ClimateBench-M: A Multi-Modal Climate Data Benchmark with a Simple Generative Method

Dongqi Fu
dongqif2@illinois.edu
University of Illinois Urbana-Champaign
IL, USA

Yada Zhu
yzhu@ibm.us.com
IBM Research
NY, USA

Zhining Liu
liu326@illinois.edu
University of Illinois Urbana-Champaign
IL, USA

Lecheng Zheng
lecheng4@illinois.edu
University of Illinois Urbana-Champaign
IL, USA

Xiao Lin
xiaol13@illinois.edu
University of Illinois Urbana-Champaign
IL, USA

Zihao Li
zihaoli5@illinois.edu
University of Illinois Urbana-Champaign
IL, USA

Liri Fang
lirif2@illinois.edu
University of Illinois Urbana-Champaign
IL, USA

Katherine Tieu
kt42@illinois.edu
University of Illinois Urbana-Champaign
IL, USA

Onkar Bhardwaj
onkarbhardwaj@ibm.us.com
IBM Research
MA, USA

Kommy Weldemariam
kommy@ibm.us.com
IBM Research
NY, USA

Hanghang Tong
htong@illinois.edu
University of Illinois Urbana-Champaign
IL, USA

Hendrik Hamann
hendrikh@ibm.us.com
IBM Research
NY, USA

Jingrui He
jingrui@illinois.edu
University of Illinois Urbana-Champaign
IL, USA

Abstract

Climate science studies the structure and dynamics of Earth's climate system and seeks to understand how climate changes over time, where the data is usually stored in the format of time series, recording the climate features, geolocation, time attributes, etc. Recently, much research attention has been paid to the climate benchmarks. In addition to the most common task of weather forecasting, several pioneering benchmark works are proposed for extending the modality, such as domain-specific applications like tropical cyclone intensity prediction and flash flood damage estimation, or climate statement and confidence level in the format of natural language. To further motivate the artificial intelligence development for climate science, in this paper, we first contribute a multi-modal climate benchmark, i.e., **ClimateBench-M**, which aligns (1) the time series climate data from ERA5, (2) extreme weather events data from NOAA, and (3) satellite image data from NASA HLS based on a unified spatial-temporal granularity. Second, under each data modality, we also propose a simple but strong generative method that could produce competitive performance in weather forecasting, thunderstorm alerts, and crop segmentation tasks in the proposed ClimateBench-M. The data and code of ClimateBench-M are publicly available at https://github.com/iDEA-iSAIL-Lab-UIUC/ClimateBench-M.

CCS Concepts

• **Applied computing** → **Earth and atmospheric sciences**; • **Computing methodologies** → **Spatial and physical reasoning**.

Keywords

Extreme Weather Forecasting, Geo-Image Segmentation

ACM Reference Format:

Dongqi Fu, Yada Zhu, Zhining Liu, Lecheng Zheng, Xiao Lin, Zihao Li, Liri Fang, Katherine Tieu, Onkar Bhardwaj, Kommy Weldemariam, Hanghang Tong, Hendrik Hamann, and Jingrui He. 2025. ClimateBench-M: A Multi-Modal Climate Data Benchmark with a Simple Generative Method. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3746252.3761647

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761647

1 Introduction

Climate science investigates the structure and dynamics of earth's climate system and seeks to understand how global, regional, and local climates are maintained as well as the processes by which they change over time,[1] In general, climate data is usually represented by a time series numerical format that covers climate features (e.g., temperature, wind, and atmospheric water content), geolocation information (e.g., longitude, latitude, and geocode), and time (e.g., hours and days). Recently, to develop artificial intelligence techniques for climate science, many interesting climate benchmarks have been proposed. For example, *WeatherBench* [14] provides a common data set and evaluation metrics to enable direct comparison between different data-driven approaches to medium-range weather forecasting (3-5 days lead time). In addition to the weather forecasting climate benchmarks, some task-specific and domain-specific benchmarks are proposed. For example, the authors in [13] present a large-scale climate dataset called *ExtremeWeather*, which is designed to encourage machine learning research in the detection, localization, and understanding of extreme weather events, to further address the problem that the existing labeled data for climate patterns like hurricanes, extra-tropical cyclones, and weather fronts can be incomplete.

