Hybrid Deep Learning Framework for Automated Weather Phenomena Recognition in Global Climate Reanalysis Data

***Mushtaq Talib Ajjah, **Mahdi Mazinani**

*Department of Computer Science - College of Education - Mustansiriya University - Baghdad - Iraq

**Department of Electrical and Electronic Engineering, ShQ.C., Islamic Azad University, Shahr-e Qods, Iran

[1]*Received: 11/06/2026; Accepted: 01/09/2026; Published: 08/09/2026*

Abstract

Climate reanalysis data describe how large-scale climate features evolve in time and space. They can be used for training machine learning models to automatically identify specific weather features. Such models have to handle complex spatial structures, and temporal variability, as well as class imbalance and small numbers of representative data. This work proposes to use ViT in combination with TCN to recognize different weather phenomena in reanalysis data. The structures of the data used in the work are analyzed, and climatological features are removed. Finally, long range temporal sequences of the data are constructed, and the models are trained to impute missing observations. ViT is used to model spatial structures in the data, and TCN to model temporal structures.

Dependence on labeled data significantly limits viability of vision transformer (ViT) architectures for climate time-series forecasting. We incorporate a SimCLR-based self-supervised contrastive technique to pre-train the ViT encoder. During fine-tuning, we perform class-weighted cross-entropy training to mitigate class-imbalance. We compare our approach against baselines that include convolutional neural networks (CNNs) with/without LSTMs and Transformer-basedConvNet (TCN) architectures. Our approach achieves a provisional classification accuracy of 96.2%, a macro-F1 score of 0.962, and a macro-ROC-AUC of 0.989, on a temporally independent test case. Temporal models, and self-supervised and meteorologically focused architecture embeddings further bolster the framework. The focus of the attribution function of the framework is on expected locations of weather phenomena based on prior knowledge, supporting the framework's focus. The framework demonstrates the capability of vision transformers to model weather in climate data, and provides a foundation for time-series forecasting of weather events.

Keywords : *Climate Reanalysis; Weather Phenomena Recognition; Vision Transformer; Temporal Convolutional Network; Self-Supervised Learning; Explainable Artificial Intelligence*.

I. INTRODUCTION

Identifying climate signals in large datasets requires the ability to extract relevant signals from large datasets. Some extreme climate and weather events are caused by the coalescence of different types of atmospheric structures. Such structures result from different physical and mechanical processes and can occur across varied spatial and temporal dimensions. For the purpose of improving weather and climate sciences as well as modeling, it is important to understand the structures, persistence and physics of coupled climate and weather extremes. Advances in machine learning allow researchers to extract complex structures from large and high-dimensional climate and earth system

[1] *How to cite the article:* Ajjah M.T., Mazinani M.; September 2026; Hybrid Deep Learning Framework for Automated Weather Phenomena Recognition in Global Climate Reanalysis Data; *International Journal of Innovations in Scientific Engineering*, Jul-Dec 2026, Vol 24, 10-40

data. Some of the remaining challenges in this area of research are inter alia the explanation and/or representation of physical processes in machine learning systems as well as incorporation of extreme climate events [1], [2].

Global reanalysis data provide an estimate of global atmospheric conditions over areas and times where observations are sparse or unavailable. Reanalysis data integrate observations with numerical data and models. ERA5 provides European Center data on global conditions from 1940, and is unique in providing data on shorter time intervals (e.g., hourly data) [3]. MERRA-2 is a more recent global reanalysis data set provided by NASA [4]. Similar to ERA5 and other reanalysis data, MERRA-2 and other data sets cover a range of environmental conditions including pressure, humidity, and temperature, and provide data on a range of vertical levels including the condensation and precipitation processes. Other data sets of this nature provide environmental conditions that are integral to the study of the processes and dynamics of weather systems. The large volume of data makes it difficult to analyze completely and over long time periods.

Evaluating extreme weather events often requires subjective determination based on various kinds of human-engineered constraints including indices, rules, and expert judgment. While definitions based on rules and expert judgment contain a lot of physical and tension meteorology, these definitions suffer from subjective analysis and depend on thresholds, the geographical region, the level of detail of the data, and rules and methods used to identify and define the event [5]. Topological data analysis, for example, can be used to detect atmospheric rivers in a climate model and other data sources, and can largely be threshold- and expert judgment-free. Furthermore, Muszynski and others showed that once automated, analysis of climatological data using deep learning can aid in detection of extreme weather events [6]. Other studies largely in the field of climate science have adopted similar methods, and some have applied deep learning in the domain of extreme event detection [7]. ClimateNet, for example, created datasets, labels, and a deep learning framework to support classification of atmospheric Rivers and Tropical Cyclones. In a separate study, deep learning methods aided in detection of atmospheric Rivers in large climate model ensembles [8].

