Proceedings of Machine Learning Research 220:292–312, 2023 NeurIPS 2022 Competition Track

Weather4cast at NeurIPS 2022: Super-Resolution Rain Movie Prediction under Spatio-temporal Shifts

Aleksandra Gruca[*ab] ALEKSANDRA.GRUCA@POLSL.PL
Federico Serva[††††a] FEDERICO.SERVA@TERRARUM.EU
Llorenç Lliso[‡] JLLISOV@AEMET.ES
Pilar Rípodas[‡] PRIPODASA@AEMET.ES
Xavier Calbet[‡] XCALBETA@AEMET.ES
Pedro Herruzo[§] PEDRO.HERRUZO@UPC.EDU
Jiří Pihrt[¶] PIHRTJIR@FIT.CVUT.CZ
Rudolf Raevskyi[¶] RAEVSRUD@FIT.CVUT.CZ
Petr Šimánek[¶] PETR.SIMANEK@FIT.CVUT.CZ
Matej Choma[‖¶] MATEJ.CHOMA@METEOPRESS.CZ
Yang Li[**] YANGLI@NUIST.EDU.CN
Haiyu Dong[††] HAIYU.DONG@MICROSOFT.COM
Yury Belousov[‡‡] YURY.BELOUSOV@UNIGE.CH
Sergey Polezhaev[§§] SPOLEZHAEV@DUBFORMER.AI
Brian Pulfer[‡‡] BRIAN.PULFER@UNIGE.CH
Minseok Seo[¶¶] MINSEOK.SEO@SI-ANALYTICS.AI
Doyi Kim[¶¶] DOYIKIM@SI-ANALYTICS.AI
Seungheon Shin[¶¶] SHSHIN@SI-ANALYTICS.AI
Eunbin Kim[¶¶] EBKIM@SI-ANALYTICS.AI
Sewoong Ahn[¶¶] ANSE3832@SI-ANALYTICS.AI
Yeji Choi[¶¶] YEJICHOI@SI-ANALYTICS.AI
Jinyoung Park[***] JINYOUNGPARK@KAIST.AC.KR
Minseok Son[***] KSOS104@KAIST.AC.KR
Seungju Cho[***] JOYGA@KAIST.AC.KR
Inyoung Lee[***] INZERO24@KAIST.AC.KR
Changick Kim[***] CHANGICK@KAIST.AC.KR
Taehyeon Kim[***] POTTER32@KAIST.AC.KR
Shinhwan Kang[***] SHINHWAN.KANG@KAIST.AC.KR
Hyeonjeong Shin[***] HYEONJEONG1@KAIST.AC.KR
Deukryeol Yoon[***] DEUKRYEOL.YOON@KAIST.AC.KR
Seongha Eom[***] DOUBLEB@KAIST.AC.KR
Kijung Shin[***] KIJUNGS@KAIST.AC.KR
Se-Young Yun[***] YUNSEYOUNG@KAIST.AC.KR
Bertrand Le Saux[†††] BERTRAND.LE.SAUX@ESA.INT
Michael K Kopp[‡‡‡] MICHAEL.KOPP@IARAI.AC.AT
Sepp Hochreiter[‡‡‡§§§] SEPP.HOCHREITER@IARAI.AC.AT
David P Kreil[‡‡‡b] DAVID.KREIL@IARAI.ORG

Editors: Marco Ciccone, Gustavo Stolovitzky, Jacob Albrecht

Abstract

Weather4cast again advanced modern algorithms in AI and machine learning through a highly topical interdisciplinary competition challenge: The prediction of hi-res rain radar movies from multi-band satellite sensors, requiring data fusion, multi-channel video frame prediction, and super-resolution. Accurate predictions of rain events are becoming ever more critical, with climate change increasing the frequency of unexpected rainfall. The resulting models will have a particular impact where costly weather radar is not available. We here present highlights and insights emerging from the thirty teams participating from over a dozen countries.

To extract relevant patterns, models were challenged by spatio-temporal shifts. Geometric data augmentation and test-time ensemble models with a suitable smoother loss helped this transfer learning. Even though, in ablation, static information like geographical location and elevation was not linked to performance, the general success of models incorporating physics in this competition suggests that approaches combining machine learning with application domain knowledge seem a promising avenue for future research.

Weather4cast will continue to explore the powerful benchmark reference data set introduced here, advancing competition tasks to quantitative predictions, and exploring the effects of metric choice on model performance and qualitative prediction properties.

[*] Silesian University of Technology, Poland

[†] Italian Space Agency, Italy

[‡] AEMET/NWC SAF, Spain

[§] Polytechnic University of Catalonia, Spain

[¶] Faculty of Information Technology, Czech Technical University in Prague, Czech Republic

[‖] Meteopress s.r.o.

[**] Nanjing University of Information Science and Technology, China

[††] Microsoft, USA

[‡‡] University of Geneva, Switzerland

[§§] DubFormer, Netherlands

[¶¶] SI Analytics, Republic of Korea

[***] Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea

[†††] European Space Agency Φ-lab, Italy

[‡‡‡] Institute of Advanced Research in Artificial Intelligence, Austria

[§§§] Machine Learning Institute, Johannes Kepler University Linz, Austria

[a] Equal Contribution.

[b] Corresponding Author.

1. Introduction

The *Weather4cast* competition at NeurIPS 2022 has advanced modern algorithms in AI and machine learning through a highly topical interdisciplinary competition challenge: The **prediction of hi-res rain radar movies from multi-band satellite sensors**, requiring data fusion of complementary signal sources, multi-channel video frame prediction, as well as super-resolution techniques. To reward models that extract relevant mechanistic patterns reflecting the underlying complex weather systems our evaluation incorporates spatio-temporal shifts: Specifically, algorithms need to forecast 8 h of ground-based hi-res precipitation radar from lo-res satellite spectral images in a unique cross-sensor prediction challenge. Models are evaluated within and across regions on Earth, with these patches reflecting diverse climate and different distributions of heavy precipitation events. Robustness over time is achieved by testing predictions on data one year after the training period.

We here present highlights and lessons learned in the *Weather4cast* competition at NeurIPS 2022.

1.1. Background and Impact

The prediction of high-resolution rainfall can be restated as a movie prediction task. This allows for the application of modern computer vision algorithms to exploit spatio-temporal correlations. This works surprisingly well, as demonstrated by the NeurIPS competitions in 2019, 2020, and 2021 for urban traffic forecasts (Kreil et al., 2020; Kopp et al., 2021), subsequently by Google Research for rainfall (Agrawal et al., 2019; Sønderby et al., 2020), and also our first *Weather4cast* competitions forecasting multiple weather variables from satellites (Gruca et al., 2021; Herruzo et al., 2021). This research contributes to a topical trend of applying machine learning in the Earth sciences, competing with traditional physical or empirical models for accuracy and speed (Bonavita et al., 2021; Schneider et al., 2021).

Now in its third edition, *Weather4cast* 2022 has moved to improve rain forecasts worldwide with a completely new data set and advanced competition tasks. Accurate predictions of rain events are becoming ever more critical for everyone, with climate change increasing the frequency of unexpected rainfall. Notably, the new models and insights will have a particular impact for the many regions on Earth where costly weather radar data are not available.