Those aforementioned benchmarks pave the way for developing possible artificial intelligence techniques for climate science from one single aspect. Then, a natural question arises: **can we provide a comprehensive climate benchmark that has multiple data modalities for chasing the artificial intelligence [1]** for climate applications? To speed up the AI development for climate science, in this paper, we first propose a multi-modal climate benchmark named ClimateBench-M, which aligns the ERA5 [7][2] time series data for weather forecasting, NOAA [3] extreme weather events records for extreme weather alerts, and HLS [8] [4] satellite image data for the crop segmentation, based on a unified spatial-temporal granularity. Moreover, we also propose a simple generative model, called SGM, for each task in the proposed ClimateBench-M. SGM is based on the encoder-decoder framework, and the choices of encoders and decoders vary for different tasks. Overall, in each task of ClimateBench-M, SGM produces a competitive performance with different baseline methods.

2 ClimateBench-M

Datasets. ClimateBench-M benchmark aligns three datasets from different modalities based on the spatial and temporal granularity. The raw data originates from public datasets *ERA5* [7][5], *NOAA* [6] and *NASA HLS* [8] [7].

- ERA5 provides hourly estimates for a large number of atmospheric, ocean-wave and land-surface quantities. The data is available from 1940 onwards.
- NOAA is National Oceanic and Atmospheric Administration that has the National Centers for Environmental Information (NCEI), which center published the Storm Events Database, currently recording the data from January 1950 to February 2024, as entered by NOAA's National Weather Service (NWS).
- The NASA HLS (Harmonized Landsat and Sentinel-2) v2.0 dataset integrates high-resolution, multi-spectral satellite images from Landsat and Sentinel-2 missions, spanning from 2013 to present.

Task 1: Weather Forecasting. We denote the weather time series data stored in $\mathcal{X} \in \mathbb{R}^{N \times D \times T}$. Note that a slice of $\mathcal{X}$, i.e., $\mathcal{X}(i, :, :) \in \mathbb{R}^{D \times T}$, $i \in \{1, \ldots, N\}$, is typically denoted as the common multivariate time series data [18, 20]. For example, in each element $\mathcal{X}(i, d, t)$ of the nationwide weather data $\mathcal{X}$, $i \in \{1 \ldots, N\}$ can be the number of spatial locations (e.g., counties), $d \in \{1 \ldots, D\}$ can be the dimension of weather features (e.g., temperature and humidity), and $t \in \{1 \ldots, T\}$ can be the timestamp (e.g., hour). Throughout the paper, we use the calligraphic letter to denote a 3D tensor (e.g., $\mathcal{X}$) and the bold capital letter to denote a 2D matrix (e.g., $\boldsymbol{X}$). Given the time series data $\mathcal{X} \in \mathbb{R}^{N \times D \times T}$, we aim to forecast the future data $\mathcal{X}' \in \mathbb{R}^{N \times D \times \tau}$, where τ is a forecasting window.

Task 2: Thunderstorm Alerts. Recall that, in the forecasting task, we aim to forecast the future data $\mathcal{X}' \in \mathbb{R}^{N \times D \times \tau}$ from the history data $\mathcal{X} \in \mathbb{R}^{N \times D \times T}$. For achieving the thunderstorm alert task, we also aim to find the anomaly in the forecast, i.e., with the forecast $\mathcal{X}'$, we aim to detect if $\mathcal{X}'$ contains abnormal values, i.e., whether thunderstorms happens in a certain location on a certain future hour based on the forecasting window.

