Weather4cast at NeurIPS 2023 – Data Fusion for Quantitative Hi-Res Rain Movie Prediction under Spatio-temporal Shifts

Aleksandra Gruca* ALEKSANDRA.GRUCA@POLSL.PL
Silesian University of Technology, Poland

Xavier Calbet XCALBETA@AEMET.ES
Pilar Rípodas PRIPODASA@AEMET.ES
Llorenç Lliso JLLISOV@AEMET.ES
Agencia Estatal de Meteorología, AEMET

Federico Serva FEDERICO.SERVA@TERRARUM.EU
CNR-ISMAR, Italy

Bertrand Le Saux BLS@IEEE.ORG
European Space Agency Φ-lab, Italy

Lu Han HANLU@LAMDA.NJU.EDU.CN
Xu-Yang Chen CHENXY@LAMDA.NJU.EDU.CN
Han-Jia Ye YEHJ@LAMDA.NJU.EDU.CN
De-Chuan Zhan ZHANDC@LAMDA.NJU.EDU.CN
School of Artificial Intelligence, Nanjing University, China
National Key Laboratory for Novel Software Technology, Nanjing University, China

P. Jyoteeshkumar Reddy JYOTEESH.PAPARI@CSIRO.AU
CSIRO, Hobart, Australia

Harish Baki H.BAKI@TUDELFT.NL
TU Delft, Delft, The Netherlands

Sandeep Chinta SANDEEPC@MIT.EDU
Massachusetts Institute of Technology, USA

Pablo Izquierdo Ayala AYALA@BIO.LMU.DE
GSN-LMU, Munich, Germany

Akshay Punjabi AKPUN@ALUMNI.UPV.ES
Independent Researcher, Spain

Ajitabh Kumar AJITABH.KUMAR@ENRFLO.COM
EnrFlo, India

Rafael Pablos Sarabia RPABLOS@CS.AU.DK
Ira Assent IRA@CS.AU.DK
Aarhus University, Denmark

Joachim Nyborg JN@CORDULUS.COM
Morten Birk MB@CORDULUS.COM
Cordulus, Denmark

David P. Kreil* W4C23@KREIL.ORG
Vienna, Austria

Editors: Tao Qin, Kun Zhang, Jes Frellsen

* Corresponding Author

© 2026 A. Gruca *et al.*

Abstract

Now in its third year, *Weather4cast* 2023 again advanced modern algorithms in AI and machine learning through a highly topical interdisciplinary challenge at the NeurIPS Competition Track. Predicting hi-res rain radar movies from multi-band satellite sensors required data fusion of complementary signal sources and video frame prediction with super-resolution techniques. This year, quantitative prediction of rain rates was assessed using a balanced multi-class metric. Accurate rain predictions are ever more critical for everyone, with climate change increasing the frequency of extreme precipitation events. *Weather4cast* models thus have a particular impact for the many regions on Earth where costly weather radar data are not available. We here present highlights and insights emerging from the competition. To reward models extracting relevant mechanistic patterns reflecting the underlying complex weather systems, our evaluation tested their responses to spatio-temporal shifts. Algorithms needed to forecast 4 h or 8 h of ground-based hi-res precipitation radar from lo-res satellite spectral images in a unique cross-sensor prediction challenge, and were evaluated within and across a variety of regions on Earth with diverse climate and different distributions of heavy precipitation events. Robustness to distributional shifts over time was achieved by testing predictions on data one year after the training period. Notably, the capability of models to correctly capture spatio-temporal patterns qualitatively varied considerably and was not always in line with leaderboard scores. This suggests that more sophisticated metrics will be required to accurately reflect the effectiveness of AI models in corresponding meteorological applications. *Weather4cast* will continue to explore the expansive data set and advance benchmark challenges in the field *en route* to validated foundation models for weather.

1. Introduction

The *Weather4cast* competition at *NeurIPS 2023* has again led to advanced new models from Artificial Intelligence and Machine Learning superseding the state of the art in a high-resolution weather prediction challenge. This year's interdisciplinary competition challenge remains both timely and relevant: The **quantitative prediction of high-resolution rain radar movies from multi-band satellite sensors**. It required the fusion of data from complementary signal sources and the prediction of 'rain movie' video frames using super-resolution techniques. In order to reward models that are capable of extracting pertinent mechanistic patterns which reflect the underlying complex weather systems, our evaluation methodology incorporated spatio-temporal shifts. Specifically, the algorithms were required to forecast 4 to 8 hours of ground-based high-resolution precipitation radar from low-resolution satellite spectral images, thereby presenting a unique cross-sensor prediction challenge. The models were evaluated both within and across various regions on Earth, with these patches representing diverse climates and different distributions of heavy precipitation events. To test robustness over time, the models' predictions were tested on data collected one year post the training period. This approach ensured that the models were not only accurate but also adaptable to evolving weather patterns and conditions.

We here present details of the *Weather4cast* 2023 competition, compare and discuss the performance of submitted models, and provide an outlook for future editions of the competition series.

1.1. Background and Impact

The *Weather4cast* competition series aims to investigate and advance AI and machine learning algorithms for tasks such as generalized video prediction under spatio-temporal shifts. We specifically frame the problem of weather forecasting as a movie prediction task, with the predicted movie frames representing high-resolution rainfall rates. This approach has proven surprisingly effective in related competitions that we have run at NeurIPS and elsewhere in the recent past (Kreil et al., 2020; Herruzo et al., 2021; Gruca et al., 2022; Eichenberger et al., 2022). Models are assessed for their generalization capability under spatio-temporal domain shifts, aiming to identify robust models that extract the most physically meaningful patterns.