Although we have made progress in this area, many of the methodological issues remain. Much of the literature that has been published about event recognition address a single phenomenon, a single region, and spatial detection and segmentation. While these particular formulations are essential for its positioning, since most atmospheric processes can't be identified with a single snapshot in space since each formation, strengthening, sustained and decay are heavily defined temporally. Additionally, an approach that is based on a particular number of phenomena or a single source of observations would potentially fail when the characteristics change, with change in the regimes, location, spatial scale or reanalysis datasets.

The recent assessment of the existing machine learning (ML) techniques in the research area of the weather and climate also emphasizes the issues of the transferability, the limited availability of training data in the extreme events, interpretability and trustworthy models with a reasonable behavior [1], [2], as the most pressing challenges ahead. For this reason, a robust approach to the recognition of the weather phenomena must adequately account for the spatial and temporal domain of the data, while avoiding the reliance on large human-annotated datasets.

Spatially Distributing Atmospheric Structure - ViTs are able to encode the distribution of spatial structures in the atmosphere. Unlike local convolutional neural networks (CNNs), which work on local neighborhoods, the image field is represented by a sequence of spatial patches, and the ViT can be used to encode the relationships of distant regions with self-attention [9]. The large ambient atmospheric structure and large- and high signature meteorological structures may extend across a large region. Unlike modeling based on a single spatial dimension or time dimension, spatially modeling structure across a large region fails to describe the evolution of the event in time.

Temporal convolutional neural networks (TCNs) provide a means of modeling such sequences of information. Based on causal (time ordered) dilated and residual convolution, TCNs can get a long effective receptive field without the long recurrent calculations of ordinary recurrent neural networks. To connect the spatial encoder of ViT with the temporal encoder of TCN would provide an idea of learning the spatial manifestation of the atmospheric flow fields and the temporal persistence of the sequence of the atmospheric flow fields that are multivariate.

A second challenge is the lack of and imbalance of reliable event labels versus the enormous number of reanalysis data. Extreme and high impact events occur by nature less frequently than the background state of the atmosphere. Expert generated labels are also expensive and can be ambiguous at times due to the definition of the event [6], [8]. Self-supervised representation learning can be leveraged to make use of unlabeled climate data before the final supervised step of fine-tuning.

We can learn representations by making two views of an image (Augmentations) of the same sample agree and be different from other views of the same image while attending to the separation of the learned representations of other samples. We can pre-train the spatial encoder with the large amount of unlabeled atmospheric data then fine-tune the end-to-end model to recognize a weather event. In this case, the structure in geophysical fields must be used to create the augmentations rather than assume that all the image augmentations used for natural images can be used with real weather data [11].

These point out the major deficiency of related work that we seek to address in this study. Breakthroughs in image classification, spatial attention, temporal modeling, self-supervised learning, have all been achieved as independent research lines. Few works have integrated these lines into a single model that can learn long-range spatial interactions and temporal development of events and simultaneously learn semantically meaningful components of the atmosphere and meaningful representations from unlabeled global reanalysis data for multi-factor phenomenon detection. Additionally, model performance should be evaluated for a large temporal gulf from training to testing, and the semantics of the learned representations should be analyzed to determine whether the model has learned the relevant structures in the data (and thus captured meaningful atmosphere elements), rather than just overfitting to the correlations.

This work provides a solution to automate our ability to detect weather phenomena by a deep learning framework that combines a Vision Transformer (ViT) with a Temporal Convolutional Network (TCN) that is trained on meteorologically guided preprocessing and atmospheric indices and combines spatial attention and temporal sequence modeling with self-supervised learning. This work makes four important advances in automated weather phenomena classification. 1) A new architecture is suggested, where the ViT model is used to encode the long range spatial information of atmospheric variables, while the TCN model is used to encode the temporal evolution of the atmospheric information over successive observations. 2) We preserve a great amount of domain knowledge in our work, through the incorporation of the real and derived variables that are embedded into the model as meaningful meteorological and climate variables. 3) To lower our reliance on costly labeled training data, we employ the technique of self-supervised learning by contrastive pretraining on our reanalysis data. 4) Our proposed models are validated with a cross domain and separate time training, validation, and testing regimes, and are compared with several baseline models: CNNs, CNN-LSTMs, ViT only, TCN only, and ViT-TCNs. Interpretation based on attention maps helps to validate the meteorologically motivated design of the model.