Task 3: Crop Segmentation. For the crop segmentation task, we collect a series of satellite images at different times but at the same place, aiming to distinguish the crop types in various regions within those images. Specifically, we denote the satellite images as $\mathcal{X} \in \mathbb{R}^{N \times D \times T}$, where N represents the number of pixels within the images, D represents the number of channels (e.g., RGB brand, near-infrared, and shortwave infrared), and T represents the number of images at the same place. We also denote the crop types as $\mathbf{y} \in \mathbb{R}^{N}$, and $\mathbf{y}(i), i \in \{1, \ldots, N\}$ represents the type of crop grown in the area corresponding to the i-th pixel. Given the image data $\mathcal{X} \in \mathbb{R}^{N \times D \times T}$, we aim to predict the crop type of each pixel $\mathbf{y} \in \mathbb{R}^{N}$, as shown in Figure 2.

3 Simple Generative Model (SGM)

We first give an overview of SGM and then induce the details of applying it to different tasks of ClimateBench-M benchmark.

Overview of ClimateBench-M. As shown in Figure 3, SGM is based on an encoder-decoder framework and has two pipelines. The upper pipeline is for time series forecasting (targeting the weather forecasting task) and anomaly detection (targeting the thunderstorm alerts). The lower pipeline is for image segmentation (targeting the temporal crop segmentation).

Deployment of SGM for Time Series Forecasting and Anomaly Detection. Here, we briefly introduce how the upper pipeline of SGM achieves time series forecasting and anomaly detection. The detailed information can also be found in our previous paper [5]. We design a simple but effective module in SGM to achieve anomaly detection along with the forecasting, i.e., an encoder-decoder model that tries to explore the distribution of normal features in $\mathcal{X}$

[1] https://plato.stanford.edu/entries/climate-science/
[2] https://cds.climate.copernicus.eu/cdsapp#!/home
[3] https://www.ncdc.noaa.gov/stormevents/ftp.jsp
[4] https://huggingface.co/datasets/ibm-nasa-geospatial/multi-temporal-crop-classification
[5] https://cds.climate.copernicus.eu/cdsapp#!/home
[6] https://www.ncdc.noaa.gov/stormevents/ftp.jsp
[7] https://huggingface.co/datasets/ibm-nasa-geospatial/multi-temporal-crop-classification

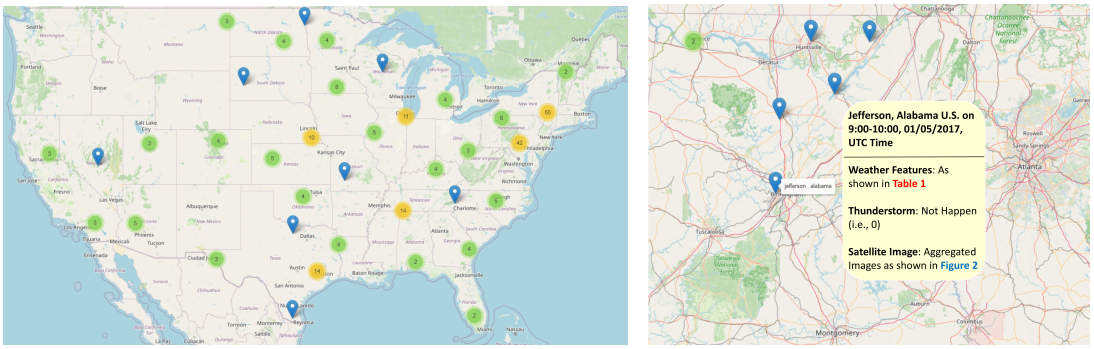

Figure 1: Left: Geographic Distribution of Covered Counties in ClimateBench-M (The number in the circle stands for the aggregation of nearby counties) Right: A Specific Example of Jefferson, Alabama U.S. on 9:00-10:00, 01/05/2017, UTC Time

Table 1: (Part of) Features with Instance Values Sampled from Jefferson, Alabama U.S. on 9:00-10:00, 01/05/2017, UTC.