Now in its third year, the competition builds on our previous successes, moving from binary classification to more quantitative predictions. This is particularly challenging in this highly unbalanced data scenario: While imbalanced data is a common challenge for classifiers, quantitative cases have received much less attention.

Participants are tasked with predicting rain fall rates as measured by high-resolution rain radar from low-resolution satellite movies only. Resultin models thus are of direct value for weather forecasts in economically challenged regions where expensive radar coverage is limited or absent. Requiring hi-res predictions from lo-res inputs, the competition thus includes a super-resolution task. It provides the first benchmark in quantitative forecasting of high-resolution rain radar movies from multi-spectral satellite image data.

1.2. Related Work

After we demonstrated the success of framing dynamic forecasts as movie frame prediction tasks for the NeurIPS 2019 competition in other domains (Kreil et al., 2020), Google Research applied this concept to rain forecasts (Agrawal et al., 2019; Sønderby et al., 2020). Subsequently, MetNet-2 (Espeholt et al., 2022) managed to outperform current physics-based operational models using Convolution Neural Networks to predict rainfall at high spatial and temporal resolution. However, accurate precipitation forecasts beyond 4 hours remained challenging. Incorporating data from satellites, radar, 942 US weather stations, and ground rain gauges MetNet-3 (Andrychowicz et al., 2023) extended predictions to other weather variables and boasted a farther prediction horizon at high resolutions. Besides integrating a diverse range of sensor data it also employed numerical weather prediction models for the initial state.

Recently, first teams have been claiming foundation model capabilities, *i. e.*, showing emergent capabilities from pre-training after fine-tuning. ClimaX extends the Transformer architecture with tokenization and aggregation blocks (Nguyen et al., 2023). After self-supervised pre-training on low-resolution data from physical models it can be fine-tuned for a range of climate and weather prediction tasks, including variables and spatio-temporal scales unseen during pre-training – and this has been demonstrated for medium scale resolutions and forecast horizons (*ca* 15 km and 24h resolutions).

The compiled high-resolution training data for MetNet-2, however, was never shared. Also the training data and the latest learned MetNet-3 model were not made available to the scientific community. While public datasets like SEVIR (Veillette et al., 2020) provide both satellite and hi-res radar observations they only cover the USA. An independent reference

benchmark for modern high-resolution spatio-temporal models is therefore required outside the USA.

2. The Weather4cast 2023 Competition

2.1. Competition Data

Weather4cast 2023 builds on the previous competitions in the series. For continuity, and to encourage the community to create new models inspired by the best solutions from the previous years, we continue to use the exact same extensive dataset.

Competition data was compiled from two sources: EUMETSAT Meteosat Second Generation SEVIRI data (Shcmetz et al., 2002) and weather radar data from OPERA, the'Operational Programme for the Exchange of Weather Radar Information' of the EUMETNET project (Huuskonen et al., 2014; Saltikoff et al., 2019). The satellite data comes in the form of 3712×3712 images representing 11 sensor channels with narrow spectral filters. The target data constitute images representing instantaneous rain rate composites calculated from OPERA radar reflectivity values. Both are provided in 15 minute intervals.

OPERA radar and MSG SEVIRI data feature different resolutions and different geographical projections. The projection for raw OPERA data is Lambert Azimuthal Equal Area with each pixel corresponding to 2000×2000 m, whereas MSG data are in geostationary projection, where the pixel size varies depending on distance from the sub-satellite point. The area covered by an MSG pixel increases from 3000×3000 m at the sub-satellite point to a non-square area with one side > 24 km over Iceland. To facilitate the superresolution prediction required in the competition we reprojected OPERA data to a geostationary grid, as described in details in (Gruca et al., 2022).

2.2. Competition Task

Last year participants predicted rain/no-rain binary events (Gruca et al., 2022) based on thresholding rain rates. This year we instead asked for quantitative predictions of rain rates in mm/hour, which is more directly relevant to rain forecasts on short time scales (nowcasting).

Competition participants are asked to predict the amount of rainfall for a sequence of 4–8 hours in 16–32 time slots of 15 mins each from an input sequence of 4 time slots of the preceding hour. The input data are satellite images representing satellite radiances, covering visible and near infrared (VIS), thermal infrared (IR), and a water vapor absorption band (WV). In addition static variables such as latitude, longitude and elevation are available fore each region. The input sequence of four images of 252×252 pixels covers a wider context area surrounding the *region of interest* of 42×42 pixels for which predictions have to be made. The required output data are *high resolution* OPERA rain rates, where the ground truth in the *region of interest* is represented by an image of size 252×252 pixels, *i. e.*, at a $6\times$ higher resolution than the input data.

For the competition task we selected 10 regions across Europe that represented a range of precipitation regimes. We divided those regions into two groups: 7 *known* and 3 *new* regions. For seven of these regions extensive training data were made available, so they were well *known* to the models. Models were also tested on three *new* regions. We provided data for

the *known* regions for two years, 2019 and 2020. Model performance was tested using three leaderboards: **core**, **nowcasting**, and **transfer learning**. The **core** leaderboard task was to predict 8 h into the future for *known* regions and on withheld test sequences from the same years used for training. For the **nowcasting** task the evaluation period was reduced to 4 h. The **transfer learning** leaderboard tests the generalization ability of models. We evaluate performance in the *new* regions for the years 2019 and 2020 (spatial shifts), in *known* regions for the year 2021 (temporal shifts), and in *new* regions for the years 2021 (spatio-temporal shifts). More details can be found in (Gruca et al., 2022).