This paper is organized as follows. Section II covers related studies on machine learning and deep learning for recognizing weather patterns and extreme events. Section III describes reanalysis datasets, preprocessing and labeling, the proposed ViT–TCN architecture, self-supervised learning, and evaluation. Section IV shows the results of experiments and baseline and ablation studies and discusses the interpretability of the model. The last section reviews the study finding limitations and suggests future directions.

(IJISE) 2026, Vol. No. 24, Jul-Dec e-ISSN: 2454-6402, p-ISSN: 2454-812X

II. RELATED WORK

Automated weather-phenomena analysis has transitioned from threshold- or rule-based detection and handcrafted feature extraction to deep neural networks capable of learning representations directly from the multivariate atmospheric state. More recently, attention-based architectures, large scale pretraining, and foundation models have shown promise for modeling such complex spatiotemporal dependencies in global weather and climate fields. Weather-phenomena recognition is fundamentally a different task than numerical weather prediction: it is intended to recognize, classify, or localize physically meaningful phenomena, rather than to predict the future atmospheric state. This is worth noting when considering which previous work may be relevant to automated multi-phenomenon recognition.

A. Machine Learning for Weather-Phenomenon Identification

The first data-driven approaches attempted to minimizemeta-data dependence of events identification to manually set thresholds. Muszynski et al. [12] presented an approach that isagnostic to the atmospheric-river detection threshold and aims at identifying the characteristic morphology of ARs. Topological data analysis (TDA) and support vector machine (SVM) classifiers were used for the recognition of ARs. Structural features are extracted from integrated water-vapor fields and classified in climate simulations CAM5.1 and reanalysis MERRA-2 at various resolutions, reporting classifers with 77% to 91% accuracy and showing that machine-learning techniques can extract relevant signatures for ARs without imposition of a physically motivated threshold. The recognition task, however, remained limited to binary recognition, used labels from an existing AR detection heuristic, and did not model event evolution through time.

Another notable step toward supervised deep-learning-based weather-event identification was carried out by Climate Net, proposed by Prabhat et al. [13]. They trained their system with an expert-labeled climate dataset and performed DeepLabv3+ semantic segmentation that identified pixels belonging to TCs, atmospheric rivers, and background fields. Rather than relying on learned threshold values, the network learned the morphology of these features from a set of time-dependent multivariate atmospheric variables. This was more of a proof-of-concept than a fully operational system; the paper showed that deep semantic segmentation could perform high-resolution, computationally-efficient analysis of large climate-model data fields. Transfer to observational and reanalysis data was left for a future study.

Higgins et al. [14] then took the next step and explored how flexible deep learning methods are for atmospheric-river detection across datasets and resolutions. Using the CG-Climate framework, they trained a portable shallow convolutional segmentation architecture on model levels of zonal and meridional winds at 850 hPa, integrated water vapor, and surface pressure to detect ARs without including integrated vapor transport as an input. Using Climate Net labels for training, they validated against multiple ARTMIP detection algorithms in MERRA-2, and applied to thousands of simulated winters from Weather@Home. Higgins et al. [14] demonstrated that a portable, lightweight neural network architecture can significantly reduce the cost of analyzing large climate ensembles, but that this approach remains limited to one meteorological phenomenon and the differences between the training and forecast spatial domain must be accounted for.

These studies established that machine learning can replace or complement rigid heuristic criteria for selected extreme-weather patterns. They also exposed two persistent difficulties: the scarcity of consistent expert labels and the limited generalization of models trained for a particular event type, geographical domain, or spatial resolution.

B. Deep Spatiotemporal Learning in Global Weather Data

Alongside event recognition research, fast advances in data-driven global weather prediction have already proven that deep neural networks can discover remarkably sophisticated relationships between the atmosphere in the legacy of several decades reanalysis data. While forecasting is not the same as classifying weather phenomena, these developments are a strong signal for choices of architecture to learn atmospheric representations at large spatial and time scales.