Feature	Unit	Description	Value
10-meter wind gust (maximum)	m s^{-1}	Maximum 3-second wind at 10 m height as defined by WMO. Parametrization represents turbulence only before 01102008; thereafter effects of convection are included. The 3 s gust is computed every time step, and the maximum is kept since the last postprocessing.	3.620435
Atmospheric water content	kg m^{-2}	This parameter is the sum of water vapor, liquid water, cloud ice, rain, and snow in a column extending from the surface of the Earth to the top of the atmosphere. In old versions of the ECMWF model (IFS), rain and snow were not accounted for.	9.287734

Table 2: Statistics of Thunderstorm Records in ClimateBench-M over 238 Selected Counties in the US from 2017 to 2021

Year	2017	2018	2019	2020	2021
Jan	26	3	2	41	7
Feb	53	6	9	50	8
Mar	85	16	26	63	62
Apr	93	44	140	170	60
May	245	207	263	175	218
Jun	770	302	348	331	452
Jul	306	291	457	453	701
Aug	294	269	415	354	435
Sep	61	80	122	29	123
Oct	32	32	82	60	55
Nov	20	22	9	114	11
Dec	5	15	11	8	58

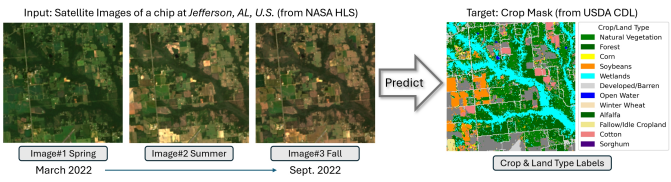

Figure 2: Example of the crop type segmentation task based on NASA HLS and USDA CDL.

as shown in Figure 3. As long as this encoder-decoder model can

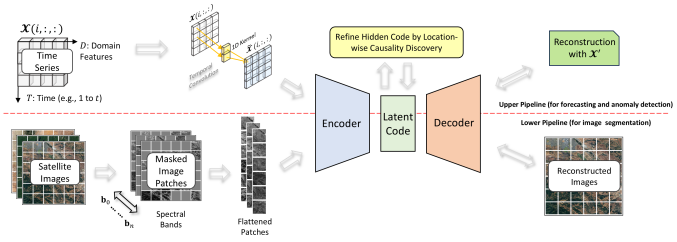

Figure 3: Simple Generative Model (SGM). Upper level of the figure is the time series forecasting pipeline, and the lower level of the figure is the image segmentation pipeline. Two pipelines have different choices of encoders and decoders.

capture the latent distribution for normal events, then the generation probability of a piece of time series data can be utilized as the condition for detecting anomaly patterns. This is because the extreme values are identified with a remarkably low generation probability. To be specific, after the forecast $\boldsymbol{H}^{(t)}$ is output, the generation probability of $\boldsymbol{H}^{(t)}$ into $\boldsymbol{X}^{(t)}$ can be used as the evidence to detect the anomalies at t. The transformation from $\boldsymbol{X}^{(t)}$ to $\boldsymbol{H}^{(t)}$ can be realized by a model-agnostic pre-trained autoencoder. Moreover, we use the mean absolute error (MAE) loss on the prediction and the ground truth, which is effective and widely applied to time-series forecasting tasks [11, 17].

$$\min_{\Theta_i, \boldsymbol{A}^{(t-1)}, \ldots, \boldsymbol{A}^{(t-l)}} \mathcal{L}_{pred} = \sum_i \sum_t |H(i,:)^{(t)} - \hat{H}(i,:)^{(t)}| \quad (1)$$

where $\Theta_i, \boldsymbol{A}^{(t-1)}, \ldots, \boldsymbol{A}^{(t-l)}$ are all learnable parameters for the prediction $\hat{H}(i,:)^{(t)}$ of variable i at time t. Note that $\boldsymbol{A}^{(t-1)}$ is a learnable parameter denoting the causal effects among all locations at time t for better forecasting performance, and the learning simply relies on the Structural Equation Model (SEM) [21].

