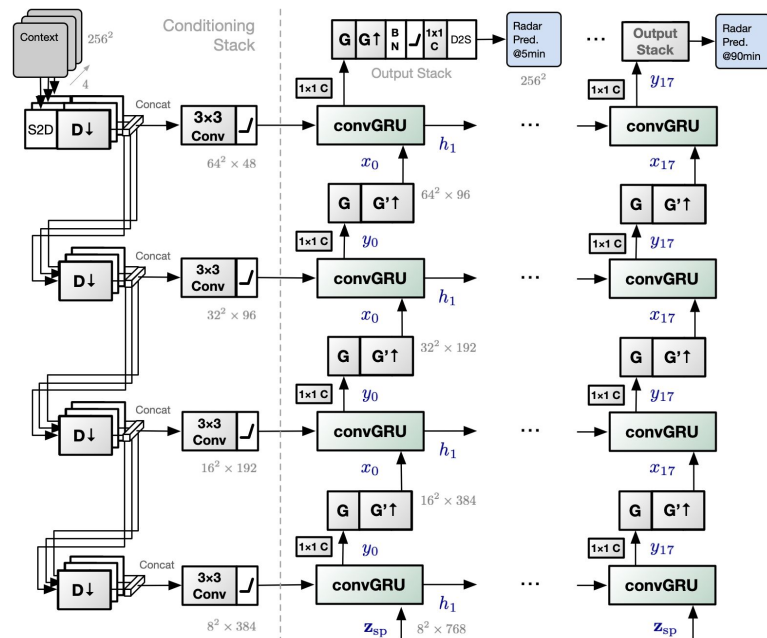


We can also think of this as the primary prediction objective, where the two GAN losses are the actual regularisation

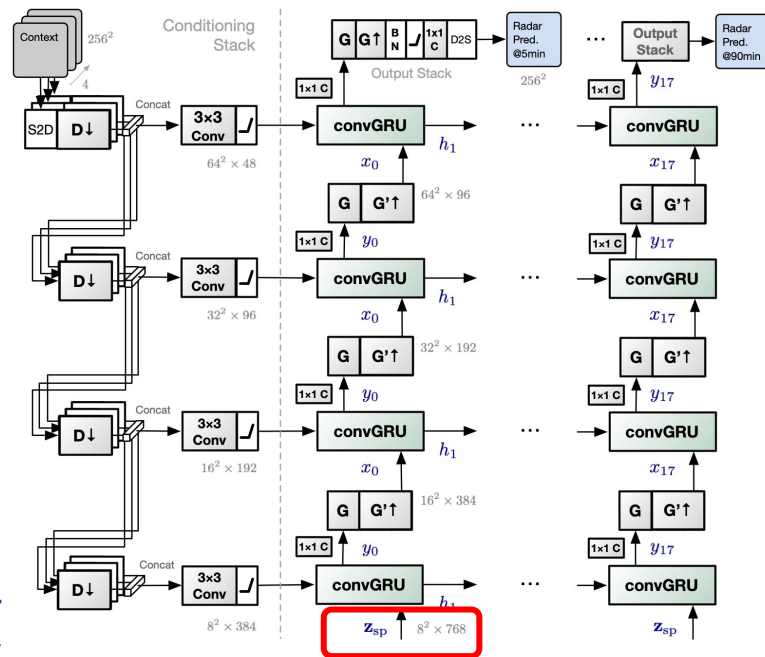
Regularization ensures location accuracy.

$$\mathcal{L}_R(\theta) = \frac{1}{HWN} \left\| (\mathbb{E}_Z[G_\theta(Z; \mathbf{X}_{1:M})] - \mathbf{X}_{M+1:M+N}) \odot w(\mathbf{X}_{M+1:M+N}) \right\|_1.$$