1.2. Related Work

Convolutional Neural Networks in MetNet (Sønderby et al., 2020) and MetNet-2 (Espeholt et al., 2021) improved on physical models for 4 and 12 hour predictions. A panel of meteorologists preferred 89% of the predictions of a deep generative model (Ravuri et al., 2021). These successes are contrasted by a lack of access to the hi-res data used to train these models. More recent models based on Graph Neural Networks (Lam et al., 2022), Transformers (Bi et al., 2022), and U-Nets (Kaparakis and Mehrkanoon, 2023) were limited to public resources like the ECMWF ERA5 reanalysis archive (Hersbach et al., 2020) providing multiple variables at multiple vertical levels across all Earth, yet at low resolutions (1 h, ~ 30 km). This is in line with other typical datasets in the domain (Rasp et al., 2020; de Witt et al., 2020). CloudCast (Nielsen et al., 2021) provides 10 different cloud related variables at

Table 1: Characteristics of the SEVIRI instrument on board of the Meteosat Second Generation (MSG) satellites from EUMETSAT.

Channel Number	Central Wavelength (μm)	Spatial Resolution (km)	Spectral Zone Characteristic	Type of Channel
1	0.635	3	Solar Visible	Window (VIS)
2	0.81	3	Solar Visible	Window (VIS)
3	1.64	3	Solar Infrared	Window (VIS)
4	3.90	3	Solar/Thermal Infrared	Window (VIS/IR)
5	6.25	3	Thermal Infrared	H_2O Absorption (WV)
6	7.35	3	Thermal Infrared	H_2O Absorption (WV)
7	8.70	3	Thermal Infrared	Window (IR)
8	9.66	3	Thermal Infrared	O_3 Absorption (IR)
9	10.80	3	Thermal Infrared	Window (IR)
10	12.00	3	Thermal Infrared	Window (IR)
11	13.40	3	Thermal Infrated	CO_2 Absorption (IR)
12	Broad Band (0.4–1.1)	1	Visible/Infrared Solar	Window (VIS)

15 min and ~ 4 km resolutions. We extended this by 22 other more general variables in the *Weather4cast* 2021 dataset and benchmark (Herruzo et al., 2021).

While SEVIR (Veillette et al., 2020) provides both satellite and hi-re radar its coverage is limited to the U. S. A. It has been used to learn satellite-to-radar translation and radar-to-radar prediction, but not satellite-to-radar prediction. Recently, it was used by Generative Adversarial Networks to improve U-Net predictions (Hu et al., 2022).

To our knowledge *Weather4cast* now, for the first time, introduces raw spectral bands from satellite sensors and ground-based hi-res precipitation radar for a wide variety of geographical regions and different time periods, allowing a first satellite-to-radar forecasting benchmark.

2. The *Weather4cast* 2022 Competition

2.1. Data Sources

Meteosat Second Generation SEVIRI data. The European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) runs the Meteosat Second Generation (MSG) geostationary meteorological series of satellites. They have on board the SEVIRI – Spinning Enhanced Visible Infra-Red Imager – instrument (Shcmetz et al., 2002). It has twelve channels with which the Earth is observed in the visible and near infrared (VIS), thermal infrared (IR) and a water vapor absorption band (WV). The spatial resolution is about 3 km at nadir for the eleven channels with narrow spectral band filters. The characteristics of each of the spectral channels are shown in Table 1. Due to its geostationary nature, it is located on the celestial equator plane, at a particular longitude, repeatedly observing the entire Earth disk from a constant perspective. It generates images of 3712 × 3712 pixels for each of the eleven channels with narrow spectral filters. In its nominal mode, these images are generated every 15 minutes. The data used in this paper belongs to the satellite located at zero degrees longitude running in nominal mode. For illustration purposes, Fig. A-1 shows an 'Air mass' RGB composite image of this satellite.

Weather Radar data from OPERA project. Weather radar are widely used to measure precipitation. A weather radar covers a relatively large area, provides the 3D structure of the precipitation systems and allows their tracking. A radar network provides the possibility of covering a larger domain. Despite these advantages, radar precipitation measurements have several sources of errors. Amongst them are the errors associated to the broadening of the beam and the higher distance to the earth surface with the increasing distance from the radar site, echoes from non-meteorological targets, beam blockage on terrain (i.e., mountains), attenuation of the signal by rain (specially by heavy rain) and anomalous propagation of the beam in certain atmospheric conditions. A more detailed overview of the pros and cons of the weather radar for measuring precipitation in comparison to other sources of precipitation data as well as a a general picture of the current state of radar research is provided in (Sokol et al., 2021).

The characteristics of the precipitation radar makes a radar network the best option for a meteorological service to perform nowcasting and warning tasks. Radar data is often complemented with other sources like rain gauges and/or satellite data. In this work the radar data is provided as a reference and is considered as the "ground truth" for the precipitation field.

The radar data provided are 2D composites of the OPERA - 'Operational Programme for the Exchange of Weather Radar Information' of EUMETNET project (www.eumetnet.eu). OPERA produces 2D composites of instantaneous surface rain rate, instantaneous maximum reflectivity and 1 hour rainfall accumulation. For this work the instantaneous rain rates composites every 15 minutes from February 2019 to 2021 have been provided. The Marshall Palmer Z–R relationship (with coefficients $a = 200$ and $b = 1.6$) (Marshall et al., 1947) is used for converting radar reflectivity into precipitation intensity in the 2D composites. More details about the OPERA project can be found in (Huuskonen et al., 2014) and (Saltikoff et al., 2019).

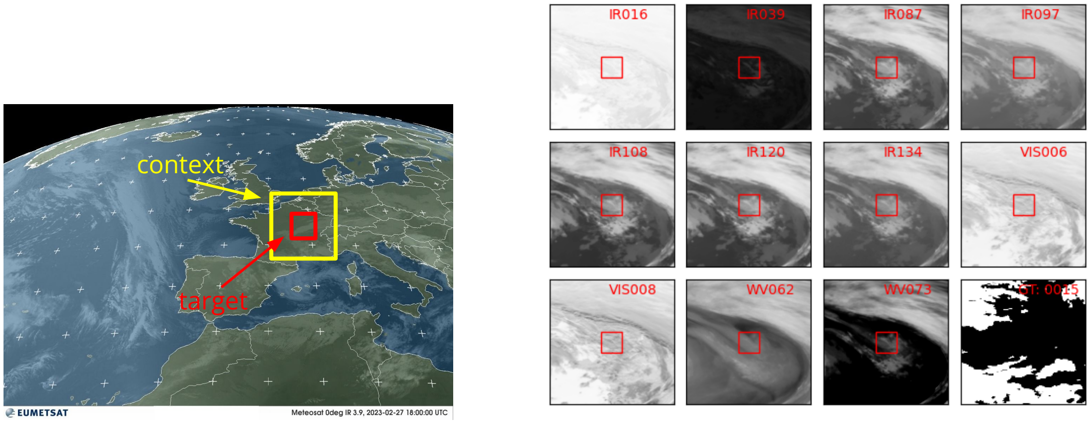
2.2. Data Compilation

The OPERA radar network data and the MSG SEVIRI data are in different geographical projections. The projection for raw OPERA data is Lambert Azimuthal Equal Area with a pixel size of 2000 × 2000 meters, this projection preserves the area with respect to the earth surface. On the other hand, the MSG data are in geostationary projection, in this projection the pixel size is bigger as pixels are farther away from the sub-satellite point. The area covered by a MSG pixel increases from 3000 × 3000 meters in the sub-satellite point to, for example, polygonal pixels with a side size larger than 24 km over Iceland.