Deployment of SGM for Image Segmentation. In the task of crop classification, we use mmsegmentation [19], an OpenMMLab Semantic Segmentation Toolbox, to segment the satellite images, following [9]. To handle the crop satellite image, we choose vision transformer [4] as the backbone of the encoder-decoder pairs for our proposed SGM. We use random crop and random flip to augment the training data.

Table 3: Forecasting Error (MAE, 10^{-2})

	ERA5-2017 (↓)	ERA5-2018 (↓)	ERA5-2019 (↓)	ERA5-2020 (↓)
GRU	1.8834 ± 0.0126	1.9764 ± 0.1466	1.6194 ± 0.2645	1.7859 ± 0.2324
DCRNN	0.0819 ± 0.0025	0.0797 ± 0.0049	0.0799 ± 0.0035	0.0826 ± 0.0033
GTS	0.0777 ± 0.0054	0.0766 ± 0.0029	0.0760 ± 0.0031	0.0742 ± 0.0021
SGM	0.0496 ± 0.0017	0.0499 ± 0.0017	0.0502 ± 0.0016	0.0488 ± 0.0019
ST-SSL	0.0345 ± 0.0051	0.0330 ± 0.0018	0.0361 ± 0.0021	0.0348 ± 0.0020
SGM++	**0.0271 ± 0.0004**	**0.0276 ± 0.0004**	**0.0282 ± 0.0003**	**0.0265 ± 0.0004**

4 Experiments

Evaluation Metrics. We measure the performance of the baselines on the ClimateBench-M with respect to the following metrics: Accuracy (Acc), Mean Absolute Error (MAE), Intersection of Union (IoU) [8], and Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

Baselines. The first category is for tensor time series forecasting: (1) GRU [3], (2) DCRNN [11], (3) GTS [17], and (4) ST-SSL [10]. The second category is for anomaly detection on tensor time series: (1) DeepSAD [16], (2) DeepSVDD [15], and (3) DROCC [6]. Since these three methods have no forecast abilities, we let them use the ground-truth observations, and our SGM utilizes the forecast features during anomaly detection experiments. The third category is for image segmentation: (1) DeepLabV3 [2] and (2) Swin [12]. Since the aforementioned baselines do not inherently incorporate temporal dependencies, we concatenate all images at the same location along the channel dimension and utilize the combined image for segmentation.

Weather Forecasting. In Table 3, we can observe a general pattern that our SGM outperforms the baselines with GTS performing better than DCGNN. An explanation is that the temporally fine-grained causal relationships can contribute more to the forecasting accuracy than non-causal directed graphs. Moreover, ST-SSL achieves competitive forecasting performance via contrastive learning on time series. Motivated by a contrastive manner, SGM++ is proposed by persistence forecast constraints. That is, the current forecast of SGM is further calibrated by its nearest past time window (i.e., the last 24 hours in our setting).

Anomaly Detection. After forecasting, we can have the hourly forecast of weather features at certain locations, denoted as $\mathcal{X}'$. Then, we use the encoder-decoder model in Figure 3 to calculate the feature-wise generation probability using mean squared error (MSE) between $\mathcal{X}'$ and its generation $\tilde{\mathcal{X}}'$. Thus, we can calculate the average of feature-wise generation probability as the condition of anomalies to identify if an anomaly weather pattern (e.g., a thunderstorm) happens in an hour in a particular location. In Table 4, we use the Area Under the ROC Curve (i.e., AUC-ROC) as the metric, repeat the experiments four times, and report the performance of ClimateBench-M with baselines.