As in previous years, a starter kit[1] with a data loader and a simple baseline model based on a modified version of a 3D U-Net with a regression loss was provided to participants.

2.3. Metric

Finding a suitable metric for rainfall prediction is challenging because 'no rain' is by far the most frequent observation and heavy rain is rare, yielding an extremely skewed distribution of rain rates. Popular metrics include the Critical Success Index (CSI), Intersection over Union (IoU), and the Root Mean Squared Error (RMSE). CSI and IoU are defined for binary classifiers. For rain events, $CSI = TP/(TP + FN + FP)$, gives the rate of correct rain event predictions over the sum of all correctly predicted rain events (true positives, TP) plus the incorrect rain predictions (false negatives, FN, and false positives, FP). Despite its wide use in the community, the CSI is dependent upon the type of events considered (Schaefer, 1990b). Like the Positive Predictive Value (PPV) it is sensitive to the real rain event prevalence (Schaefer, 1990a; Mbizvo and Larner, 2024). The IoU is the Jaccard Index. Both do not allow a decomposition of scores by subregions, impeding parallel computation and statistics (Gruca et al., 2022; Kreil et al., 2020). While the RMSE does not suffer these limitations, adjusting an RMSE computation to equally weight errors at different rain rates requires an expensive calculation of these weights. While these could be used to assess predictions it would be more costly in a loss function. After some testing we therefore decided this year to assess models for their average performance at predicting rain events at different intensities. This is still cheap to compute while dealing with the extremely skewed rain rate distribution.

We compute the CSI for multiple thresholds: 0.2, 1, 5, 10 and 15 mm/h. Any precipitation below 0.2 mm/h can be considered as 'no rain' due to uncertainties in the reference OPERA rain radar data and artefacts introduced by their re-projection, 1 mm/h was chosen as threshold for typical light rain, 5 and 10 mm/h cover moderate rain, and 15 mm/hour or more correspond to heavy rain. Notably, qualitative prediction properties like getting the location and time of rainfall right, are here assessed by averaging across all locations and times.

3. Results, Model Highlights and Insights

Links to source code and learned parameters for all the models described in this section are provided in the Appendix.

1. https://github.com/agruca-polsl/weather4cast-2023

3.1. Temporal Frame Interpolation

Team SandD contributed the Temporal Frame Interpolation (TFI) Model which significantly enhances the robustness and accuracy of precipitation forecasting by creating synthetic samples between adjacent frames of satellite and ground radar data. The TFI model, incorporating a novel Multi-Level Dice (ML-Dice) loss function (Han et al., 2023), achieved first place in all the leaderboards of the Weather4cast'23 competition.

The model employs a 2D U-Net architecture (Ronneberger et al., 2015), which was chosen over the initial 3D U-Net (Çiçek et al., 2016) due to its faster convergence and superior performance in our testing phases. This change was instrumental in enhancing model robustness and computational efficiency. Input images are specifically cropped to 126×126 to maintain focus on relevant features without overloading the model with redundant data. They also constrain the model to output a small 42×42 patch and use a simple upsampling strategy to restore the 252×252 prediction. During training, Temporal Frame Interpolation (TFI) was used to augment the training dataset by interpolating between adjacent frames, which helps the model better understand the dynamics between sequential satellite images and reduces overfitting. Apart from TFI, specific geometric augmentations such as vertical and horizontal flips are employed to both input and target images to enhance the diversity. In addition, a Multi-Level Dice (ML-Dice) loss takes the ordinal relationship between rainfall levels into account, achieving better forecasting results.

In all, the integration of TFI and ML-Dice loss in the training process allowed the model to effectively predict rainfall by understanding the nuanced differences between various rainfall intensities. The code and trained weights for the model are publicly available for further research and development.

3.2. 2D U-Net

Team RainAI proposed utilizing the original 2D U-Net (Ronneberger et al., 2015), due to its numerous successes, with residual connections to mitigate the vanishing gradient issue in deep neural network (Sarabia et al., 2023). To address the absence of direct temporal management, the approach stacks time on top of channels, processing input dimensions of $T \times C \times H \times W$ as $T * C \times H \times W$. Similarly, the output of the 2D U-Net of size $T' * C' \times H \times W$ is unstacked to $T' \times C' \times H \times W$, with T' representing forecasted timesteps and C' the output channels for each timestep.

Importance sampling (Ravuri et al., 2021) balances the dataset, crucial for training efficiency and improving model accuracy considering the highly unbalanced dataset. Dataset preparation includes incorporating static data like latitude, longitude, and topological height for each position, while input patches are cropped to provide just the required spatial context for an 8-hour prediction.

The team explored an alternative cross-entropy loss function for precipitation nowcasting (Espeholt et al., 2021), enhancing performance over standard mean squared error and enabling probabilistic outputs. This classification approach categorizes rain intensities into predefined classes, facilitating informative probabilistic outputs while also supporting intensity forecasts when required.

The introduction of crop and upsample optimization, inspired by previous work (Espeholt et al., 2021), allows the U-Net architecture to focus on future lead times without sacrificing

target region resolution. NinaSR, a sophisticated super-resolution architecture, yields the best results for upsampling. However, additional work may be required to optimize the architecture and loss function.