Pangu-Weather: A 3D Earth-specific Transformer for High-Resolution Global Weather Forecasting By: Bi et al. [15] Pangu-Weather was a global medium-range weather forecast system using a 3D Earth-specific Transformer model. The model was trained on 39 years of ERA5 data (0.25 resolution) and predicted several atmospheric variables across pressure levels simultaneously. Using the 3D attention mechanism, the model was able to identify long-term dependencies in the forward and sideways directions, among pressure levels, and across many atmospheric variables, using hierarchical temporal aggregation to ease error accumulation in multi-time step predictions. Pangu-Weather was also able to accurately track tropical cyclones and provides strong evidence that transformer-based models are capable of sensing long-range dependencies in the atmosphere; however, Pangu-Weather is designed to forecast the future atmospheric states, not classify the presence of multiple weather phenomena.

Building on the performance of Pangu-Weather, in [16], Chen et al. Introduced a U-Transformer setup combined with a cascaded machine-learning approach. This proposed FuXi system makes predictions up to 15 days, alleviating error propagation through multiple models trained on target windows of increasing length. Learning on 0.25 validation of the ERA5 reanalysis spanning 39 years, Chen et al. Showed that explicit training on temporal dynamics benefits atmospheric ML modeling. As in Pangu-Weather, however, the system explicitly predicts fields rather than events, and thus does not directly tackle classification with class imbalance or limited class data.

The detection of these systems suggests that a good model for weather-phenomenon recognition should not consider each atmospheric map as an independent image following a statistical distribution. The progression of the formation of systems over time and the properties such as intensification, persistence, translation, and decay could hold discriminative information that is not present in a spatial snapshot alone. This urges an architecture that explicitly decouples learning of spatial representations from temporal sequences.

C. Transformer-Based and Pretrained Models for Weather and Climate

The other direction is the adoption of general-purpose Transformer model architectures trained on diverse sets of atmospheric variables. Nguyen et al. [17] have proposed ClimaX, a foundation-model architecture for weather and climate tasks. ClimaX extends the Transformer paradigm with the incorporation of variable-specific tokenization and aggregation mechanisms, enabling it to learn from heterogeneous combinations of atmospheric variables and downstream tasks. Results suggested that pretraining could produce representations that were usable by the system to multiple weather and climate tasks. While the design of ClimaX was not specifically aimed at multi-class recognition of weather phenomena, it was specified as a generalist model.

Standardized, carefully time series-based assessment has further been underscored by Rasp et al. [18] in Weather Bench 2, an open set of datasets, use-case driven metrics, and a common evaluation protocol for contemporary data-driven forecasting systems. Weather Bench 2 demonstrated how quickly Transformer, graph-based, and hybrid systems have improved, and stood for reproducibility and systematic train-test partitioning. While much of the focus in Weather Bench 2 lay on forecasting (versus recognition), the approach underscores the importance of chronological evaluation and identical experiment settings between learning systems on atmospheric data.

More recently, Bodnar et al. [20] proposed Aurora, a large-scale foundation model of the Earth-system pretrained on more than one million hours of heterogeneous geophysical data. Aurora trains a 3-dimensional Swin Transformer architecture with flexible encoders and decoders, which can then be adapted to a variety of tasks: global weather prediction, the concentration of atmospheric constituents, ocean-waves, and tropical-cyclone prediction, among others. The authors also showed that large-scale pretraining yielded improvements compared to training comparable models from scratch. This result provides further evidence for the general assumption that unlabeled or weakly labeled atmospheric data is useful as a general-purpose representation in combination with adaptation to the desired task.

However, these foundation models are computationally far more expensive than the target task-specific weather-recognition system.

D. Labeled Data, Generalization, and Explainability

One of the major limitations of the original Climate Net dataset was the challenge of generating reliable labels. Due to the complex nature of the atmospher structures involved, experts could diverge on what exactly constitutes a climate classification [13]. To enable further development of detection methods on larger datasets, Kim et al. [19] published Climate Net Large in 2025. This dataset provides 49,184 expert-guided labels of individual frames from ERA5, including segmentation masks of tropical cyclones, atmospheric blocking events, and atmospheric rivers. Each frame of each event was labeled independently by two different experts according to the expert guidelines. Climate Net Large provides a substantial increase in the amount of labeled climate events, but it also emphasizes the challenging effort required to generate expert-quality labels and further incentivizes learning methods that can harness much larger volumes of unlabeled reanalysis.