[8]It measures the ratio of the intersection of two sets over the union of two sets

Table 4: Anomaly Detection Performance (AUC-ROC)

	NOAA-2017 (↑)	NOAA-2018 (↑)	NOAA-2019 (↑)	NOAA-2020 (↑)
DeepSAD	0.5305 ± 0.0481	0.5267 ± 0.0406	0.5563 ± 0.0460	0.6420 ± 0.0054
DeepSVDD	0.5201 ± 0.0045	0.5603 ± 0.0111	**0.6784 ± 0.0112**	0.5820 ± 0.0205
DROCC	0.5319 ± 0.0661	0.5103 ± 0.0147	0.6236 ± 0.0992	0.5630 ± 0.1082
SGM	**0.5556 ± 0.0010**	**0.5685 ± 0.0011**	0.6298 ± 0.0184	**0.6745 ± 0.0185**

Crop Classification. In addition to the first two tasks, we also assess the quality of ClimateBench-M in the crop classification task. Table 5 presents the results of baseline methods. We have the following observations: (1) All methods achieve good performance on some class, such as Open Water, Soybeans, Corn, Forest, etc, indicating the high quality of our benchmark. (2) These methods tend to perform worse in other classes, such as Sorghum, Other, Alfalfa. By investigation, we attribute this observation to the limited samples for these classes, comparing with the rich samples for the classes with good performance. (3) Our proposed method SGM outperforms baseline methods, demonstrating the effectiveness.

Table 5: Crop Classification

Baselines	SGM		Swin		DeepLabV3	
Classes	IoU (↑)	Acc (↑)	IoU (↑)	Acc (↑)	IoU (↑)	Acc (↑)
Natural Vegetation	39.23	46.86	45.66	71.80	47.31	64.28
Forest	42.44	61.07	34.47	41.63	46.50	77.10
Corn	53.30	63.56	52.00	62.53	52.30	72.81
Soybeans	54.35	69.76	56.53	72.78	47.96	72.54
Wetlands	40.17	59.55	42.15	69.57	35.42	43.62
Developed/Barren	34.88	52.25	40.19	56.08	44.04	58.88
Open Water	69.49	91.89	76.09	57.81	76.39	88.85
Winter Wheat	55.54	75.96	48.21	86.41	47.75	54.32
Alfalfa	24.78	55.51	20.99	54.64	29.39	34.84
Fallow/ Idle Cropland	38.32	61.75	37.14	23.23	17.55	19.45
Cotton	33.53	66.66	24.38	65.86	35.80	66.38
Sorghum	33.48	68.93	33.95	28.85	23.40	24.85
Other	28.27	42.81	28.72	45.56	27.14	41.58
Average	**42.14**	**62.81**	41.57	55.67	40.84	55.34

5 Conclusion

In conclusion, we provide a multi-modal climate benchmark named ClimateBench-M, integrating diverse datasets and assessing the quality of this benchmark by conducting experiments with various tasks. Our experimental results demonstrate the high quality of ClimateBench-M. Additionally, we propose SGM, a simple encoder-decoder-based generative model, which demonstrates competitive performance across various tasks.

Acknowledgement

This work is supported by National Science Foundation under Award No. IIS-2416070, and IBM-Illinois Discovery Accelerator Institute - a new model of an academic-industry partnership designed to increase access to technology education and skill development to spur breakthroughs in emerging areas of technology. The content of the information in this document does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

GenAI Usage Disclosure

In this paper, authors do not use generative AI software tools to create the content.

References

[1] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *CoRR* abs/2303.12712 (2023). doi:10.48550/ARXIV.2303.12712 arXiv:2303.12712

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017).

[3] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *CoRR* abs/1412.3555 (2014). arXiv:1412.3555 http://arxiv.org/abs/1412.3555

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

[5] Dongqi Fu, Yada Zhu, Hanghang Tong, Kommy Weldemariam, Onkar Bhardwaj, and Jingrui He. 2024. Generating Fine-Grained Causality in Climate Time Series Data for Forecasting and Anomaly Detection. *CoRR* abs/2408.04254 (2024). doi:10.48550/ARXIV.2408.04254 arXiv:2408.04254

[6] Sachin Goyal, Aditi Raghunathan, Moksh Jain, Harsha Vardhan Simhadri, and Prateek Jain. 2020. DROCC: Deep Robust One-Class Classification. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 3711–3721. http://proceedings.mlr.press/v119/goyal20c.html

[7] Hans Hersbach, Bill Bell, Paul Berrisford, Gionata Biavati, András Horányi, Joaquín Muñoz Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Iryna Rozum, et al. 2018. ERA5 hourly data on single levels from 1979 to present. *Copernicus climate change service (c3s) climate data store (cds)* 10, 10.24381 (2018).