To ease the training of models the OPERA data has been reprojected to a geostationary grid. With this, both OPERA and MSG data (2D image-like data) match geographically and can be combined. The detailed description of the reprojection schema is provided in section A.2 of the Appendix.

2.3. Dynamic Input and Static Data

Seasonality is a major source of variability of precipitation in the European domain (Zveryaev, 2004), but further modulation is driven by multiple local and remote climate patterns (Karagiannidis et al., 2008). Global warming due to anthropogenic activities is also expected to increase the severity of rainfall extremes over the continent (King and Karoly, 2017), therefore prediction of high-impact events is a very relevant task.



(a) Context and target areas

(b) Input radiances and binary ground truth

Figure 1: (a) Schematic representation of the spatial context (yellow), for which satellite radiances are provided, and the target region (red), where rainfall prediction are sought. (b) Snapshot of longitude-latitude maps for the eleven MSG band radiances for the context of patch 15 and OPERA binary mask ground truth (GT) with a threshold of 0.2 mm/hour (bottom right). In MSG images, dark means lower values, black indicates rain in the OPERA image.

For a given area of interest (size 252 × 252 OPERA reprojected to geostationary pixels), a wider context is provided. The context area covers a square with sides six times bigger than the area of interest, in the same projection, as shown in Fig. 1a. The relative size of the context is chosen in order to capture weather systems nearby the target patch which may bring rain to it in the following hours. In Fig. 1b we provide a visual comparison of satellite radiances and binary rainfall mask for OPERA for a given timestep. It is clear how the various bands provide different perspectives of the same scene, and how the binary mask is not trivially related to the radiance patterns.

Before defining target patches for the competition we wanted to characterize them based on monthly frequencies of rain events. Our goal was to ensure that the selected regions include rain events that are typically infrequent. A detailed description of the target patches selection process is provided in section A.3 of the Appendix. Location of the selected regions is given in Fig. A-4.

2.4. Tasks

Competition participants are given a task to predict rainfall events for the next 8 hours in 32 time slots from an input sequence of 4 time slots of the preceding hour. The input sequence consists of four 11-band spectral satellite images. These 11 channels represents satellite radiances covering VIS, WV, and IR bands with a small amount of random noise added. In addition static variables such as latitude, longitude and elevation are available.

Each satellite image covers a 15 minute period with the region of interest of the size 42×42 pixels in the satellite resolution surrounded by the spatial context of the size 252×252 pixels. The prediction output is a sequence of 32 images representing rain events from ground-radar reflectivities. Output images also have a temporal resolution of 15 minutes but have higher spatial resolution, with the image size of 252×252 pixels in the OPERA resolution covering the same region of interest given in the input sequence. So in addition to predicting the weather in the future, converting satellite inputs to ground-radar outputs, this adds a super-resolution task due to the coarser spatial resolution of the satellite data as one pixel in the satellite resolution corresponds to six pixels in OPERA resolution.

In addition, a starter kit with a data loader and a baseline were provided to the participants[*]. It contained all the necessary code to train and explore a modified version of a 3D variant of the U-Net.[†]

The challenge was organized in two stages. In Stage 1 (period: from August 1 to November 18, 2022), data for 2019 and three regions are provided to start model development and allow participants to test the baseline model. A public leaderboard was shared to give participants rapid dynamic feedback on their submissions in relation to the baseline and submissions by others. Stage 2, which took place from October 14 to November 18, 2022, consisted of two challenges: The Extended Core challenge incorporated the release of 2020 data for the Stage-1 regions as well as four additional regions for both 2019 and 2020. The Transfer Learning challenge provided data for 2021 for all regions, as well as three unknown locations, to test models capability for transfer learning. Public leaderboards for both Stage 2 challenges were made available. The final evaluation, however, was made on held out data which were kept undisclosed.

2.5. Metric

This year the Weather4cast competition has introduced a new challenge of precipitation prediction, and we have also provided new dataset. Therefore to simplify the problem we have asked participants to predict rain events only, instead of specifying the exact amount of rainfall. For that reason, the competition leaderboard metric is IoU (Intersection over Union), a common evaluation metric used in computer vision to measure the accuracy of object detection and segmentation models. It is calculated as the ratio of the area of overlap between the predicted and ground truth to the area of union between them.

Since rain events are rare, we decided to use a metric that specifically targets them. During the evaluation of submissions, we focused only on correctly predicted rain events (pixels). Finally, to obtain a single value for each submission, we calculated IoU values for each region separately and then averaged obtained values across all regions. The detailed information on calculating the final metric and rain rates threshold is provided in section A.4 of the Appendix.

[*] Weather4Cast starter toolkit: github.com/iarai/weather4cast-2022

[†] ELEKTRONN3 - Neural Network Toolkit: elektronn3.readthedocs.io

3. Results and Model Highlights

We scored over $1,600$ submissions by thirty teams from over a dozen countries with over 900 entries beating a simple 3D U-Net baseline. After a first pilot stage, about half as many submissions were received in the larger comprehensive competition. About 100 submissions were filed to the dedicated transfer-learning leaderboard. About 60 models were finally scored on held-out test datasets. Teams were invited to submit brief research papers for presentation at the NeurIPS conference; see the competition website www.weather4cast.org. Highlights from the best or most interesting models are presented below.

3.1. WeatherFusionNet

The model developed by the team of *FIT-CTU* is designed to estimate rainfall from satellite data. It consists of three separate modules, each designed to process the data from a different angle (Pihrt et al., 2022).

The first module, called sat2rad, is a U-Net that is trained to estimate rainfall from a single satellite frame, allowing it to extract information about the current rain situation without having to predict the future.

The second module is a recurrent convolutional network called PhyDNet (Guen and Thome, 2020), which is designed to disentangle physical dynamics from other complementary visual information. Its architecture consists of two branches. The first branch is responsible for the physical dynamics and features a recurrent physically constrained cell called PhyCell, which performs PDE-constrained prediction in latent space. The second branch extracts other residual information, such as visual appearance and details, using ConvLSTM cells. This module was intended to be trained on radar frames, but due to limitations in the data, it was trained only on satellite data in this case. PhyDNet's role in the model is to extend the input sequence of satellite frames, with a limited output sequence length of 10.