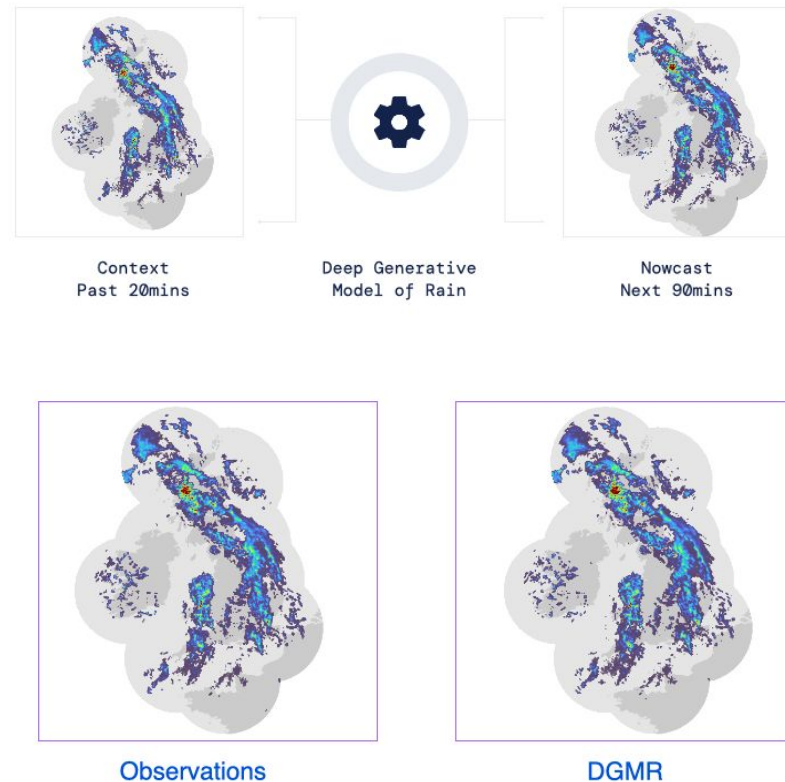
# Generator



# Generator

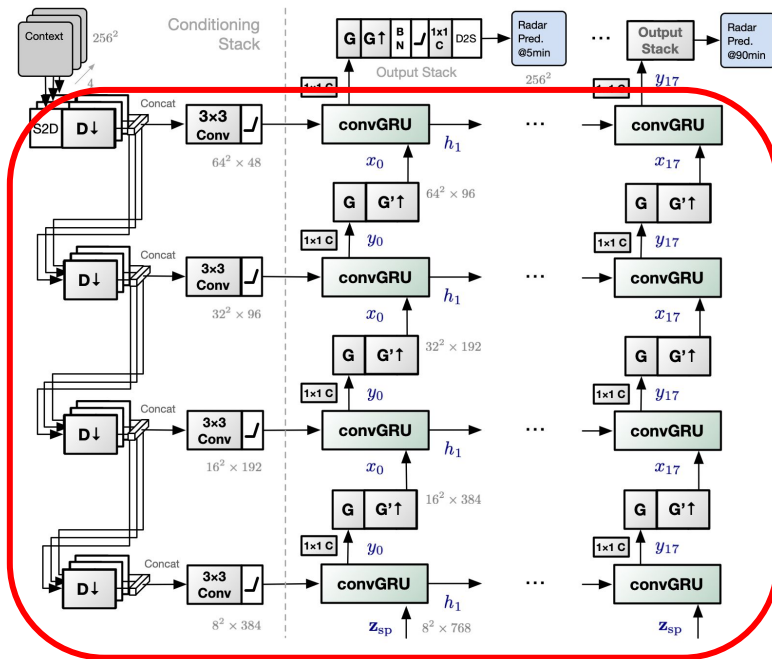


Latents  
account for  
uncertainty

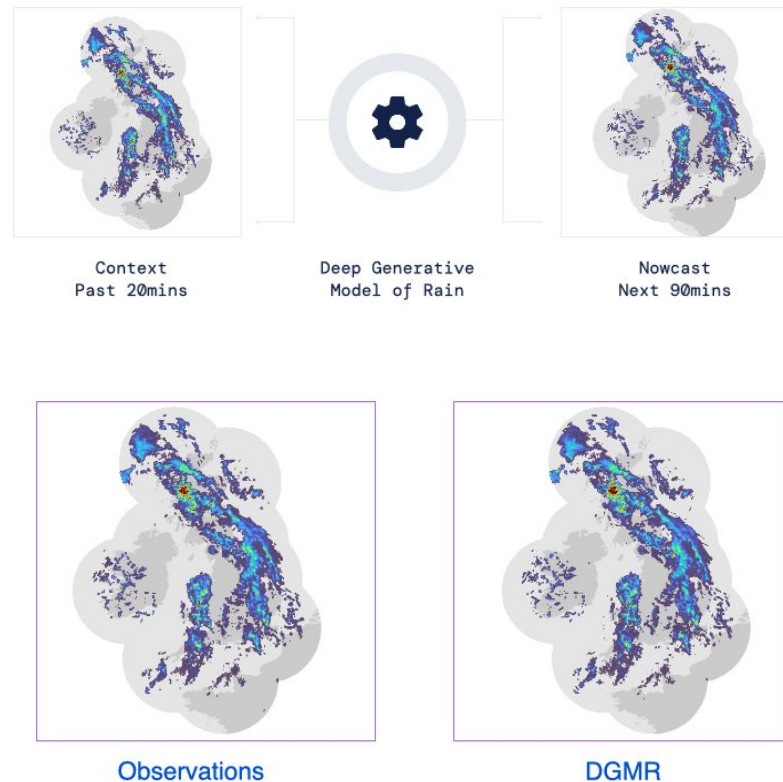




# Generator

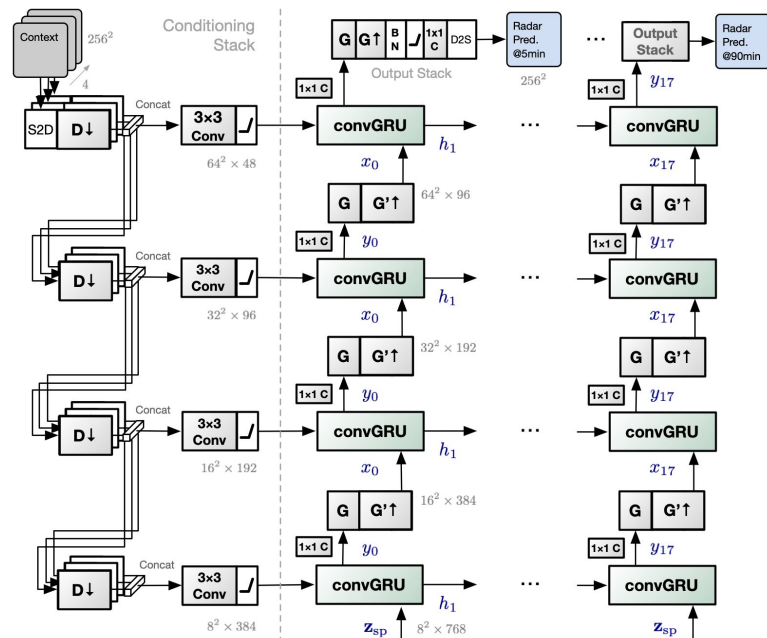


Architecture  
for multiple  
scales



# Generator

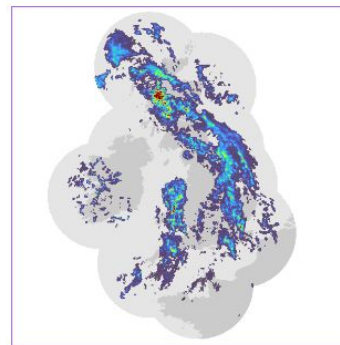
But what  
about rare  
events?



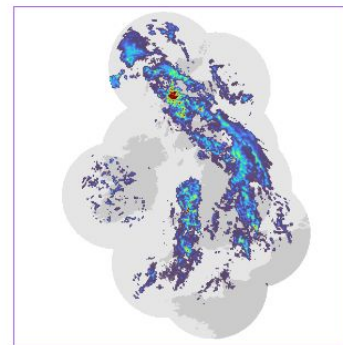
Context  
Past 20mins

Deep Generative  
Model of Rain

Nowcast  
Next 90mins



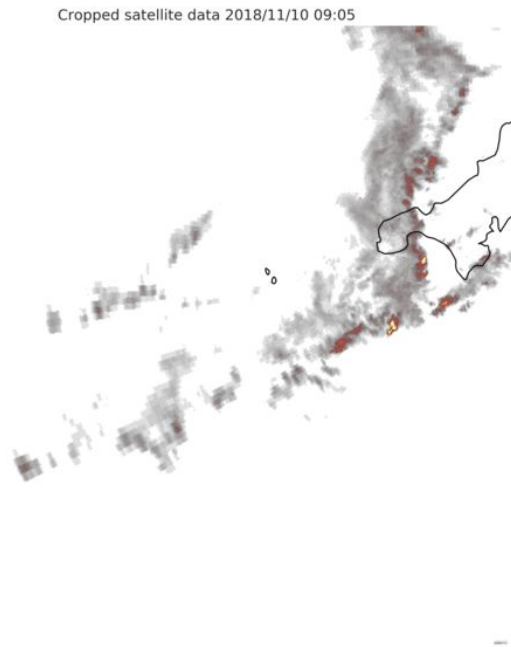
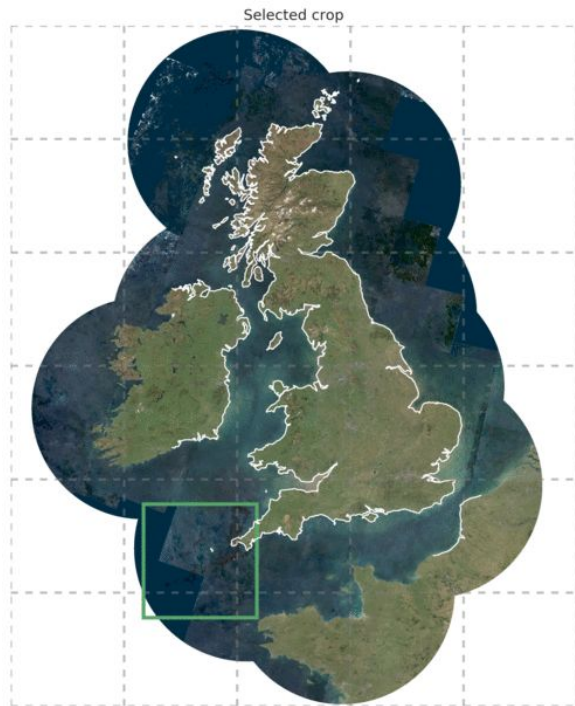
Observations



DGMR

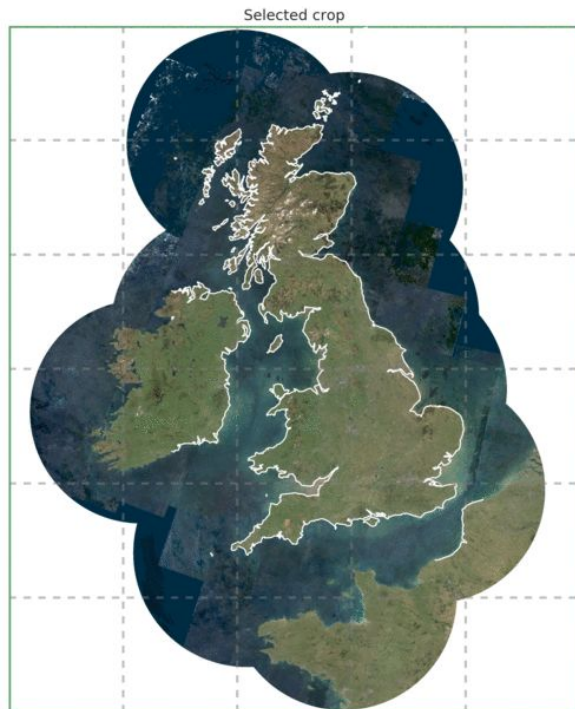
# Met Office RadarNet4 Data

Bias sampling towards high average precipitation for training.