This is also closely related to the generalization of the model: a network can be quite accurate when the training and test sets are composed of samples with similar spatial distributions, events definitions, and reanalysis data, but break when the network is applied to another region, resolution, period, or reanalysis data. Published work on atmospheric-river detection has even shown sensitivity to mismatches between the training and inference domain [14]. Thus, separating the training and test data temporally is not sufficient to assess generalization, and evaluation based on data set or reanalysis type provides a more robust test of whether the model has captured meteorological structure and not dataset statistics.

Interpretability is also a desideratum for scientific use case. Camps-Valls et al. [21] summarized the advances of artificial-intelligence methods applied for extreme weather and climate analysis and enumerated major open challenges including explainability, uncertainty, limited labels, data heterogeneity, and robustness under distribution shift. In atmospheric applications, predictive accuracy does not imply physical consistency, so attribution or attention maps need to be validated against meteorological understanding, such as pressure maxima/minima, organized moisture fluxes, wind fields, or stationary temperature anomalies. This is especially relevant when using high-capacity attention models with opaque decision mechanisms.

E. Research Gap and Positioning of the Proposed Framework

What is still lacking: Despite the progress made, methodological gaps remain as shown by the related work surveyed here: 1) early ML methods, as well as several deep-learning detection schemes, have targeted individual phenomena, mostly atmospheric rivers [12], [14], or spatial semantic segmentation [13]. While Climate Net extended detection to tropical cyclones and atmospheric rivers [13], the temporal evolution of events was not explicitly modelled as a sequence-learning task. 2) Transformer-based weather models showcase the ability to learn global spatial dependencies [15], [16], but are geared towards atmospheric-state prediction rather than multi-phenomenon recognition. 3) Large-scale pretrained models point to the usefulness of learning transferables from large-scale atmospheric archives [17], [20], but task-specific methods which combine such representation learning with explicit temporal classification are still comparatively underdeveloped. 4) Label scarcity, class imbalance, cross-dataset generalization, and physical interpretability remain classic issues [19], [21].

The present work lies at the convergence of these research agendas. Rather than using atmospheric fields solely as individual images or directly optimizing for future-state prediction across entire fields, the approach recasts weather-phenomenon recognition as a spatiotemporal representation-learning task. It applies a Vision Transformer to capture long-range spatial interactions within each multivariate atmospheric snapshot, coupled with a Temporal Convolutional Network to model how the learned representations evolve over sequential measurements. We incorporates self-supervised contrastive pretraining before the supervised classification task to leverage unlabeled reanalysis data and minimize the need for infrequently available event labels. Meteorologically meaningful features and attention-based explanations are used to enhance physical interpretability. Chronological separation of train and test periods, comparison against CNN, CNN-LSTM, ViT only, and TCN-only baselines is used to delineate the influences of spatial attention, temporal modeling, and self-supervised initialization.

The key references above have been summarized in Table I. Moreover, the comparison further illustrates the evolution of our paper, from single-event pattern recognition to deep segmentation, from a Transformer-based atmospheric representations learning to large-scale pretrained models. The comparison further makes a distinction between the former forecasting-oriented Transformer systems and the general multi-phenomenon recognition goal.