3.3. PAUNet: attention-based U-Net architecture

Team CFG proposed a solution (Reddy et al., 2023a) for precipitation prediction from satellite radiance data using an attention-based U-Net architecture (PAUNet). The submitted approach primarily focused on data preprocessing, model architecture, and loss function. In data preprocessing, all input multi-band satellite and radar precipitation data were converted to Zarr format chunking, which enabled rapid loading of selected samples using CPUs in parallel as part of a Dask cluster. Since the distribution of radar precipitation data is highly skewed with many no-rain samples, only samples with at least rain in 100 grid points over the selected time sequence were considered, based on previous studies (Park et al., 2022). Further, the precipitation data was log-transformed and normalized with a maximum (considered here as 128, as per previous studies), i.e., $\log_{10}(1+x)/\log_{10}(1+128)$, and then clipped to the range 0-1. All values below 0.2 were transformed to zero value due to radar data artifacts, as suggested by previous studies (Gruca et al., 2022). However, most values lie in the range 0.2-2. To ensure the training encompassed all categories of precipitation (0.2-1, 1-5, 5-10, 10-15, and >15), the team used the Focal Precipitation Loss (Huang et al., 2023) with exponential weighting based on the precipitation rate, which was defined as Exponential Focal Precip Loss (e-FPL). Additionally, to focus on the center of the domain, a center crop in the model architecture was applied to extract critical contextual spatial features, based on previous studies (Gruca et al., 2022; Li et al., 2022; Park et al., 2022). Furthermore, the architecture included an attention layer to better capture the relationship between extracted encoded spatial features (Vaswani et al., 2017). The model was also quantitatively evaluated using a frequency-based metric, the 2D spatial Fast Fourier Transform (Reddy et al., 2023b). PAUNet exhibited good performance, surpassing the official leaderboard baseline scores by notable margins of 10% in the core, 18% in nowcasting, and 25% in transfer learning challenges.

3.4. NowcastNet and Swin UNETR

Team Ernflo submitted two different models: NowcastNet and Swin UNETR to forecast precipitation. NowcastNet (Zhang et al., 2023; Kumar, 2023b) has two main modules: an evolution network which learns underlying physics using neural operators, and a generative network which makes final prediction using context from the evolution network. Noise projector is added in generative network which transforms the latent vector into required spatial size. Thus, the model includes a deterministic evolution network parameterized by ϕ and a stochastic generative network parameterized by θ. Nowcasting procedure is thus physics-conditional generation from latent random vectors z, given as:

$$P(\hat{X}_{1:T}|Y_{-T_0:0},\phi;\theta) = \int P(\hat{X}_{1:T}|Y_{-T_0:0},\phi(Y_{-T_0:0}),z;\theta)P(z)dz \tag{1}$$

where $Y_{-T_0:0}$ denotes given past satellite images and $\hat{X}_{1:T}$ denotes predicted precipitation image.

Evolution network is a two path U-Net with shared evolution encoder for learning context representation. Motion decoder learns the two-dimensional velocities for $x-$ and $y-$directions, while intensity decoder learns intensity residuals for the desired times. Spectral normalization technique is applied for each convolutional layer. Input and output values are concatenated on temporal dimension for skip connection.

Swin UNETR (Hatamizadeh et al., 2022; Kumar, 2023a) architecture contains two main modules: a swin transformer-based encoder and a CNN-based decoder. Both are connected in a U-shaped network via skip connections at different resolutions. Attention mechanism in swin transformer is quite effective as it uses hierarchical windows of multiple sizes to handle different scales within feature set. $3D$ patches at one level is shifted in the next level, thus learning multi-scale contextual representations and long-range dependencies. Encoder has 4 stages, each with 2 transformer blocks for regular and shifted partitioning. Extracted feature representations of the encoder are used in the CNN-based decoder via skip connection at each resolution. Both models are competitive with respect to the base UNet model.

3.5. SmaAt-UNet: Small Attention U-Net

Final model submitted by Team Akpun has an architecture based on a Small Attention U-Net (SmaAt-UNet)(Trebing et al., 2021; Punjabi and Ayala, 2021, 2023), an efficient variant of the widely used U-Net(Ronneberger et al., 2015). It replaces standard convolutions with depthwise separable convolutions and incorporates channel attention (CBAM) to emphasize informative features(Woo et al., 2018). The encoder contains 5 stages with max pooling for downsampling, while the decoder upsamples to recreate the input resolution.

To handle the imbalanced rain distribution, non-rainy samples below a threshold were filtered out, the input satellite images to 126×126 pixels were center-cropped to reduce training cost, and the 126×126 output predictions were upsampled to 252×252 for evaluation using bilinear interpolation. The water vapor bands were also removed as they contributed little to performance.

The team started with the baseline SmaAt-UNet architecture and trained per region and year, taking $\sim$2 hours per model. They also experimented with creating an ensemble across regions but found the independent regional approach performed best. Other architectural variations like XCAttention (Ali et al., 2021) and translator blocks did not improve performance. Models were trained using the AdamW optimizer with a learning rate of 1e-3, batch size 32, and early stopping on validation loss. Training only required two NVIDIA T4 GPUs and already converged after 10 epochs.

The efficient model achieved a CSI score of 0.05157, outperforming the 0.04822 baseline using $\sim$4.1M parameters compared to $\sim$22M in the baseline, at a fraction of the training time and cost. This highlights the need to balance model performance with computational/environmental efficiency.