# Fully convolutional architecture, enables full-frame evaluation

Test on full 1536 x 1280 UK radar frames (2019) – (for US data: 3584 x 7168 frames)



Cropped satellite data 2018/10/29 22:30



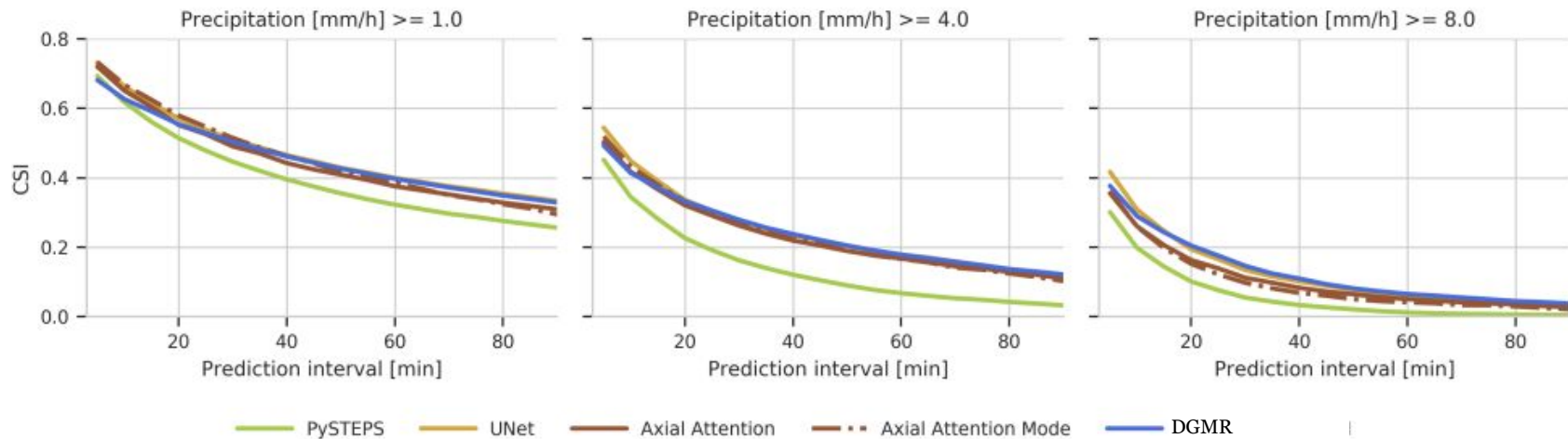


5

# Quantitative Verification and its Limits



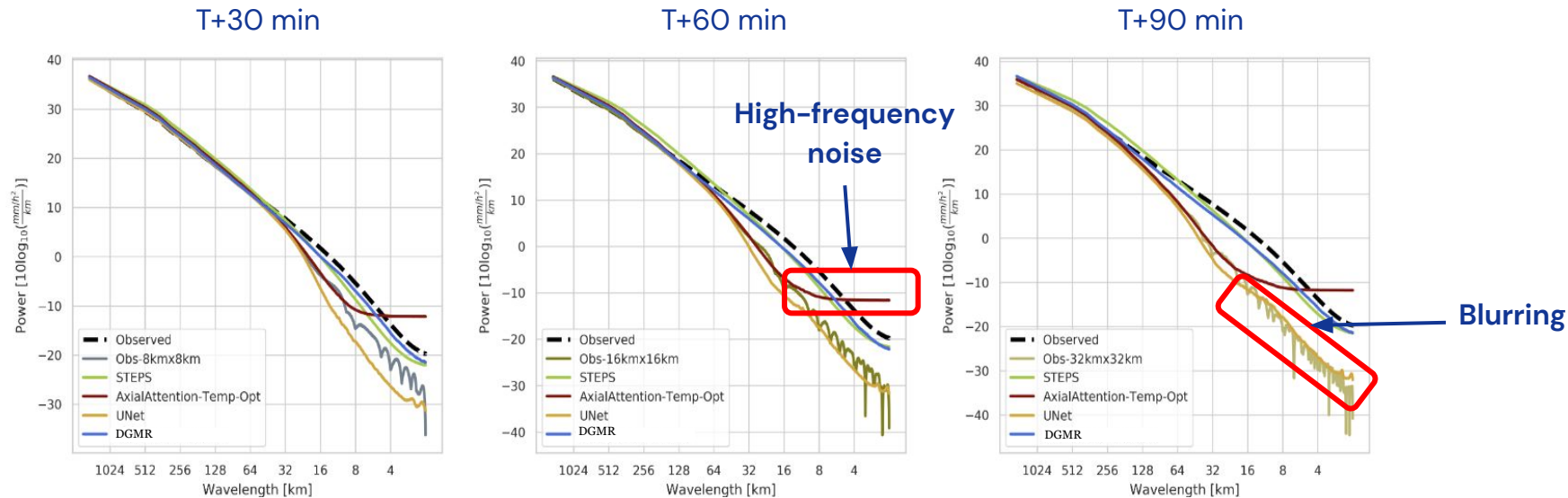
# Critical Success Index (CSI)



- Deep learning methods outperform PySTEPS.
- CSI does not account for blurry predictions.



# Power Spectral Density (PSD)

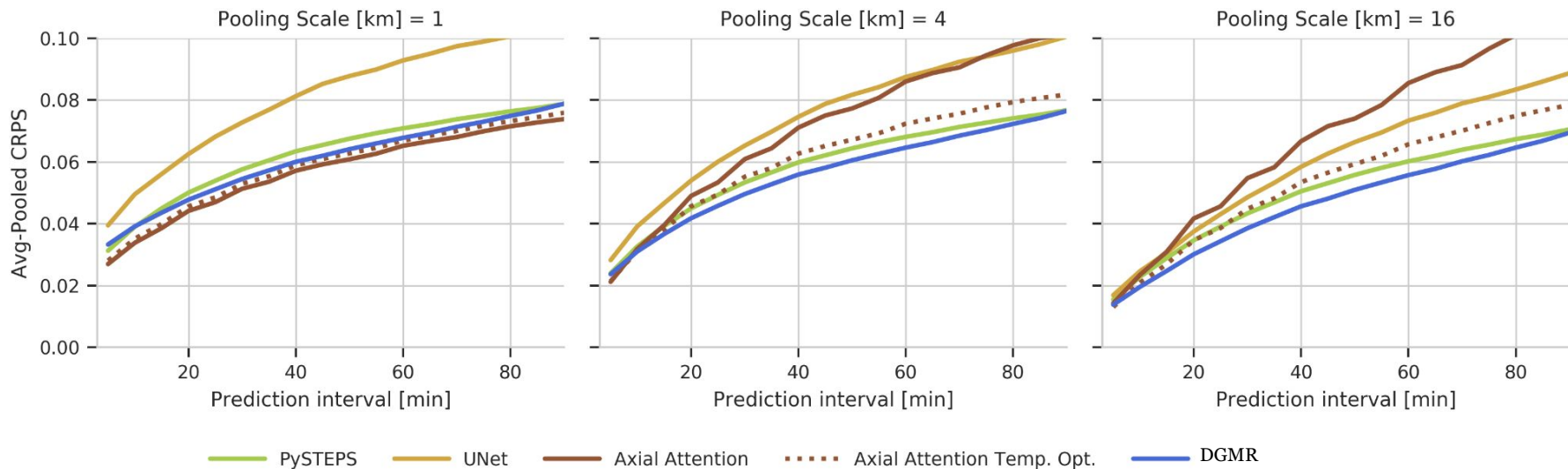


30 minute prediction from other DL models has the same resolution as 8kmx8km observations