TABLE I COMPARISON OF REPRESENTATIVE STUDIES RELATED TO DATA-DRIVEN WEATHER-PHENOMENA ANALYSIS

Ref.	**Study**	**Primary Task**	**Data**	**Main Method**	**Principal Contribution**	**Main Limitation Relative to This Study**
[12]	Muszynski *et al.* (2019)	Atmospheric-river recognition	CAM5.1, MERRA-2	TDA + SVM	Threshold-independent feature representation during classification; evaluated across multiple resolutions	Binary AR recognition; heuristic-derived labels; no explicit temporal sequence modeling
[13]	Prabhat *et al.* (2021), ClimateNet	TC and AR segmentation	Expert-labeled CAM5.1	DeepLabv3+	Expert-labeled climate dataset and pixel-level deep segmentation of extreme events	Primarily spatial snapshot segmentation; only two event types; reanalysis transfer not comprehensively tested
[14]	Higgins *et al.* (2023)	Atmospheric-river detection/tracking	ClimateNet, MERRA-2, Weather@Home	Lightweight CNN/CG-Climate	Efficient AR detection for very large climate ensembles; comparison with ARTMIP algorithms	Single phenomenon; sensitivity to training/inference resolution and domain
[15]	Bi *et al.* (2023), Pangu-Weather	Global weather forecasting and TC tracking	ERA5	3D Earth-specific Transformer	Demonstrated strong global spatial/vertical representation learning and efficient medium-range prediction	Forecasting rather than multi-event recognition; no event-specific classification framework
[16]	Chen *et al.* (2023), FuXi	15-day global weather forecasting	ERA5	Cascaded U-Transformer	Explicit handling of temporal forecast evolution and accumulated error	Designed for atmospheric-state prediction rather than weather-event classification
[17]	Nguyen *et al.* (2023), ClimaX	General weather/climate	Heterogeneous	Transformer foundati	Transferable representations across	General-purpose model; not specifically designed

		downstream tasks	climate datasets	on model	variables, datasets, and tasks	for multi-phenomenon temporal recognition
[18]	Rasp *et al.* (2024), WeatherBench 2	Benchmarking data-driven weather models	ERA5 and operational forecast datasets	Standardized benchmark	Reproducible evaluation framework and operationally motivated comparison metrics	Benchmark rather than an event-recognition architecture
[19]	Kim *et al.* (2025), ClimateNetLarge	Extreme-event segmentation dataset	ERA5	Expert-guided dense annotation	49,184 labeled timesteps for ARs, TCs, and blocking events	Dataset contribution rather than an explicit spatiotemporal recognition architecture
[20]	Bodnar *et al.* (2025), Aurora	Multi-domain Earth-system forecasting	Multiple geophysical datasets	3D Swin-Transformer foundation model	Large-scale pretraining and adaptation across diverse Earth-system tasks	Very high computational scale; downstream focus primarily forecasting rather than weather-event classification
[21]	Camps-Valls *et al.* (2025)	Review of AI for extreme events	Multiple Earth/climate data sources	Review of AI/DL/XAI methods	Identified limited labels, generalization, interpretability, uncertainty, and robustness as key challenges	Provides methodological synthesis rather than a specific recognition model

III. METHODOLOGY

A. Methodological Overview and Problem Formulation

This paper presents an automatic weather-phenomena recognition system as a supervised spatiotemporal classification framework based on self-supervised representation learning. The proposed framework combines a Vision Transformer (ViT) for spatial representation learning with a Temporal Convolutional Network (TCN) for learning the temporal evolution of the obtained spatial features of the atmospheric states. Multivariate atmospheric states collected from the ERA5 and MERRA-2 global reanalysis datasets are spatially (statistically) aligned, temporally synchronized, normalized and further enriched with relevant meteorological information, namely physically meaningful variables. We then extract consecutive atmospheric states as fixed-length spatiotemporal sequences and feed them into the proposed ViT-TCN model.

For a weather sample i, the input is represented as

$$\mathbf{X}_i = \{\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \dots, \mathbf{x}_{i,T}\}, \qquad \mathbf{x}_{i,t} \in \mathbb{R}^{C\times H\times W} \qquad (1)$$

Where T is the temporal length, C is the number of meteorological channels, and H and W denote the spatial dimensions of the analysis window. In the current implementation, T=24 observations at 3-h intervals, thus we consider a 72-h sequence of the atmosphere.

The objective is to estimate the posterior probability of the weather-phenomenon class

$$\widehat{\mathbf{p}}_i = f_\theta(\mathbf{X}_i) \qquad (2)$$

where f_θ denotes the complete ViT-TCN network parameterized by θ, and

$$\widehat{\mathbf{p}}_i = [p_{i,1}, p_{i,2}, \dots, p_{i,K}], \qquad \sum_{k=1}^{K} p_{i,k} = 1 \qquad (3)$$

For the four-class formulation considered in this study, $K = 4$, representing tropical cyclone, atmospheric river, heatwave, and no-event/background conditions. The predicted class is

$$\hat{y}_i = \operatorname*{argmax}_k p_{i,k} \qquad (4)$$

The training sample data: as the classifier trains via mutually exclusive softmaxes, we can't treat samples with multiple phenomena at once as single-label training samples. Multi-phenomena overlap is detected when creating the labels for the spectrum and is eliminated from the main single-label experiment or considered explicitly on its own. This avoids physically overlapping phenomena contaminating ambiguous training labels.