3.6. Qualitative Evaluation of Selected Model Predictions

Here we provide some qualitative analysis of the model predictions, represented by an example illustrating the evolution of a precipitation system moving from the West to the North-East in the *boxi_0015* region on September 2019 at 03:00 UTC. In particular, we focus on the predictions by two teams, SandD, highest ranked on the leaderboard, and Akpun, who score

Team	core (8h)	nowcasting (4h)	transfer-learning
SandD	1	1	1
Ernflo	2	2	2
RainAI	3	3	**4**
Ernflo_t2	4	4	**3**
CFG	—	5	5
Akpun	—	6	6
ALI_BDIL	**disqualified**, no model weights published and does not respond		
FIT-CTU	**withdrawn**, no conference presentation		

Table 1: Consolidated leaderboard ranks: ErnFlo entered a 2nd model, 'swin-UNETR,' labelled 't2'

lower. Figure 1 shows the qualitative comparison of the ground truth to the predictions for these two selected models. Both teams correctly predict the rain starting from the middle left side of the image in the first hour, with precipitation then moving to the top part of the image in the second hour with the maximum amount of rainfall in the 3rd and 4th hours. For the second part of the prediction period Team SandD correctly captures spatial patterns of precipitation moving across the area from the left bottom corner to the top right. The model of team Akpun only forecasts the first 4 h (nowcasting).

3.7. Comparative discussion

Four teams out of five used a 2D U-Net, a reliable standard architecture. One team submitted two different models, NowcastNet and Swin-UNETR. NowcastNet (Zhang et al., 2023) has two main modules: an evolution network which learns underlying physics using neural operators, and a generative network which makes the final prediction using context from the evolution network. Comparing the best-performing models allows some insights into the mechanisms successful for the prediction and super-resolution tasks. As observed by Team SandD, rapidity of convergence is likely to be a decisive benefit: one of the main challenges to overcome is the size of the data to ingest, so a fast model may have benefits. Another aspect, considering the use of a U-Net like is the importance of the superresolution contributed for which convolutions remain a popular technique. the inventive mechanisms that the various teams employed to enhance their models. There are data-centric approaches, such as the different data augmentation strategies utilized by Teams RainAI and SandD. The Temporal Frame Interpolation (TFI) contributed to the winning entry by team SandD. Given the highly-skewed distribution of extreme-weather events, re-balancing by importance sampling for instance may make sense. Model-centric tricks included attention mechanisms, present in the transformer encoder of the UNETR of team Enrflo, but also in the U-Net variants of

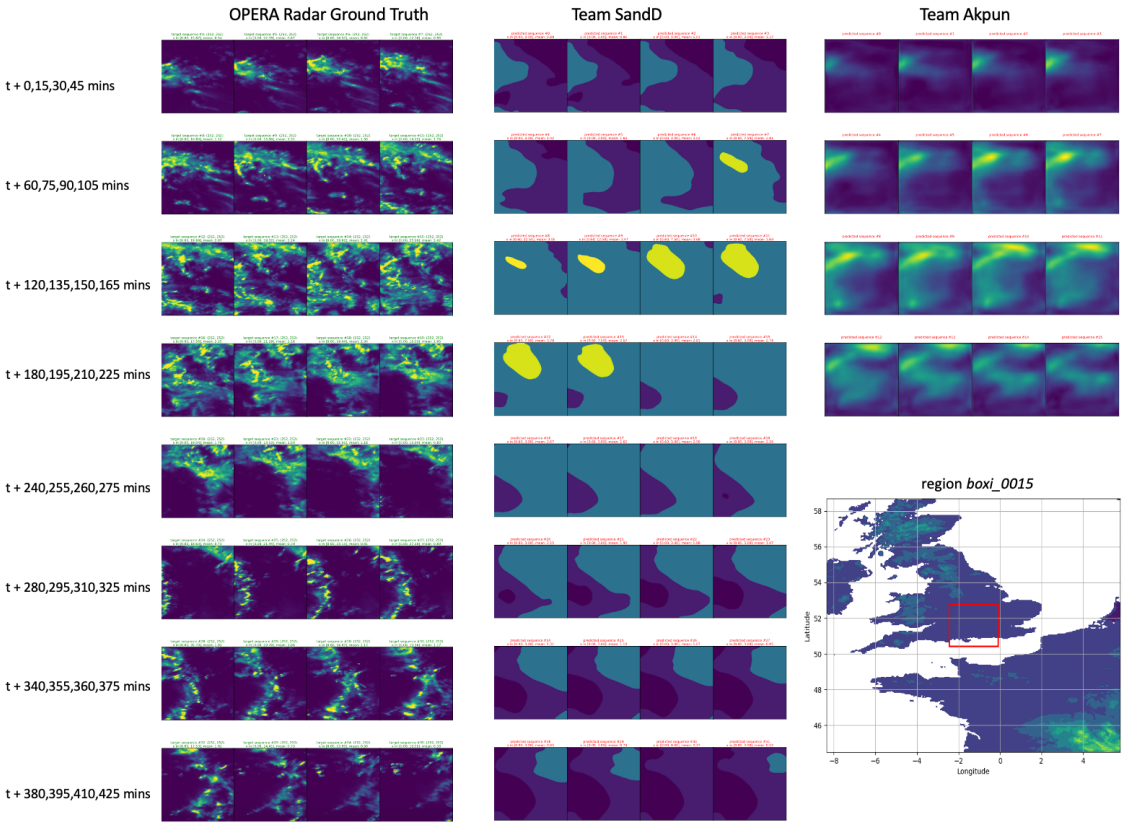


Figure 1: An example of the qualitative comparison of the ground truth to predictions of selected models. The left group shows the ground truth, the middle group shows the 8 h prediction by team SandS, and the right group shows the 4 h prediction by team Akpun. Note that team SandD directly predicts and plots rain classes instead of actual rain rates. Each frame corresponds to a future 15 min slot, where the 16 (32) frames correspond to 4 (8) hours. The bottom right insert shows the relevant region *boxi_0015*. This entire region forms the larger context area provided as input to the model. The central red square highlights the central *region of interest* for which predictions are made. Images are aligned so that the top is North.

teams CFG and Akpun. Loss engineering can also often be key to success, and a Multi-Level Dice loss (team SandD) and an Exponential Focal Precipitation Loss (team CFG) seems suitable for the nature of the problem.