90 minute prediction from other DL models has the same resolution as 32kmx32km observations

DGMR/PySTEPS has roughly the same resolution as the original 1kmx1km observations

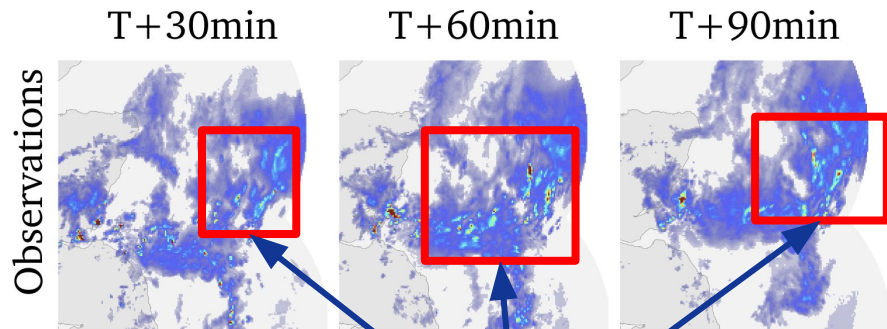
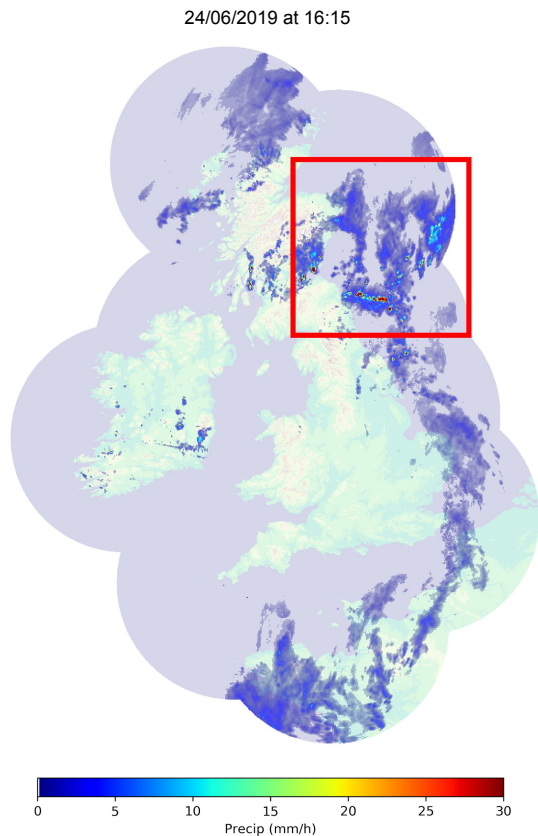
# Continuous Ranked Probability Score (CRPS)



## Probabilistic verification using CRPS

- Show CRPS aggregated over different scales (like FSS).
- Discriminative deep learning methods lack spatial consistency.
- Generative methods are spatially consistent.

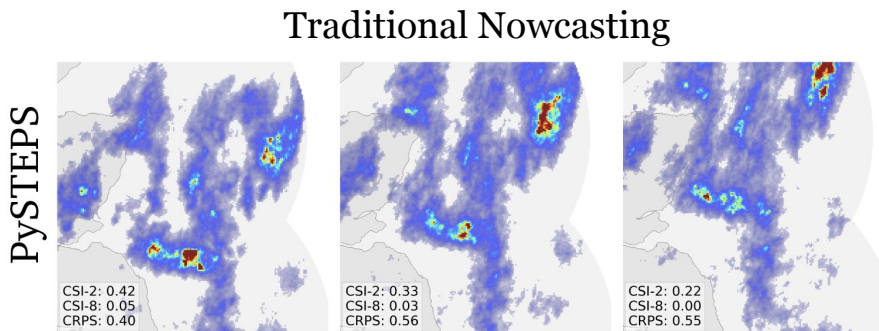
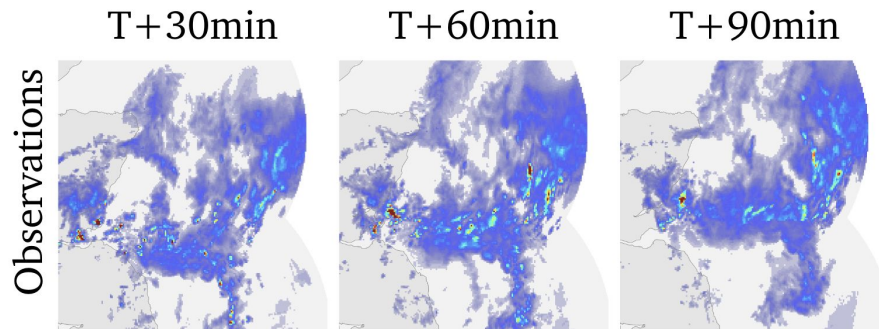
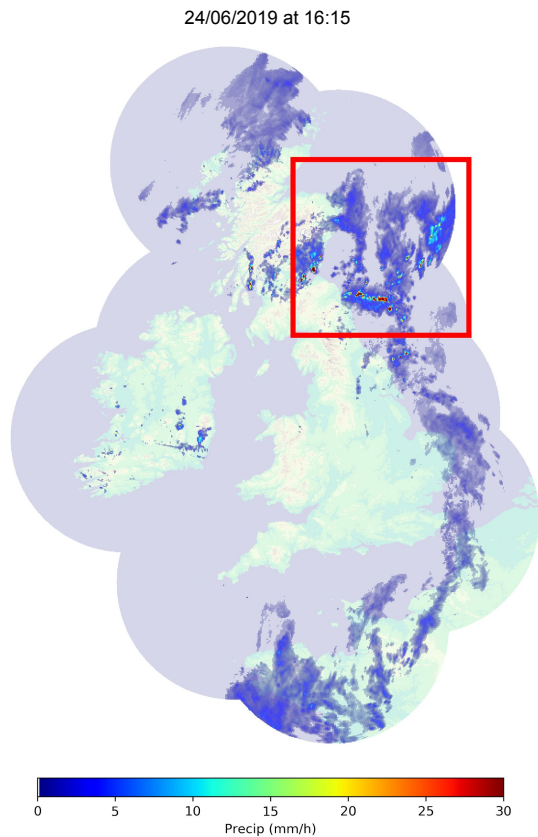
# Intercomparison case study