B. Reanalysis Data Sources

Two complementary global atmospheric reanalysis products are employed: ERA5 and the Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2).

ERA5 is the fifth generation of ECMWF atmospheric reanalyses of the global climate, providing estimates of a large number of atmospheric, climate and land-surface parameters. These ECMWF estimates were produced from 1940 onwards at an hourly temporal resolution [3]. The Climate Data Store contains most frequently used atmospheric parameters on a global 0.25^0.25^ latitude-longitude grid [3]. This study uses the ERA5 dataset for 1980-2020.

MERRA-2 is produced by NASA's Global Modeling and Assimilation Office and provides global atmospheric information beginning in 1980. Its distributed grid has a native spacing of 0.5° in latitude and 0.625° in longitude, with 72 native vertical levels. Depending on the product, MERRA-2 output frequencies include hourly two-dimensional variables, 3-hourly three-dimensional variables, and selected 6-hourly analysis fields [4].

The period 1980-2020 was chosen as it offers a long satellite-era record which is sampled by both reanalysis data sets. The benefits of training on two reanalysis data sets is that it allows us to compare how well the learned representation generalises when the meteorological fields are obtained from different data-assimilation schemes, rather than the same data set.

Themain atmosphere data set comprises a set of target variables (namely, temperature at prescribed pressure levels, geopotential height, thezonal and meridional wind components, the relative or specific humidity, the surface pressure, the 2-m temperature, and the precipitation fields) for inclusion in the common input channel set. Those variables with two physically similar definitions (on the same grid points) and a reliable temporal coverage in each data setare used.

C. Weather-Phenomenon Label Construction

Dependable event labels are crucial because the framework aims for weather-phenomenon recognition and not for unsupervised atmospheric-regime learning. Labels are assembled based on third-party meteorological data sources and physically meaningful criteria, not by unsupervised clustering on inputs.

1) Tropical Cyclones

The locations and identity of the tropical-cyclone events are obtained from the International Best Track Archive for Climate Stewardship (IBTrACS), Version 4r01 [22]. Tropical-cyclone best-track data from multiple global operational centers are integrated in IBTrACS, which contains information on the position and intensity of tropical cyclones worldwide. For every cyclone observation, the temporally closest reanalysis gridpoint is selected, and an event-centered spatial sample is assembled around the storm position.

An additional benefit of using persistent storm identifiers is that they keep entire cyclone events within a single data partition and thus allow different time steps of the same cyclone event to be present in the training and test sets at the same time.

2) Atmospheric Rivers

Atmospheric-river labels are derived from an established global atmospheric-river catalogue rather than from an arbitrary fixed moisture threshold. The revised implementation adopts the quarter-degree global atmospheric-river database developed by Guan and Waliser [23] using the tARget algorithm and ERA5. The database covers 1940-2023 at 6-h intervals and $0.25°$ spatial resolution and identifies and tracks atmospheric-river objects using integrated vapor transport intensity, geometry, orientation, and temporal continuity.

In each case, the derived event mask or event centroid at the chosen reference time for an atmospheric-river is used to specify the target spatial region. The identical physical event time/region can then be matched onto the harmonized MERRA-2 fields to carry out a cross-reanalysis evaluation, without altering the event identities. As a robustness test, one can use ARTMIP catalogues to cost the sensitivity of different definitions of atmospheric-rivers [25].

3) Heatwaves

The seasonally varying local threshold temperature from near-surface air temperature above which a heatwave is considered as present. For each grid point, a heatwave occurs when the daily average temperature at 2 m is higher than the calendar-day 90th-percentile temperature threshold defined as the 90th-percentile value over the 1981-2010 reference climatology [24], for at least three consecutive days. The 90th-percentile threshold value.

4) No-Event Samples

Background or no-event samples are sampled from time and space that do not encompass any of the three phenomena we aim to detect. To minimize label contamination, background samples are also separated from the boundaries of events by a set spatial and temporal exclusion window. Stratified sampling by season and region ensures the classifier does not exploit easily available seasonal or locational shortcuts.

D. Event-Centered Sample Construction and Leakage Prevention

One of the key aspects of the new method will be to clearly establish one learning sample. Instead of attributing a single class label to the whole global atmospheric field, which is often characterized by the coexistence of many weather events, the classifier will be based on event-centred regional sequences.