4. Discussion and Outlook

Model performance on all three tasks was in general aligned (8 h rainfall prediction, 4 h rainfall now-casting, 4 h rainfall transfer learning), with the exception of the Swin UNETR (Ernflo_t2) submission doing better in transfer learning (Rank 2, Table 1). The other

submission by the same team Enrflo successfully applied NowcastNet to the *Weather4cast* challenge tasks, a recent model demonstrating advances in extreme event forecasting (Zhang et al., 2023), which must be commended for publishing both model and model weights, a critical good practice that is regrettably absent for other leading models 'published' in the field. Team SandD contributed a Temporal Frame Interpolation Model, which significantly enhanced the robustness and accuracy of precipitation forecasting by creating synthetic samples between adjacent frames of satellite and ground radar data, while incorporating a novel Multi-Level Dice loss. Their submission wins this year's competition and scored better than all state-of-the-art models available. It did particularly well in the transfer learning challenge and also captured the qualitative weather patterns very well.

While most models captured some patterns of precipitation and the broad movements in their predictions over time, only some models provided predictions that could give valuable insights into qualitative precipitation patterns that might actually be employed in meteorological applications. In line with earlier observations for state-of-the-art models (Ayzel et al., 2020; Ravuri et al., 2021), forecast patterns were blurred / smoothed out. Progressing further along forecast time, precipitation patterns became more and more diffuse. Even though that makes the forecasts seem less realistic, this actually does correctly reflect the fact that predictions are necessarily becoming less precise as they run further into the future. A lack of detailed dynamics exhibited by some models may actually be a consequence of models getting trained to minimize a loss that averages over time and space. While it is known that atmospheric conditions evolve in a stochastic manner, this materializes fast for precipitation, whereas deterministic short-term prediction is feasible for some other atmospheric variables, such as temperature. This overtly stochastic nature of rainfall in practice justifies probabilistic approaches to precipitation forecasting that some models and metrics directly support.

One notable conclusion from the observations reported here is that rankings obtained from training models on our rain prediction task to maximize an average CSI score, even if balanced, does not reflect the models' capability of qualitatively capturing spatio-temporal patterns. This justifies an exploration of other task / metric combinations to also capture these performance aspects critical to relevant metereological applications.

Considering first claims to foundation models with emerging capabilities for weather / climate (Chen et al., 2023), the search for models featuring the ability to perform well for multiple tasks or by multiple criteria has become a highly topical research area. Just last month, Microsoft Research described Aurora, their 'Foundation model of the Atmosphere' (Bodnar et al., 2024), that combines 3D-Perceiver encoder/decoders with a 3D-Swin-Transformer U-Net pre-trained by RMSE to predict a 6h lead for a variety of atmospheric data with a range of resolutions. Interestingly, pre-training also incorporated expansive datasets generated by numerical physical models. This is in line with our observation from this and last year's competitions that data driven models which also incorporate physical knowledge tend to score well. For Aurora such knowledge is likely absorbed in pre-training. The model could then be fine-tuned to outperform state-of-the-art alternatives for the prediction of pollution and weather on *different* scales (resolutions of 10–40 km and 24h for multi-day forecasts).

Although code has been released for foundation-class models, pre-trained weights have not been made available, and recreating these requires considerable manual and computational efforts. It is therefore still unclear how and to what degree these reported successes can

be transferred to hi-res now-casting (1 km and 15 min resolutions for multi-hour forecasts), or other meteorologically relevant downstream tasks that test emerging model capabilities. There is hence a need now for a high-resolution multi-task benchmark challenge for weather foundation models, supported by solid public baselines, which we will explore in *Weather4cast* at NeurIPS 2024 (www.weather4cast.net).

References

Shreya Agrawal, Luke Barrington, Carla Bromberg, John Burge, Cenk Gazen, and Jason Hickey. Machine Learning for Precipitation Nowcasting from Radar Images. *arXiv:1912.12132 [cs, stat]*, December 2019. URL http://arxiv.org/abs/1912.12132. arXiv: 1912.12132.

Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. *Advances in neural information processing systems*, 34:20014–20027, 2021.

Marcin Andrychowicz, Lasse Espeholt, Di Li, Samier Merchant, Alexander Merose, Fred Zyda, Shreya Agrawal, and Nal Kalchbrenner. Deep learning for day forecasts from sparse observations, 2023.

G. Ayzel, T. Scheffer, and M. Heistermann. Rainnet v1.0: a convolutional neural network for radar-based precipitation nowcasting. *Geoscientific Model Development*, 13(6):2631–2644, 2020. doi: 10.5194/gmd-13-2631-2020. URL https://gmd.copernicus.org/articles/13/2631/2020/.

Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan Weyn, Haiyu Dong, Anna Vaughan, Jayesh K. Gupta, Kit Tambiratnam, Alex Archibald, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. Aurora: A foundation model of the atmosphere, 2024.

Shengchao Chen, Guodong Long, Jing Jiang, Dikai Liu, and Chengqi Zhang. Foundation models for weather and climate data understanding: A comprehensive survey, 2023.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In Sébastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gözde B. Ünal, and William M. Wells III, editors, *MICCAI*, volume 9901 of *Lecture Notes in Computer Science*, pages 424–432, 2016.