Finally, the outputs from the sat2rad and PhyDNet modules are concatenated with the input sequence and fused by another U-Net to generate the final prediction. The prediction covers a large area, but only the center part is needed, so the prediction is cropped and upscaled for the final output. The upscale operation uses simple bilinear interpolation and there is room for further research for improvement. The code and trained parameters are publicly available.[‡]

3.2. Model Ensemble for Probabilistic Rain Prediction

Team *meteoai* presents a solution (Li et al., 2022) for probabilistic rain prediction using the model ensemble method from the baseline 3D U-Net and EarthFormer (Gao et al., 2022) models based on multi-channel satellite measurements. The team focused on data preprocessing, training strategy, and post-processing instead of modifying model structure to maximize the performance of the baseline models. For data preprocessing, considering large synoptic-scale context can carry useful circulation information for precipitation prediction in advance hours, particularly for heavy precipitation, and a characteristic synoptic-scale motion of 10 m/s (Holton, 1973; Pan et al., 2019) and storm-motion velocity of 15 m/s (Wapler, 2021), the center cropping is performed to crop the input image size by half, discarding

[‡] github.com/Datalab-FIT-CTU/weather4cast-2022

redundant information in the input satellite images to ensure the model focuses on more important context information improving the model's performance. In addition, the combined loss functions (IoUDice and IoUDiceFocal) were introduced to cope with rain and no rain class imbalance. Finally, considering the differences in rain characteristics across different regions, a probability threshold optimization method was introduced to search for the optimal probability threshold for rain prediction in each region to classify rain or no rain pixels. With just data preprocessing, combined loss functions and post-processing, the approach achieved second place in stage 2. The source code and trained model weights are available online.[§]

3.3. Vision Transformers for Weather4cast

The approach of team *team-name* (Belousov et al., 2022) is based on Vision Transformers (Dosovitskiy et al., 2020), where the input can be represented as either a volumetric image with time as a depth dimension, or as a video, specifically, a SWIN-UNETR Transformer (Hatamizadeh et al., 2022) for 3D medical images and a VIVIT (Arnab et al., 2021) model for video input. The team also introduces a set of configurations that can be applied to enhance results for various models as well as baseline-specific improvements. The findings reveal that utilizing half-precision and gradient checkpointing during training does not compromise performance while reducing GPU memory requirements. Furthermore, the choice of loss function was found to be of critical importance, regardless of the underlying model architecture. Interestingly, the results indicate that optimizing for the test metric (IoU) leads to inferior performance compared to optimizing for BCE.

Consistent with prior years, the authors report that combining multiple trained models yields the most competitive results. The majority voting algorithm, which combines their top models, achieved a tie for 3rd place in the final competition, indicating that utilizing transformer-based approaches for weather forecasting constitutes a potentially valuable area of research warranting further exploration. The code and corresponding model have been made publicly accessible online.[¶]

3.4. Simple Baseline for Weather Forecasting Using Spatiotemporal Context Aggregation Network

Team *SI-Analytics* proposed a SImple baseline for weather forecasting using spatiotemporal context Aggregation Network (SIANet) (Seo et al., 2022b) and training strategy (Seo et al., 2022a). SIANet is an end-to-end model composed only of CNNs to which network decomposition technology is applied. It consists of Large Context Aggregation (LCA) and Spatiotemporal Refinement Module (STR), and has the same shape as U-Net. LCA is an element that composes all CNNs blocks of SIANet and has the same structure as inception blocks, but the amount of computation is less compared to inception blocks because network decomposition technology is applied. STR serves to refine the output of SIANet through spatio-temporal modeling, and it exploits the fact that weather patterns show strong spatio-temporal correlations.

In addition, SIANet has a different strategy from the general training strategies used by existing weather forecasting frameworks. It is motivated by the fact that performance

[§] github.com/bugsuse/weather4cast-2022-stage2

[¶] github.com/bruce-willis/weather4cast-2022 is degraded when the same training strategy is used because weather data has different characteristics from general computer vision recognition task data. SIANet introduces a data augmentation strategy that considers wind direction, a smoother loss that considers spatio-temporal correlation, and a test-time geometric augmentation ensemble that performs inverse augmentation again during inference. As a result of applying these technologies, SIANet achieved first place in stage 1 and stage 3 of the W4C22 challenge, and achieved third place in stage 2.

3.5. RainUnet

KAIST-CILAB (Park et al., 2022) presents a hierarchical U-shaped network, RainUnet, that utilizes the Temporal-wise Separable block (TS block). This block helps capture interframe correlations by decomposing the standard 3D convolution into spatial and temporal components, increasing the receptive field and enabling the network to learn long-range spatio-temporal dependencies.

In addition, various data preprocessing strategies are utilized to further enhance the capabilities of RainUNet. The authors conducted a toy experiment and selected the bands composed of IR and VIS, as they produced the best results based on metrics such as IoU, precision, and accuracy. To balance the rain/no rain classes, the authors filtered the data sequences by removing non-rainy sequences where the count of positive pixels in 32 ground-truth future frames is less than 100. Lastly, the authors perform center cropping to improve the model's performance by allowing it to focus on the important context information for the target region. Crop size affects the prediction performance over time evolution of forecast. It is necessary to focus on close surrounding information when predicting the near future while utilizing a larger region is recommended for predicting the far future. Thus, it is crucial to exploit the appropriate context information for effective future frame prediction.

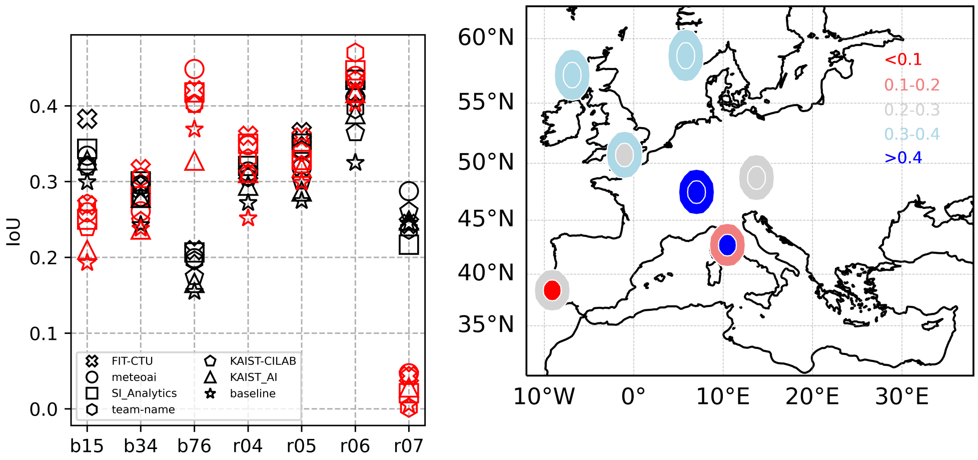
3.6. Region-Conditioned Orthogonal 3D U-Net

KAIST AI (Kim et al., 2022) proposes a simple yet efficient method that involves injecting region information into the feature maps during propagation. The architecture is a modified 3D U-Net architecture using a new Region Conditioned Network (RCN) that generates region-conditioned context. In specific, RCN takes a one-hot encoded categorical input to generate a region-conditioned context that is added to the feature maps in the encoder block. The authors demonstrate that this module helps to ensure that the model is able to better distinguish between different regions in the input images and capture regional differences in the segmentation output.