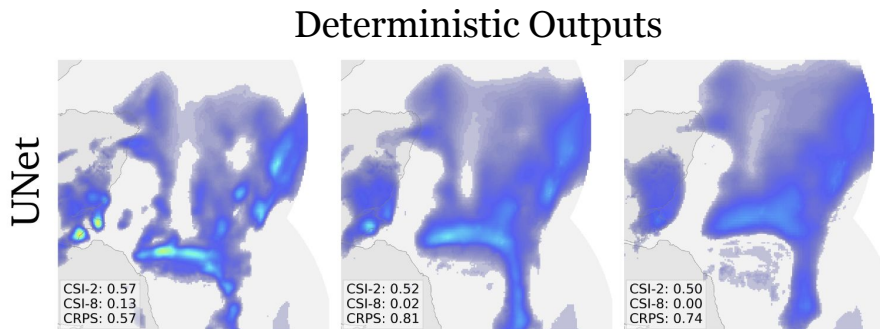
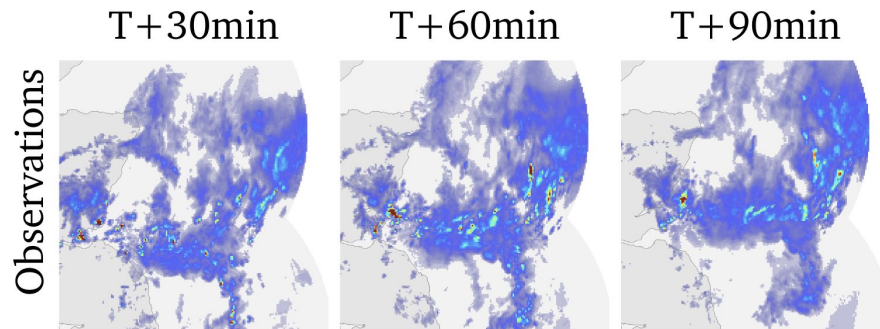
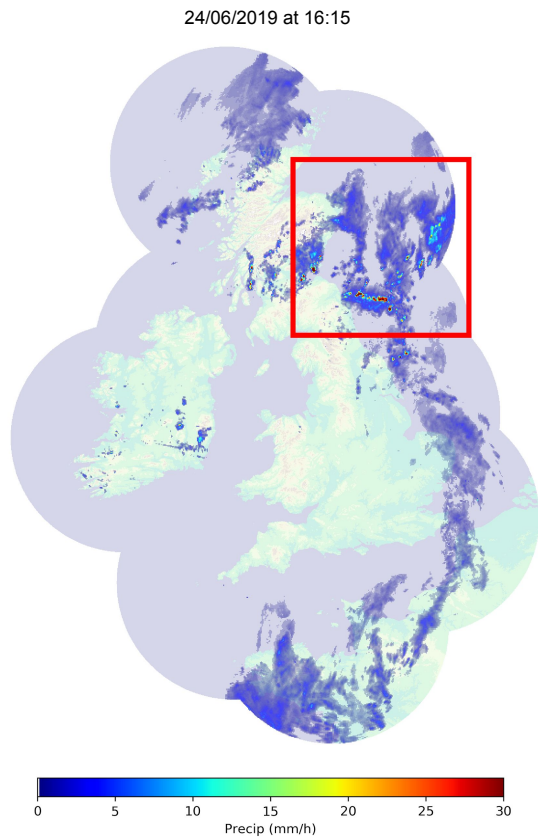
Important (and difficult) to predict  
convective cells

Difficult case chosen by the Chief Forecaster  
(independent of the project team)

# Intercomparison case study

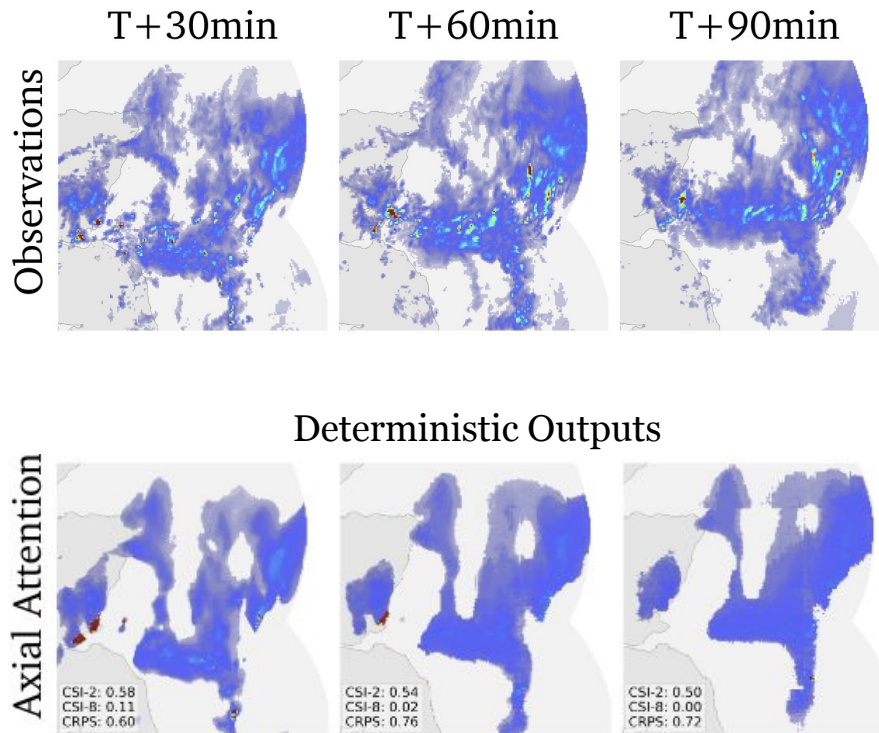
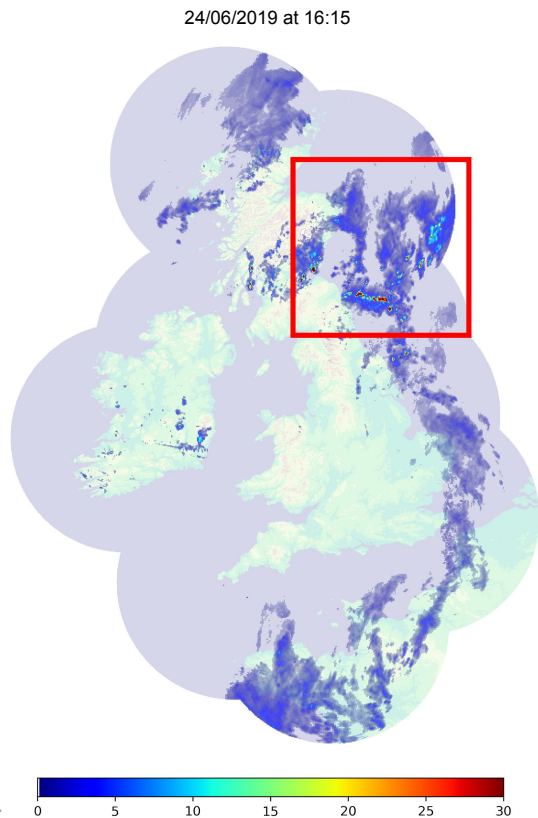