Christian Eichenberger, Moritz Neun, Henry Martin, Pedro Herruzo, Markus Spanring, Yichao Lu, Sungbin Choi, Vsevolod Konyakhin, Nina Lukashina, Aleksei Shpilman, Nina Wiedemann, Martin Raubal, Bo Wang, Hai L. Vu, Reza Mohajerpoor, Chen Cai, Inhi Kim, Luca Hermes, Andrew Melnik, Riza Velioglu, Markus Vieth, Malte Schilling, Alabi Bojesomo, Hasan Al Marzouqi, Panos Liatsis, Jay Santokhi, Dylan Hillier, Yiming Yang, Joned Sarwar, Anna Jordan, Emil Hewage, David Jonietz, Fei Tang, Aleksandra Gruca, Michael Kopp, David Kreil, and Sepp Hochreiter. Traffic4cast at neurips 2021 – temporal

and spatial few-shot transfer learning in gridded geo-spatial processes, 2022. URL https://arxiv.org/abs/2203.17070.

Lasse Espeholt, Shreya Agrawal, Casper Sønderby, Manoj Kumar, Jonathan Heek, Carla Bromberg, Cenk Gazen, Jason Hickey, Aaron Bell, and Nal Kalchbrenner. Skillful Twelve Hour Precipitation Forecasts using Large Context Neural Networks, November 2021. URL http://arxiv.org/abs/2111.07470. arXiv:2111.07470 [physics].

Lasse Espeholt, Shreya Agrawal, Casper Sønderby, Manoj Kumar, Jonathan Heek, Carla Bromberg, Cenk Gazen, Rob Carver, Marcin Andrychowicz, Jason Hickey, Aaron Bell, and Nal Kalchbrenner. Deep learning for twelve hour precipitation forecasts. *Nature Communications*, 13(1):5145–, 2022. ISSN 20411723. doi: 10.1038/s41467-022-32483-x. URL https://doi.org/10.1038/s41467-022-32483-x.

Aleksandra Gruca, Federico Serva, Llorenç Lliso, Pilar Rípodas, Xavier Calbet, Pedro Herruzo, Jiří Pihrt, Rudolf Raevskyi, Petr Šimánek, Matej Choma, Yang Li, Haiyu Dong, Yury Belousov, Sergey Polezhaev, Brian Pulfer, Minseok Seo, Doyi Kim, Seungheon Shin, Eunbin Kim, Sewoong Ahn, Yeji Choi, Jinyoung Park, Minseok Son, Seungju Cho, Inyoung Lee, Changick Kim, Taehyeon Kim, Shinhwan Kang, Hyeonjeong Shin, Deukryeol Yoon, Seongha Eom, Kijung Shin, Se-Young Yun, Bertrand Le Saux, Michael K Kopp, Sepp Hochreiter, and David P Kreil. Weather4cast at neurips 2022: Super-resolution rain movie prediction under spatio-temporal shifts. In Marco Ciccone, Gustavo Stolovitzky, and Jacob Albrecht, editors, *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pages 292–313. PMLR, 28 Nov–09 Dec 2022.

Lu Han, Xu-Yang Chen, Han-Jia Ye, and De-Chuan Zhan. Learning robust precipitation forecaster by temporal frame interpolation. *arXiv preprint arXiv:2311.18341*, 2023.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. *arXiv preprint arXiv:2201.01266*, 2022.

Pedro Herruzo, Aleksandra Gruca, Llorenç Lliso, Xavier Calbet, Pilar Rípodas, Sepp Hochreiter, Michael Kopp, and David P. Kreil. High-resolution multi-channel weather forecasting – First insights on transfer learning from the Weather4cast Competitions 2021. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 5750–5757, December 2021. doi: 10.1109/BigData52589.2021.9672063.

Ping-Chia Huang, Yueh-Li Chen, Yi-Syuan Liou, Bing-Chen Tsai, Chun-Chieh Wu, and Winston H. Hsu. Stamina (spatial-temporal aligned meteorological information attention) and fpl (focal precip loss): Advancements in precipitation nowcasting for heavy rainfall events. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 854–863, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701245.

A. Huuskonen, E. Saltikoff, and I. Holleman. The operational weather radar network in europe. *Bulletin of the American Meteorological Society*, 95:897–907, 2014. doi: https://doi.org/10.1175/BAMS-D-12-00216.1.

David P. Kreil, Michael K. Kopp, David Jonietz, Moritz Neun, Aleksandra Gruca, Pedro Herruzo, Henry Martin, Ali Soleymani, and Sepp Hochreiter. The surprising efficiency of framing geo-spatial time series forecasting as a video prediction task – Insights from the IARAI Traffic4cast Competition at NeurIPS 2019. In *NeurIPS 2019 Competition and Demonstration Track*, pages 232–241. PMLR, August 2020. URL http://proceedings.mlr.press/v123/kreil20a.html. ISSN: 2640-3498.

Ajitabh Kumar. Precipitation nowcasting with spatial and temporal transfer learning using swin-unetr. *arXiv preprint arXiv:2312.00258*, 2023a.

Ajitabh Kumar. Skilful precipitation nowcasting using nowcastnet. *arXiv preprint arXiv:2311.17961*, 2023b.

Yang Li, Haiyu Dong, Zuliang Fang, Jonathan Weyn, and Pete Luferenko. Super-resolution probabilistic rain prediction from satellite data using 3d u-nets and earthformers. *arXiv preprint arXiv:2212.02998*, 2022.

Gashirai K. Mbizvo and Andrew J. Larner. On the dependence of the critical success index (csi) on prevalence. *Diagnostics*, 14(5), 2024. ISSN 2075-4418. URL https://www.mdpi.com/2075-4418/14/5/545.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. Climax: A foundation model for weather and climate, 2023.