Additionally, the authors also introduce the use of orthogonal 1x1x1 convolution and residual units to help reduce redundancy in the filter response and capture more fine-grained features from the latent representations. The orthogonality is applied through the extension of spectral restricted isometry property (Kim and Yun, 2022). This property helps to preserve the magnitude of the propagation signal and improve the quality of the segmentation output. To further improve the performance, the authors also apply several training strategies, including mixup, self-distillation, and feature-wise linear modulation – FiLM (Perez et al., 2018).

Table 2: Top ranked teams and key features. **ex-aequo* – see Appendix A.5

Rank	Team	avg IoU	Preprocess	Ensemble	Physics-based	Transformer
1	FIT-CTU	.316	✓	×	✓	×
2	meteoai	.307	✓	✓	×	✓
3*	SI Analytics	.305	✓	×	✓	×
3*	TEAM-NAME	.300	×	✓	×	✓
4	KAIST-CILAB	.287	✓	×	×	×
5	KAIST-AI	.274	✓	×	×	×
-	Baseline	.254	×	×	×	×



(a) IoU in 2019 (black) and 2020 (red)

(b) Average IoU among top-ranked models

Figure 2: (a) IoU for each region and team in 2019 (black) and 2020 (red). (b) Averaged IoU for each year (outer circle, 2019; inner circle; 2020) and region of the core challenge.

4. Discussion and Outlook

In Table 2 we report the average IoU scores for the top ranking teams in an overview of the various features of the winning models in comparison to the simple baseline model. Dedicated preprocessing steps were adopted by most of the top performing models. They were either based on Earth observation domain knowledge or on standard machine learning techniques. For instance, standard data augmentation was adopted to mitigate the strong imbalance between the rain and no-rain classes. On the other hand, domain knowledge was exploited in application specific data augmentation, such as estimating physical characteristics like wind speed from cloud motion as an additional model input. Domain knowledge was also used to exclude one or several satellite bands from the VIS or WV channels as less relevant, or to discard extended context by cropping the input data. These procedures also improved the computational efficiency of the models, as less data was required in the training phase.

While team *meteoai* (Li et al., 2022) reported that the addition of static information such as geographical coordinates and elevation did not provide any substantial improvement the other top ranked teams did not report on similar tests. Interestingly, recent models for rain forecasting (Espeholt et al., 2021; Sønderby et al., 2020) do include static information yet report no ablation studies to test their relevance. Whether or not static information might already be encoded within the dynamics of weather patterns thus remains an open research question.

Several state-of-the-art machine learning techniques featured prominently in the top ranking models. For instance, transformers were chosen for spatio-temporal modelling in the second and third* best-ranking solutions. Ensemble models are well known to increase robustness and performances, and again proved to be efficient. On the other hand, incremental improvements of the baseline model could also already sufficiently increase prediction performance for joining the top ranking teams without any need for more complex architectures. Team *KAIST AI* (Kim et al., 2022), for instance, demonstrated the power of adding Feature-wise Linear Modulation (FiLM, Perez et al., 2018) layers, which alter the output of neural networks with an affine transformation applied to intermediate features.

The top-ranked models can be further evaluated in terms of the best and worst performances achieved across all available regions and over the different years tested, as shown in Fig. 2. The models presented here generally outperformed the baseline solution, with a few exceptions. For 2019, region r06, located in Central Europe, was the region with the best scores for all models, while region b76, located in the Tyrrhenian Sea between Corsica and Italy, was the most difficult one in 2019. Interestingly the hardest region for 2019 was the best or second-best performing one for all models in 2020, reflecting high year-to-year variability. The worst performing region for 2020, r07, was located in Southern Portugal. This region and also b76 are both located in southern Europe, and are characterized by dry and hot summers and higher probabilities of developing heavy thunderstorms (Merheb et al., 2016). Year-to-year climate variation was also significant, as the relative IoU differences between the best and the worst performing regions were about 50% in 2019, while in 2020 results for the most difficult region were ten/twenty times worse than for the easiest regions.

Interestingly, the core and transfer learning challenges had different winners. Most of the teams applied the same model to both challenges. Team *SI-Analytics*, however, approached the transfer learning challenge independently, applying geometric data augmentation and test-time ensemble models with a spatio-temporal smoother loss to better capture regional differences (Seo et al., 2022a), netting the team both the first position on the leaderboard and the special Transfer Learning Award by the Scientific Committee.

It is striking that models incorporating physics yielded the best and third best results in the core prediction challenge as well as the best performance in the transfer-learning challenge. The development of new classes of models combining the best of machine learning and application domain knowledge thus seem a promising avenue for future research.

In the coming year, *Weather4cast* will further explore the powerful benchmark reference data set introduced here, advancing competition tasks to quantitative predictions, and exploring the effects of metric choice on model performance and qualitative prediction properties.

Acknowledgments

We wish to thank the data providers EUMETSAT and the OPERA project from EUMETNET. We are also grateful to Annakaisa von Lerber at OPERA for her support.

References

Shreya Agrawal, Luke Barrington, Carla Bromberg, John Burge, Cenk Gazen, and Jason Hickey. Machine learning for precipitation nowcasting from radar images. arXiv preprint arXiv:1912.12132, 2019. URL https://arxiv.org/abs/1912.12132.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In Proceedings of the IEEE/CVF international conference on computer vision, pages 6836–6846, 2021.

Yury Belousov, Sergey Polezhaev, and Brian Pulfer. Solving the weather4cast challenge via visual transformers for 3d images, 2022. URL https://arxiv.org/abs/2212.02456.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast, 2022. URL https://arxiv.org/abs/2211.02556.

Massimo Bonavita, Rossella Arcucci, Alberto Carrassi, Peter Dueben, Alan J. Geer, Bertrand Le Saux, Nicolas Longépé, Pierre-Philippe Mathieu, and Laure Raynaud. Machine learning for earth system observation and prediction. Bulletin of the American Meteorological Society, 102(4):E710 – E716, 2021. doi: 10.1175/BAMS-D-20-0307.1. URL https://journals.ametsoc.org/view/journals/bams/102/4/BAMS-D-20-0307.1.xml.

Christian Schröder de Witt, Catherine Tong, Valentina Zantedeschi, Daniele De Martini, Freddie Kalaitzis, Matthew Chantry, Duncan Watson-Parris, and Piotr Bilinski. Rainbench: Towards global precipitation forecasting from satellite imagery. CoRR, abs/2012.09670, 2020. URL https://arxiv.org/abs/2012.09670.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. Journal of Machine learning research, 7(Jan):1–30, 2006.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

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. arXiv preprint arXiv:2111.07470, 2021. URL https://arxiv.org/abs/2111.07470v1.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. arXiv preprint arXiv:2207.05833, 2022. URL https://arxiv.org/abs/2207.05833.