# Intercomparison case study



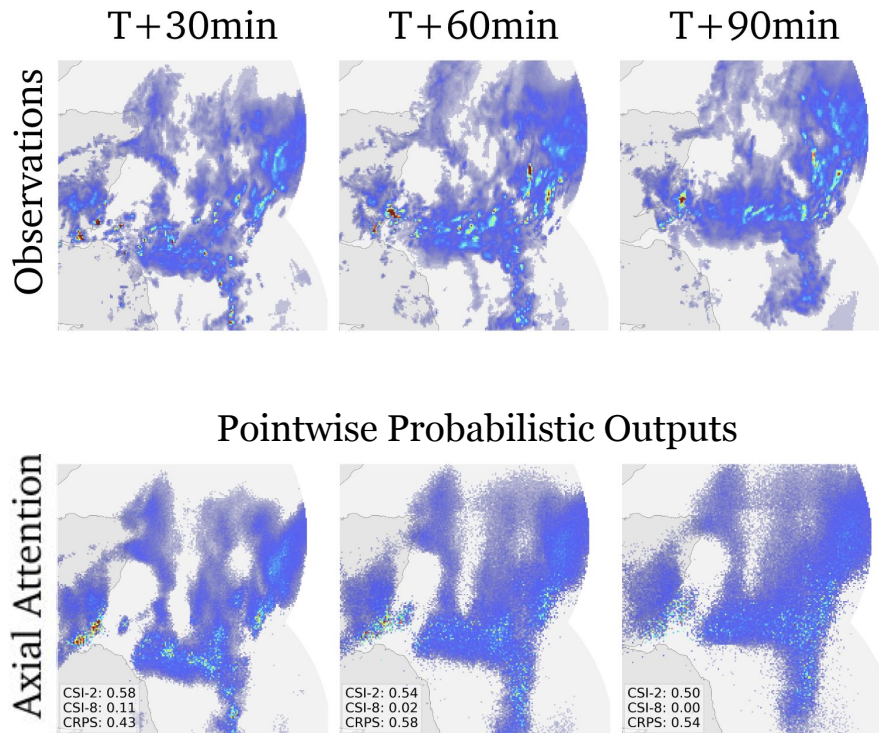
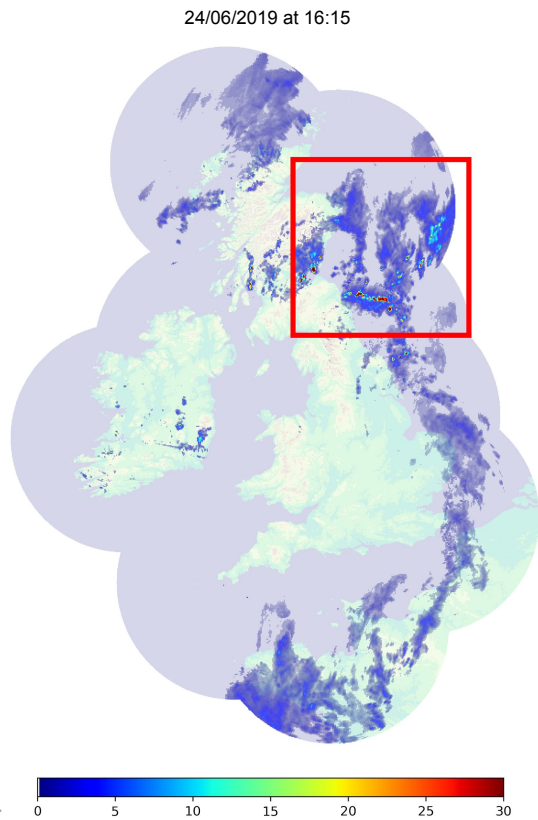


# Intercomparison case study

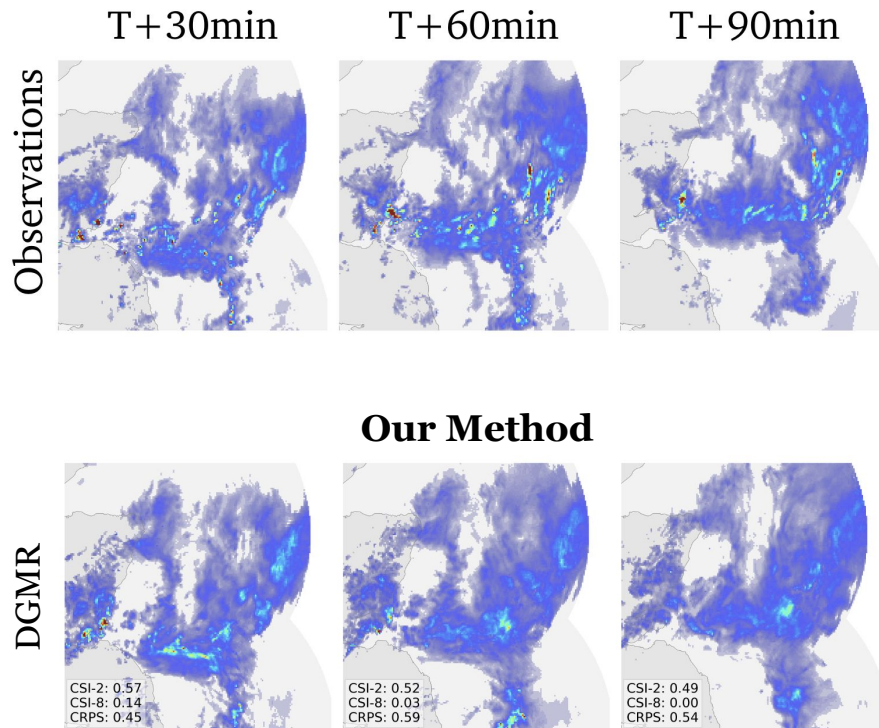
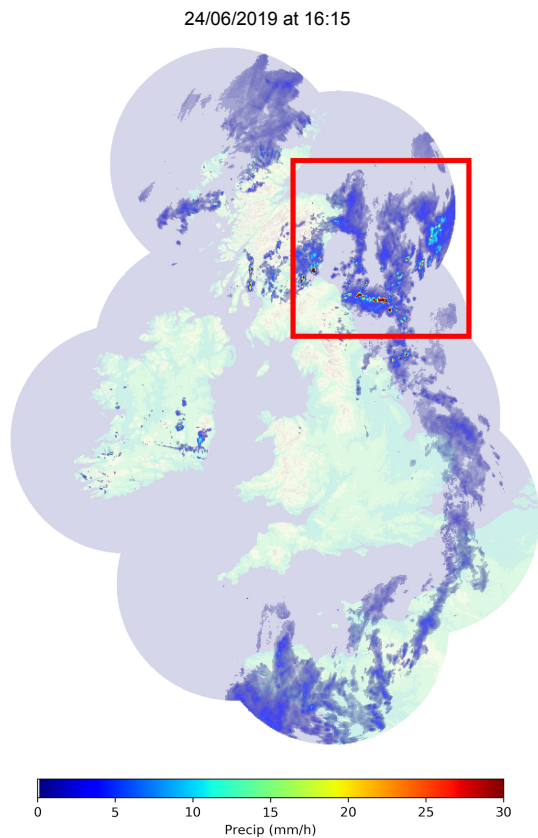




# Intercomparison case study

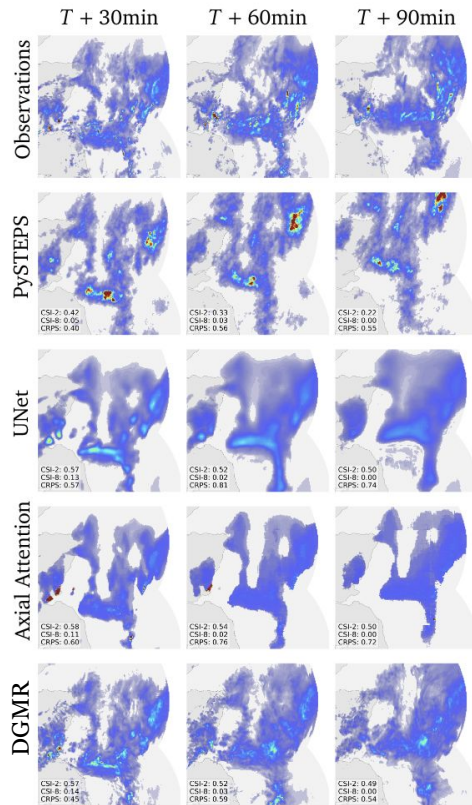
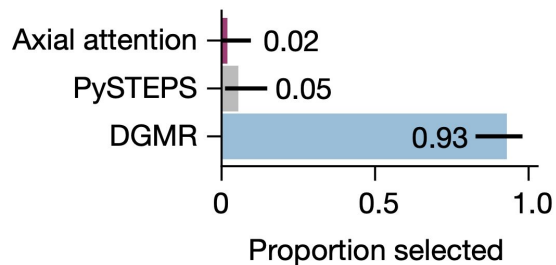


# Intercomparison case study



# Limitation of quantitative verification

We showed these nowcasts  
to expert forecasters...