Jinyoung Park, Minseok Son, Seungju Cho, Inyoung Lee, and Changick Kim. Rainunet for super-resolution rain movie prediction under spatio-temporal shifts. *arXiv preprint arXiv:2212.04005*, 2022.

Akshay Punjabi and Pablo Izquierdo Ayala. Efficient spatio-temporal weather forecasting using u-net. *arXiv preprint arXiv:2112.06543*, 2021.

Akshay Punjabi and Pablo Izquierdo Ayala. Efficient baseline for quantitative precipitation forecasting in weather4cast 2023. *arXiv preprint arXiv:2311.18806*, 2023.

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, Rachel Prudden, Amol Mandhane, Aidan Clark, Andrew Brock, Karen Simonyan, Raia Hadsell, Niall Robinson, Ellen Clancy, Alberto Arribas, and Shakir Mohamed. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, September 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03854-z. URL https://www.nature.com/articles/s41586-021-03854-z.

P Jyoteeshkumar Reddy, Harish Baki, Sandeep Chinta, Richard Matear, and John Taylor. Paunet: Precipitation attention-based u-net for rain prediction from satellite radiance data. *arXiv preprint arXiv:2311.18306*, 2023a.

P Jyoteeshkumar Reddy, Richard Matear, John Taylor, Marcus Thatcher, and Michael Grose. A precipitation downscaling method using a super-resolution deconvolution neural network with step orography. *Environmental Data Science*, 2:e17, 2023b.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

E. Saltikoff, G. Haase, L. Delobbe, N. Gaussiat, M. Martet, D. Idziorek, H. Leijnse, P. Novák, M. Lukach, and K. Stephan. Opera the radar project. *Atmosphere*, 10,320, 2019. doi: https://doi.org/10.3390/atmos10060320.

Rafael Pablos Sarabia, Joachim Nyborg, Morten Birk, and Ira Assent. Rainai – precipitation nowcasting from satellite data, 2023.

Joseph T. Schaefer. The critical success index as an indicator of warning skill. *Weather and Forecasting*, 5(4):570 – 575, 1990a. doi: 10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/wefo/5/4/1520-0434_1990_005_0570_tcsiaa_2_0_co_2.xml.

Joseph T. Schaefer. The Critical Success Index as an Indicator of Warning Skill. *Weather and Forecasting*, 5(4):570–575, December 1990b. ISSN 1520-0434, 0882-8156. doi: 10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/wefo/5/4/1520-0434_1990_005_0570_tcsiaa_2_0_co_2.xml.

Johannes Shcmetz, Paolo Pili, Stephan Tjemkes, Johan Kerkmann, Sergio Rota, and Alain Ratier. An introduction to meteosat second generation (msg). *Bulletin of the American Meteorological Society*, 83(7):977–992, 2002.

Casper Kaae Sønderby, Lasse Espeholt, Jonathan Heek, Mostafa Dehghani, Avital Oliver, Tim Salimans, Shreya Agrawal, Jason Hickey, and Nal Kalchbrenner. MetNet: A Neural Weather Model for Precipitation Forecasting. *arXiv:2003.12140 [physics, stat]*, March 2020. URL http://arxiv.org/abs/2003.12140. arXiv: 2003.12140.

Kevin Trebing, Tomasz Stanczyk, and Siamak Mehrkanoon. Smaat-unet: Precipitation nowcasting using a small attention-unet architecture. *Pattern Recognition Letters*, 145: 178–186, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Mark Veillette, Siddharth Samsi, and Chris Mattioli. Sevir : A storm event imagery dataset for deep learning applications in radar and satellite meteorology. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22009–22019. Curran Associates, Inc., 2020.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

Yuchen Zhang, Mingsheng Long, Kaiyuan Chen, Lanxiang Xing, Ronghua Jin, Michael I. Jordan, and Jianmin Wang. Skilful nowcasting of extreme precipitation with nowcastnet. *Nature*, 619:526—532, 2023.

Appendix

List of the repositories with source code of the described models and trained parameters:

Team SandD

Temporal Frame Interpolation – https://github.com/Secilia-Cxy/UNetTFI
Detailed description of the solution: https://arxiv.org/abs/2311.18341
Training: A single GPU was used with batch size eight, running models for a total of 90 epochs. The initial learning rate was established at 0.0001, with a weight decay set at 0.02 and a dropout rate maintained at 0.4. The optimization was managed via AdamW, featuring a learning rate scaled down by a factor of 0.9 whenever the validation loss exceeded that of the previous epoch. Furthermore, an early stopping mechanism was enforced to halt training if the validation loss plateaued.
All experiments were executed on 4 Nvidia RTX 4090 GPUs, each with 32 GB VRAM.

Team RainAI

2D U-Net: https://github.com/rafapablos/w4c23-rainai
Detailed description of the solution: https://arxiv.org/abs/2311.18398
Training: 1 NVIDIA A10G for a maximum of 5 epochs with 15k samples (<20% of the available samples).

Team CFG

PAUNet: https://github.com/HarishBaki/PAUNet-Weather4cast-2023
Detailed description of the solution: https://arxiv.org/abs/2311.18306
Training: 80 Nvidia V100 GPUs, each with 32 GB memory.

Team Ernflo

NowcastNet: https://arxiv.org/abs/2311.17961
Swin UNETR: https://arxiv.org/abs/2312.00258
Detailed description of the solution: https://github.com/ajikmr/weather4cast-2023-submission
Training: Nvidia V100 GPU with 32GB RAM.

Team Akpun

SmaAt-UNet: https://github.com/Dauriel/w4c_repo
Detailed description of the solution: https://arxiv.org/abs/2311.18806
Training: 2 NVIDIA T4 GPUs with a maximum of 10 epochs.