Aleksandra Gruca, Pedro Herruzo, Pilar Rípodas, Andrzej Kucik, Christian Briese, Michael K. Kopp, Sepp Hochreiter, Pedram Ghamisi, and David P. Kreil. Cdceo'21 - first workshop on complex data challenges in earth observation. In Proceedings of the 30th

ACM International Conference on Information Knowledge Management, CIKM '21, page 4878–4879, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384469. doi: 10.1145/3459637.3482044. URL https://doi.org/10.1145/3459637.3482044.

Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11474–11484, 2020.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I, pages 272–284. Springer, 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.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The era5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730):1999–2049, 2020. doi: https://doi.org/10.1002/qj.3803. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803.

James R Holton. An introduction to dynamic meteorology. American Journal of Physics, 41 (5):752–754, 1973. doi: https://doi.org/10.1119/1.1987371.

Yuan Hu, Lei Chen, Zhibin Wang, Xiang Pan, and Hao Li. Towards a more realistic and detailed deep-learning-based radar echo extrapolation method. Remote Sensing, 14 (1), 2022. ISSN 2072-4292. doi: 10.3390/rs14010024. URL https://www.mdpi.com/2072-4292/14/1/24.

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.

Christos Kaparakis and Siamak Mehrkanoon. Wf-unet: Weather fusion unet for precipitation nowcasting, 2023. URL https://arxiv.org/abs/2302.04102.

A. Karagiannidis, A. Bloutsos, P. Maheras, and Ch. Sachsamanoglou. Some statistical characteristics of precipitation in Europe. Theor Appl Climatol, 91:193–204, 2008. doi: 10.1007/s00704-007-0303-7.

Taehyeon Kim and Se-Young Yun. Revisiting orthogonality regularization: A study for convolutional neural networks in image classification. IEEE Access, 10:69741–69749, 2022. doi: 10.1109/ACCESS.2022.3185621.

Taehyeon Kim, Shinhwan Kang, Hyeonjeong Shin, Deukryeol Yoon, Seongha Eom, Kijung Shin, and Se-Young Yun. Region-conditioned orthogonal 3d u-net for weather4cast competition, 2022. URL https://arxiv.org/abs/2212.02059.

Andrew D King and David J Karoly. Climate extremes in europe at 1.5 and 2 degrees of global warming. Environmental Research Letters, 12(11):114031, 2017. doi: 10.1088/1748-9326/aa8e2c.

Michael Kopp, David Kreil, Moritz Neun, David Jonietz, Henry Martin, Pedro Herruzo, Aleksandra Gruca, Ali Soleymani, Fanyou Wu, Yang Liu, Jingwei Xu, Jianjin Zhang, Jay Santokhi, Alabi Bojesomo, Hasan Al Marzouqi, Panos Liatsis, Pak Hay Kwok, Qi Qi, and Sepp Hochreiter. Traffic4cast at neurips 2020 - yet more on the unreasonable effectiveness of gridded geo-spatial processes. In Hugo Jair Escalante and Katja Hofmann, editors, Proceedings of the NeurIPS 2020 Competition and Demonstration Track, volume 133 of Proceedings of Machine Learning Research, pages 325–343. PMLR, 06–12 Dec 2021. URL https://proceedings.mlr.press/v133/kopp21a.html.

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.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Alexander Pritzel, Suman Ravuri, Timo Ewalds, Ferran Alet, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Jacklynn Stott, Oriol Vinyals, Shakir Mohamed, and Peter Battaglia. Graphcast: Learning skillful medium-range global weather forecasting, 2022. URL https://arxiv.org/abs/2212.12794.

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. URL https://arxiv.org/abs/2212.02998.

J.S. Marshall, R.C. Langille, and W.M.K. Palmer. Measurement of rainfall by radar. Journal of Meteorology, 4, 1947. doi: https://doi.org/10.1175/1520-0469(1947)004<0186:MORBR>2.0.CO;2.

Mohammad Merheb, Roger Moussa, Chadi Abdallah, François Colin, Charles Perrin, and Nicolas Baghdadi. Hydrological response characteristics of mediterranean catchments at different time scales: a meta-analysis. Hydrological Sciences Journal, 61(14):2520–2539, 2016. doi: 10.1080/02626667.2016.1140174.

A. H. Nielsen, A. Iosifidis, and H. Karstoft. Cloudcast: A satellite-based dataset and baseline for forecasting clouds. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, pages 1–1, 2021. doi: 10.1109/JSTARS.2021.3062936.

Baoxiang Pan, Kuolin Hsu, Amir AghaKouchak, and Soroosh Sorooshian. Improving precipitation estimation using convolutional neural network. Water Resources Research, 55(3):2301–2321, 2019. doi: https://doi.org/10.1029/2018WR024090.

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.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.

Jiří Pihrt, Rudolf Raevskiy, Petr Šimánek, and Matej Choma. Weatherfusionnet: Predicting precipitation from satellite data, 2022. URL https://arxiv.org/abs/2211.16824.

Stephan Rasp, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Soukayna Mouatadid, and Nils Thuerey. WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting. Journal of Advances in Modeling Earth Systems, 12(11):e2020MS002203, 2020. ISSN 1942-2466. doi: https://doi.org/10.1029/2020MS002203. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002203.

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, and et al. Skilful precipitation nowcasting using deep generative models of radar. Nature, 597(7878): 672–677, Sep 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03854-z. URL http://dx.doi.org/10.1038/s41586-021-03854-z.

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.

Rochelle Schneider, Massimo Bonavita, Alan Geer, Rossella Arcucci, Peter Dueben, Claudia Vitolo, Bertrand Le Saux, Begum Demir, and Pierre-Philippe Mathieu. Esa-ecmwf report on recent progress and research directions in machine learning for earth system observation and prediction. npj Climate and Atmospheric Science, 5(51):2397–3722, 2021. doi: 10.1038/s41612-022-00269-z. URL https://doi.org/10.1038/s41612-022-00269-z.

Minseok Seo, Doyi Kim, Seungheon Shin, Eunbin Kim, Sewoong Ahn, and Yeji Choi. Domain generalization strategy to train classifiers robust to spatial-temporal shift. arXiv preprint arXiv:2212.02968, 2022a.

Minseok Seo, Doyi Kim, Seungheon Shin, Eunbin Kim, Sewoong Ahn, and Yeji Choi. Simple baseline for weather forecasting using spatiotemporal context aggregation network. arXiv preprint arXiv:2212.02952, 2022b.

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.

Z. Sokol, J. Szturc, J. Orellana-Alvear, J. Popová, A. Jurczyk, and R Célleri. The role of weather radar in rainfall estimation and its application in meteorological and hydrological modelling—a review. Remote Sensing, 13, 351, 2021. doi: https://doi.org/10.3390/rs13030351.

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.

Mark Veillette, Siddharth Samsi, and Chris Mattioli. Sevir: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. Advances in Neural Information Processing Systems, 33:22009–22019, 2020.

Kathrin Wapler. Mesocyclonic and non-mesocyclonic convective storms in germany: Storm characteristics and life-cycle. Atmospheric Research, 248:105186, 2021. doi: https://doi.org/10.1016/j.atmosres.2020.105186.