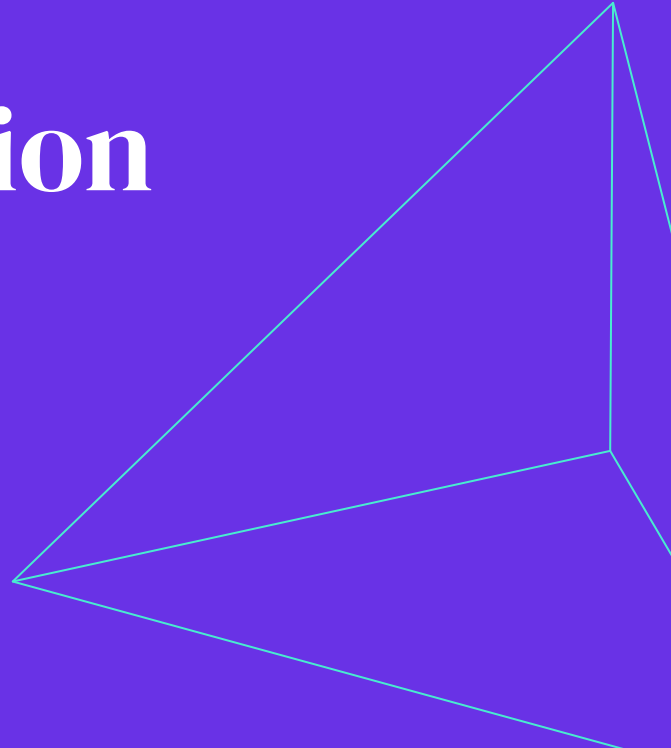
Existing metrics do not capture  
all the salient properties of a nowcast.

- ML methods **do not incorporate physical knowledge**
- ML methods are much more flexible and can **game the metrics**
- Existing metrics do not give us the ability to detect the differences between different approaches.
- Expert meteorologists have this knowledge



6

# Expert Evaluation



# Experimental psychology for operational meteorology

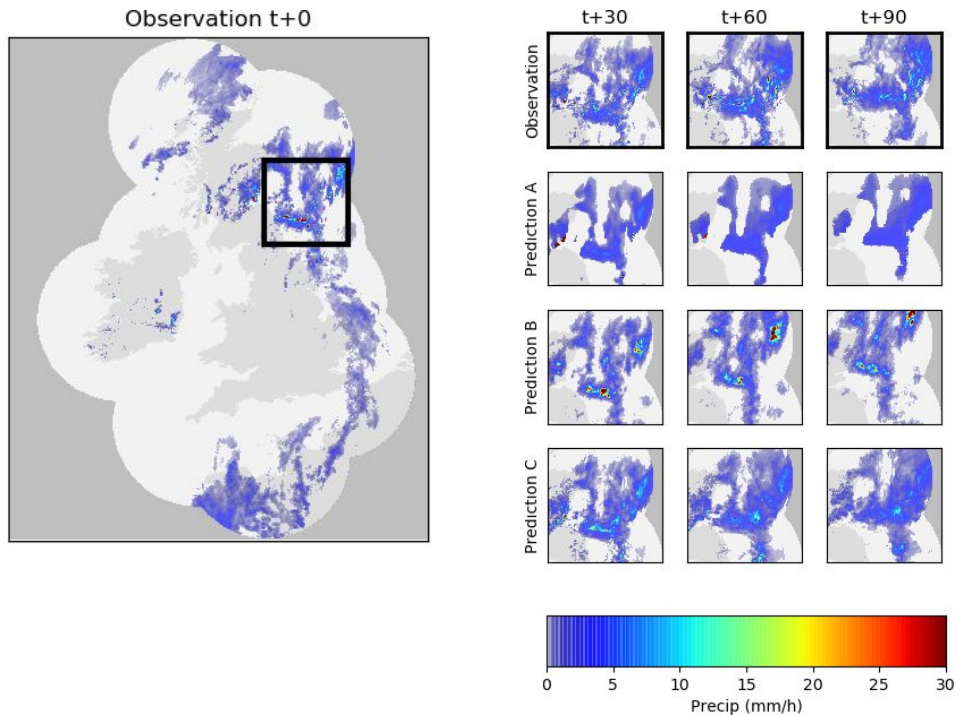
- Controlled study of expert judgments to assess performance of ML models.
- Worked with the Met Office Chief Forecaster to design the study and to adjust the type of assessments made.
- 56 UK Met Office professional meteorologists (anonymised) participated in this study.
- Insight:
  - Are our approaches useful in an operational setting?
  - Can they inform decisions on the development of new ML approaches?

# Protocol Design

- Our experiment protocol consisted of two phases
  - **Phase One:** a browser based preference rating of images depicting real and simulated radar data.
  - **Phase Two:** a retrospective recall and justification of decisions made. Of all participants to complete Phase One, 20% were selected at random to complete Phase Two.
- Due to current circumstances, the protocol was completed remotely
  - Phase One via a browser based form that was generated uniquely for each participant.
- We worked with the Chief Forecaster to design the study and to adjust the type of assessments made.



# Protocol (phase one): preference ranking



Please provide your ranking. \*

	Most preferred (Ranked ...)	2nd	Least preferred (Ranke...
Prediction A	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Prediction B	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Prediction C	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[OPTIONAL] Please type any notes about your observations or reasoning below.

Long-answer text

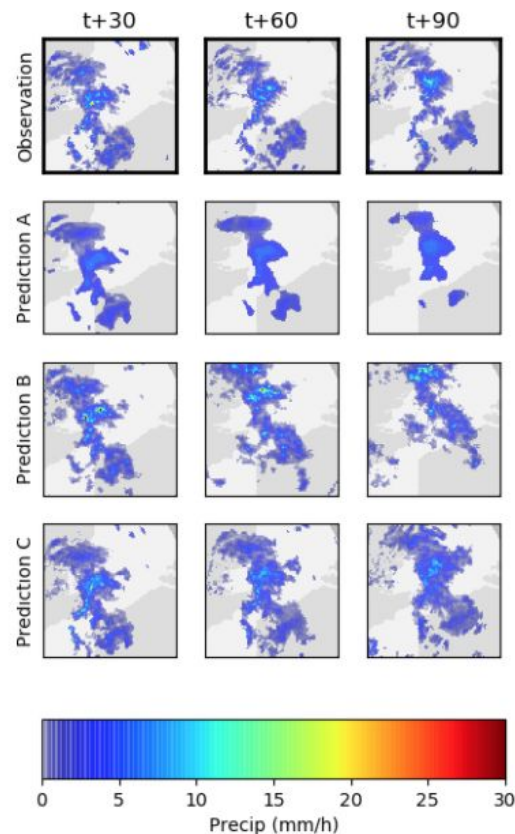
.....