[8] Johannes Jakubik, Linsong Chu, Paolo Fraccaro, Ranjini Bangalore, Devyani Lambhate, Kamal Das, Dario Oliveira Borges, Daiki Kimura, Naomi Simumba, Daniela Szwarcman, Michal Muszynski, Kommy Weldemariam, Bianca Zadrozny, Raghu Ganti, Carlos Costa, Campbell Watson, Karthik Mukkavilli, Sujit Roy, Christopher Phillips, Kumar Ankur, Muthukumaran Ramasubramanian, Iksha Gurung, Wei Ji Leong, Ryan Avery, Rahul Ramachandran, Manil Maskey, Pontus Olofossen, Elizabeth Fancher, Tsengdar Lee, Kevin Murphy, Dan Duffy, Mike Little, Hamed Alemohammad, Michael Cecil, Steve Li, Sam Khallaghi, Denys Godwin, Maryam Ahmadi, Fatemeh Kordi, Bertrand Saux, Neal Pastick, Peter Doucette, Rylie Fleckenstein, Dalton Luanga, Alex Corvin, and Erwan Granger. 2023. *HLS Foundation*. doi:10.57967/hf/0952

[9] Johannes Jakubik, Sujit Roy, CE Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, et al. 2023. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660* (2023).

[10] Jiahao Ji, Jingyuan Wang, Chao Huang, Junjie Wu, Boren Xu, Zhenhe Wu, Junbo Zhang, and Yu Zheng. 2023. Spatio-Temporal Self-Supervised Learning for Traffic Flow Prediction. In *AAAI 2023*.

[11] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2018. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=SJiHXGWAZ

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.

[13] Evan Racah, Christopher Beckham, Tegan Maharaj, Samira Ebrahimi Kahou, Prabhat, and Chris Pal. 2017. ExtremeWeather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 3402–3413. https://proceedings.neurips.cc/paper/2017/hash/519c84155964659375821f7ca576f095-Abstract.html

[14] Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey. 2020. WeatherBench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems* 12, 11 (2020), e2020MS002203.

[15] Lukas Ruff, Nico Görnitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Robert A. Vandermeulen, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep One-Class Classification. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 4390–4399. http://proceedings.mlr.press/v80/ruff18a.html

[16] Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. 2020. Deep Semi-Supervised Anomaly Detection. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=HkgH0TEYwH

[17] Chao Shang, Jie Chen, and Jinbo Bi. 2021. Discrete Graph Structure Learning for Forecasting Multiple Time Series. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=WEHSlH5mOk

[18] Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. 2019. Robust Anomaly Detection for Multivariate Time Series through Stochastic Recurrent Neural Network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis (Eds.). ACM, 2828–2837. doi:10.1145/3292500.3330672

[19] Jiarui Xu, Kai Chen, and Dahua Lin. 2020. MMSegmenation. https://github.com/open-mmlab/mmsegmentation.

[20] Hang Zhao, Yujing Wang, Juanyong Duan, Congrui Huang, Defu Cao, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, and Qi Zhang. 2020. Multivariate Time-series Anomaly Detection via Graph Attention Network. In *20th IEEE International Conference on Data Mining, ICDM 2020, Sorrento, Italy, November 17-20, 2020*, Claudia Plant, Haixun Wang, Alfredo Cuzzocrea, Carlo Zaniolo, and Xindong Wu (Eds.). IEEE, 841–850. doi:10.1109/ICDM50108.2020.00093

[21] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. 2018. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). 9492–9503. https://proceedings.neurips.cc/paper/2018/hash/e347c51419ffb23ca3fd5050202f9c3d-Abstract.html