Igor I. Zveryaev. Seasonality in precipitation variability over europe. Journal of Geophysical Research: Atmospheres, 109(D5), 2004. doi: https://doi.org/10.1029/2003JD003668.

Appendix A. Further Details on the *Weather4cast* 2022 Competition

A.1. Meteosat Second Generation SEVIRI data

In coordination with EUMETSAT, a small percentage of noise was added to the original Meteosat sensor data to allow re-distribution of the compiled dataset, with multiplicative noise best suited to the range of data considered. Sensitivity tests showed that the impact on the baseline predictions were negligible.

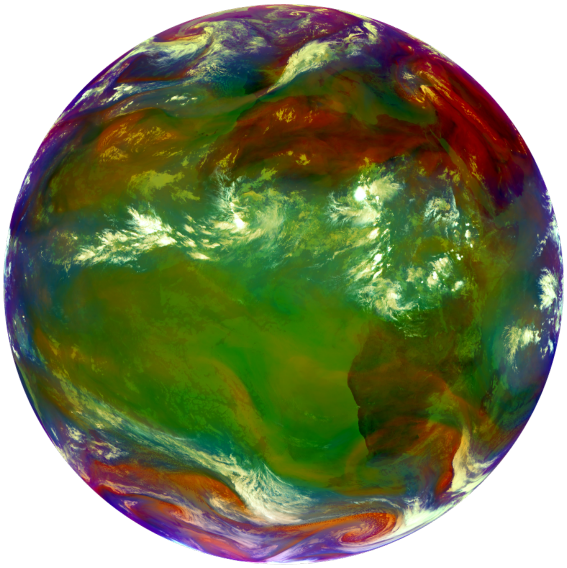


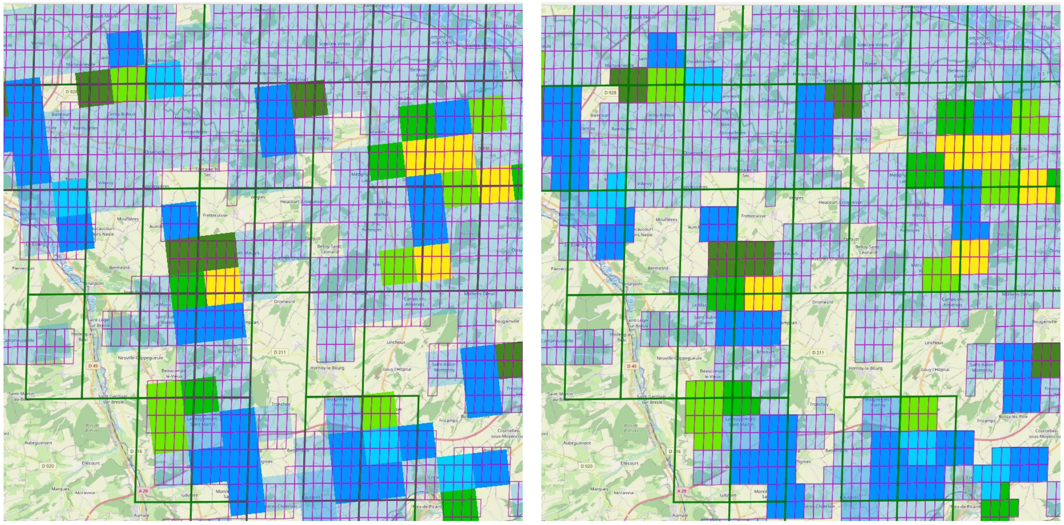
Figure A-1: 'Airmass' RGB composite using a combination of four channels (5, 6, 8 and 9) from SEVIRI. Obtained from MSG located at zero degrees longitude on August 20, 2019 at 10 UTC.

.

A.2. Data Compilation

In the reprojection process some information can be changed, this is due to the bigger and changing size of the MSG pixels with respect to the OPERA pixels. To cope with this underlying problem, a dense destination grid was chosen: each MSG pixel is divided into 36 smaller pixels, dividing each side of the satellite pixels by 6. Figure A-2 illustrates the reprojection, it corresponds to a scene of approximately 30 × 30 km near Amiens (France). Presenting the MSG grid, the OPERA final grid and the result of reprojecting the OPERA pixels to the final grid (outlined in cyan).

The size of the dense grid has been chosen evaluating the loss of information in a forward and backward reprojection of the OPERA data, the 6 × 6 chosen grid has a negligible loss even in the more unfavorable areas.
It should be emphasized that in the data provided for the competition, the MSG pixels are not divided into 6 × 6 smaller pixels, corresponding each MSG pixel with 6 × 6 reprojected OPERA pixels.



(a) The two grids superimposed over OPERA original data.

(b) OPERA reprojected to dense grid.

Figure A-2: Reprojection schema: Green lines outline MSG pixels, magenta lines outline the pixels of the destination grid (where OPERA data is reprojected), The colored pixels are the OPERA pixels.

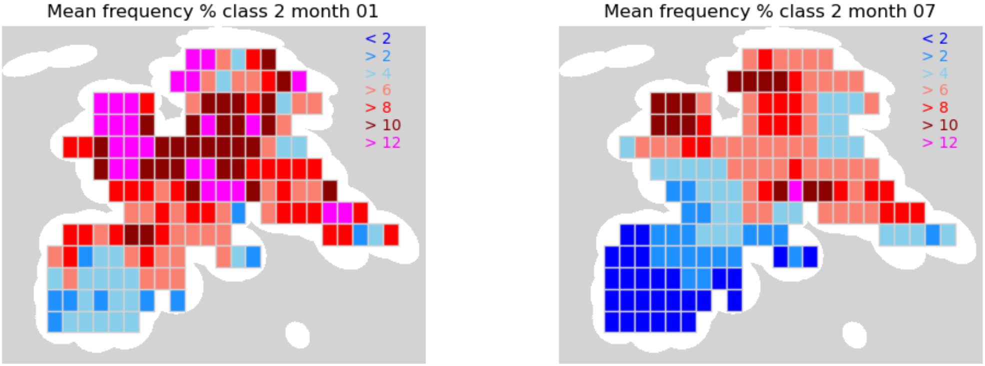
A.3. Target Patch Selection

To characterize possible target patches, we followed standard meteorological rain rates classification (World Meteorological Organization, 2018), similar to the approach of other works, such as Ravuri et al. (2021), and defined the following classes for OPERA derived rain rates: no rain between 0-0.1 mm/hour, low between 0.1-2.5 mm/hour, moderate for 2.5-7 mm/hour and heavy when above 7 mm/hour. This allowed us to compare different regions and to provide consistent splits for training, testing and the transfer challenge.

In particular, for each region of interest we calculated monthly frequencies of rain events according to four rain rates classification mentioned above from February 2019 until December 2021. Then we examined the number of rain events accumulated monthly during this time for each region focusing on regions with the highest amount of rain events. We noted that rain events belonging to moderate or heavy precipitation rate class are quite rare across the years and their monthly occurrences are highly variable over the year reflecting seasonal precipitation patterns. Finally, for the competition we picked regions as to balance between regions having more events with low rainfall rates and regions having less frequent but more intense rainfall rates.