Design decision to limit additional information to control variables.

# Case Selection

Participants were shown a total of 23 cases. Comprising

- 10 cases where precipitation levels were  $>5\text{mm/hour}$ .
- 10 cases where precipitation levels were  $>10\text{mm/hour}$ .
- 3 'special' cases, selected by an expert meteorologist. These cases featured severe or unusual weather events during the past year. (Included in the paper)

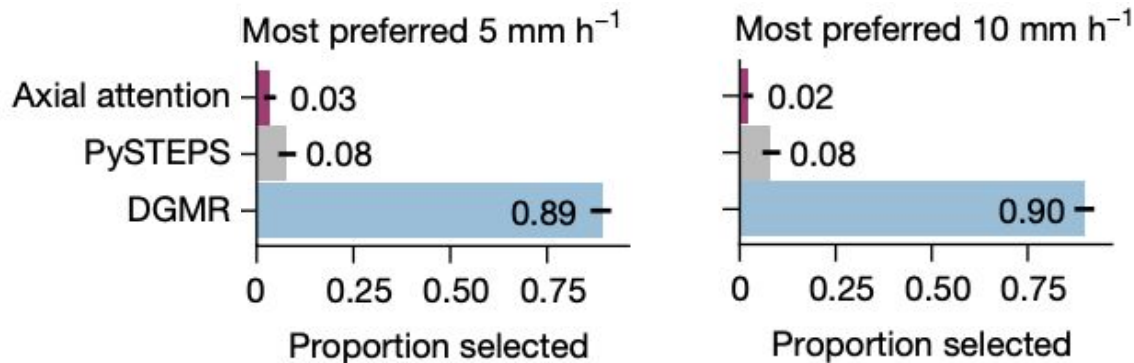


# Design Decisions

- **Question phrasing:** The language used within the preference ranking question can influence the way in which participants interact with the study. We opted to phrase the question in terms of 'preference is based on [their] opinion of accuracy and value'.
- **Additional Information:** In an operational setting, forecasters would often use multiple sources of information to make decisions. We opt to show radar based plots only, with limited additional information, to ensure we can limit and control variables.
- **Participant Selection Criteria:** Participants were selected based on the following criteria; must be a professional UK meteorologist, qualified for 6+ months

# Results

- Key factor: ability to capture well the extent of rain.
- Results contrast with the qualitative verification scores:
  - that did not differentiate between competing approaches
  - that often favoured models that blur the forecast over time



# Transcript Extracts

- “I would prefer the model to underdo intensities but get a much better spatial variation.”
- “I like things to look slightly realistic even if they’re not in the right place so that I can put some of my own physics knowledge into it.”
- “Lower resolution and unrealistic ones can still be useful, but a lot of the time (Axial Attention) didn’t even get the shape right”
- “Anything close to reality is really useful, any that track movement is good.”
- “This looks much higher detail (sic) compared to what we’re used to at the moment. I’ve been really impressed with the shapes compared with reality. I think they’re probably better than what we’re currently using. The shapes in particular, some of them do look really high resolution (sic).”

# Transcript Extracts

- "I would prefer the model to underdo intensities but get a much better spatial variation."
- "I like things to look slightly realistic even if they're not in the right place so that I can put some of my own physics knowledge into it."
- "Lower resolution and unrealistic ones can still be useful, but a lot of the time (Axial Attention) didn't even get the shape right"
- "Anything close to reality is really useful, any that track movement is good."
- "This looks much higher detail (sic) compared to what we're used to at the moment. I've been really impressed with the shapes compared with reality. I think they're probably better than what we're currently using. The shapes in particular, some of them do look really high resolution (sic)."



# Transcript Extracts

- "I would prefer the model to underdo intensities but get a much better spatial variation."
- "I like things to look slightly realistic even if they're not in the right place so that I can put some of my own physics knowledge into it."
- "Lower resolution and unrealistic ones can still be useful, but a lot of the time (Axial Attention) didn't even get the shape right"
- "Anything close to reality is really useful, any that track movement is good."
- "This looks much higher detail (sic) compared to what we're used to at the moment. I've been really impressed with the shapes compared with reality. I think they're probably better than what we're currently using. The shapes in particular, some of them do look really high resolution (sic)."

# Transcript Extracts

- "I would prefer the model to underdo intensities but get a much better spatial variation."
- "I like things to look slightly realistic even if they're not in the right place so that I can put some of my own physics knowledge into it."
- "Lower resolution and unrealistic ones can still be useful, but a lot of the time (Axial Attention) didn't even get the shape right"
- "Anything close to reality is really useful, any that track movement is good."
- "This looks much higher detail (sic) compared to what we're used to at the moment. I've been really impressed with the shapes compared with reality. I think they're probably better than what we're currently using. The shapes in particular, some of them do look really high resolution (sic)."

# Transcript Extracts

- "I would prefer the model to underdo intensities but get a much better spatial variation."
- "I like things to look slightly realistic even if they're not in the right place so that I can put some of my own physics knowledge into it."
- "Lower resolution and unrealistic ones can still be useful, but a lot of the time (Axial Attention) didn't even get the shape right"
- "Anything close to reality is really useful, any that track movement is good."
- "This looks much higher detail (sic) compared to what we're used to at the moment. I've been really impressed with the shapes compared with reality. I think they're probably better than what we're currently using. The shapes in particular, some of them do look really high resolution (sic)."

# Future for expert assessments

- Successful study and insight on:
  - how we run experimental psychology studies in Atmospheric Sciences
  - how we gain insight into meteorological decision mechanisms
- Operational-driven design
- Ensure that ML based products deliver value in an operational context



7

# Conclusion




# Final Thoughts


- **Nowcasting fills a gap in performance of NWP**s in the first 2 forecast hours, highlighting a role for ML fill gaps in existing physics-driven/simulation-based approaches.
- **Deep Generative models that improves upon the currently used nowcasting methods.**  
It is suitable for high-intensity events and is also able to provide the probabilistic ensembles required to estimate the evolution of chaotic systems.
- **Provides genuine decision-making value for use by real-world experts.**  
First deep learning model significantly preferred to an operational system by professional forecasters.
- **Interest to the many sectors and the public.** Many key environment and climate meetings this year, and we hope to add to those opportunities.




# Thank you!


Reach out: [ravuris@deepmind.com](mailto:ravuris@deepmind.com)


 master ▾ [deepmind-research](#) / nowcasting / Go to file

 **ravurisDM** and **diegolascasas** Added journal information to the citation info. ... 3257aa3 on Sep 30 History

..

 Open\_sourced\_dataset\_and\_model\_snapshot\_for\_pr... Adding nowcasting to deepmind-research repo. last month


 README.md Added journal information to the citation info. last month

 [README.md](#)

## Skillful Precipitation Nowcasting Using Deep Generative Models of Radar

This repository is a supplement to "Skillful Precipitation Nowcasting using Deep Generative Models of Radar" and provides necessary code for loading data from a large scale nowcasting dataset and obtaining predictions with the pretrained model.

Please see the Colab notebook for further details:

 [Open in Colab](#)

UK model, data, and Colab available at:

<https://github.com/deepmind/deepmind-research/tree/master/nowcasting>

## Parting Thought

“All models are wrong, but  
some are useful”

- George Box (Statistician)

## Parting Thought

“All ~~models~~ forecasts are  
wrong, but some are useful”

- We should think about how to better measure data-driven approaches!