Figure A-3 shows the frequency of occurrence of low rain rates across the OPERA domain in boreal winter and summer. In winter the probability of rain is relatively high (>12%) in many patches in the British Isles, Northern Europe, the Alps and Eastern Europe.

Dry conditions are dominant in boreal summer in southern Europe, with the exception of the Alpine region, and the frequency of rain events is lower also at higher latitudes. This comparison indicates that seasonal variability is very large in the European domain (Karagiannidis et al., 2008) and that the distribution between rain/no rain is unbalanced


Mean frequency % class 2 month 01
< 2
> 2
> 4
> 6
> 8
> 10
> 12
Mean frequency % class 2 month 07
< 2
> 2
> 4
> 6
> 8
> 10
> 12

(a) Low rain January

(b) Low rain July

Figure A-3: Longitude-latitude probability maps (%) for low rain rates measured by the OPERA network between 2019 and 2021 for the months of January (a) and July (b). Values are shown for square areas of the same size of the outputs to be predicted. Grey shading indicates areas outside the OPERA coverage.

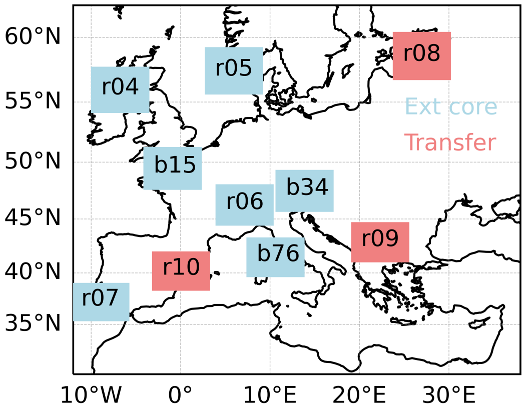


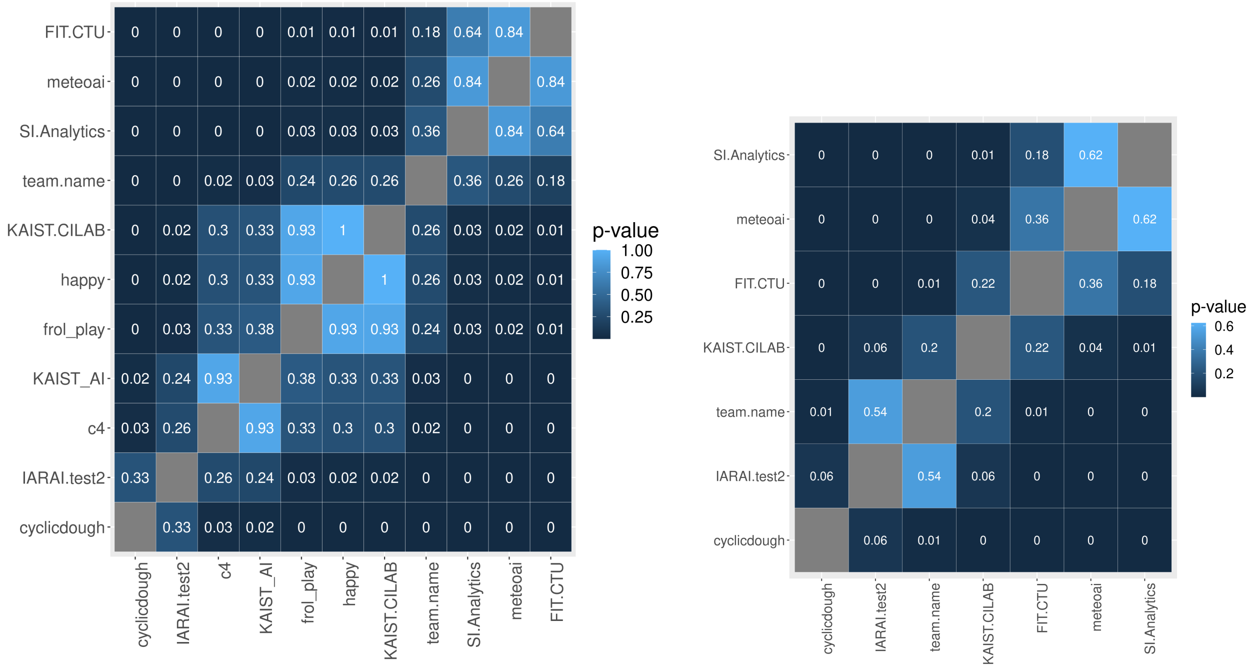
Figure A-4: Location of competition regions across Europe. Core regions used at Stage 1 and 2 (denoted by b), Extended Core used at Stage 2 only (denoted by r) in blue, and Transfer learning regions in red.

especially in southern countries. From this analysis, a number of target patches for the core and the transfer learning challenge are selected, for which data is provided to participants. Location of the selected regions is given in Fig. A-4.

A.4. Metric Considerations

Before calculating the final metric, from each region we removed pixels out of OPERA coverage or with missing data, coded as –9,999,000 and –8,888,000 values, respectively. The first case happens when a region of interest encompasses a sea area that is beyond the coverage of a ground radar. The second one is related to errors that occur during the collection of OPERA radiance data. To prevent the misidentification of cluttering echoes and artefacts, and to account for the satellite's limitations in detecting precipitation, a threshold of 0.2 mm/h was introduced in the second stage of the competition. As a result, our dataset became even more imbalanced, leading to a decreased occurrence of rain events and making the prediction task more challenging. This difficulty is reflected in lower evaluation metric values on the leaderboard.

A.5. Leaderboard Significance and Awards



(a) Core Leaderboard

(b) Transfer-Learning Leaderboard

Figure A-5: All-*vs*-all rank comparisons. We show estimates of the False Discovery Rate of pairwise *post-hoc* tests for significant differences in ranking across all tested regions and years.

All leaderboard rankings were highly significant overall (Friedman test). Pairwise *post-hoc* mean-rank statistics based on Tukey's honestly significant differences (Demšar, 2006) validated the *ex-aequo* ranking of two teams at rank 3 in the core leaderboard. Notably, the top ranked teams performed significantly better than the others, justifying the prizes and recognition awarded to the winners. For the transfer learning leaderboard the top ranked team did significantly better than teams at rank 4 or lower. The Scientific Committee awared the special transfer learning award to the top ranked team based on its leading performance and deeper coverage of the transfer learning aspect in their conference research paper. The pairwise similarity structures for both leaderboards are shown in Fig. A-5.

References (Appendix)

A. Karagiannidis, A. Bloutsos, P. Maheras, and Ch. Sachsamanoglou. Some statistical characteristics of precipitation in Europe. Theor Appl Climatol, 91:193–204, 2008. doi: 10.1007/s00704-007-0303-7.

S. Ravuri, K. Lenc, M. Willson, and et al. Skilful precipitation nowcasting using deep generative models of radar. Nature, (597):672–677, 2021. doi: 10.1038/s41586-021-03854-z.

World Meteorological Organization. Guide to instruments and methods of observation. Technical Report WMO No. 8, 2018.