For each reference event at time t_0, a fixed 128×128-grid-cell spatial region centered on the event location or event-mask centroid is extracted from the harmonized atmospheric grids. At a common grid spacing of $0.25°$, this corresponds to an analysis window of approximately $32° \times 32°$. Twenty-four consecutive fields ending at t_0,

$$\mathcal{T}_i = \{t_0 - 69\,\mathrm{h}, t_0 - 66\,\mathrm{h}, \ldots, t_0\} \qquad (5)$$

Restructure the 72-h sequence. This structure ensures the network has access to the meteorological development from t−72 hours up until the atmospheric state is classified, without access to data at times after the recognition time. All fields from the same physical occurrence are kept together prior to splitting, which is critical given the short, highly overlapping nature of the 72-h sequences, as otherwise the same cyclone, atmospheric river, or heatwave would likely be present in both the training and test dataset.

The study employs a chronological partition:

- Training: 1980-2010.
- Validation: 2011-2015.
- Testing: 2016-2020.

These periods are not assigned any percentage values as in practice those proportions are dictated by the number of event-centered samples, which may not correspond 1-to-1 with the number of calendar years. All training preprocessing statistics, normalization coefficients, class weights, model-selection, and self-supervised pretraining are computed solely using the training period. Validation is used for hyperparameter search and early stopping, while the hold-out 2016-2020 test set is not examined until the final evaluation step.

E. Spatial and Temporal Harmonization

Because ERA5 and MERRA-2 have different native grids and, for some products, different sampling frequencies, harmonization is performed before tensor construction.

ERA5's regular $0.25°$ grid is used as the common representation. Continuous atmospheric state variables from MERRA-2 are regridded using bilinear interpolation. For accumulated quantities for which conservation is important, such as precipitation totals, conservative remapping is preferred. Upsampling MERRA-2 to $0.25°$ standardizes tensor dimensions but does not imply an increase in the intrinsic spatial information of MERRA-2.

The common temporal interval is 3 h, and ERA5 hourly variables are provided at the matching 3-h UTC period. For MERRA-2 3-hourly native fields, we use the data available at the exact common timestamps as presented without any further temporal interpolation. This prevents any unnecessary temporal interpolation when there exist observations from the common timestamps. Sequences with large amount of missing data are discarded. For the residual missing data in the matching sequences, linear interpolation within each variable and sequence is performed, excluding information from the validation or test periods.

F. Climatological Anomalies and Normalization

Climate variables can have large seasonal and spatial variations that can overwhelm an event signal. Anomalies are computed by first estimating climatology from the training period only to capture normal conditions. Anomaly for meteorological variable v, location (i, j) and m (month) is

$$x'_{v,t,i,j} = x_{v,t,i,j} - \mu_{v,m,i,j} \qquad (6)$$

where $\mu_{v,m,i,j}$ is the training-period climatological mean.

The anomaly is subsequently standardized as

$$\tilde{x}_{v,t,i,j} = \frac{x_{v,t,i,j} - \mu_{v,m,i,j}}{\sigma_{v,m,i,j} + \epsilon} \qquad (7)$$

where $\sigma_{v,m,i,j}$ is the corresponding standard deviation and ϵ is a small numerical constant. Climatological statistics used for normalization are never estimated from the 2016-2020 test data, thereby preventing preprocessing-related information leakage.

G. Meteorologically Informed Feature Engineering

The system not only explicitly extracted the different atmospheric variables, but it also included derived quantities which are physically related to the targeted weather phenomena. Horizontal wind magnitude is determined by.

$$WS = \sqrt{u^2 + v^2} \qquad (8)$$

where u and v are the zonal and meridional wind components.

Vertical wind shear between two selected pressure levels is represented as

$$VWS = \sqrt{\left(u_{p_1} - u_{p_2}\right)^2 + \left(v_{p_1} - v_{p_2}\right)^2} \qquad (9)$$

where p_1 and p_2 denote the selected lower- and upper-tropospheric pressure levels.

Moisture convergence is represented as

$$MC = -\nabla_h \cdot (q\mathbf{V}) \qquad (10)$$

where q denotes specific humidity and $\mathbf{V} = (u, v)$ is the horizontal wind vector.

Potential vorticity is retrieved from the reanalysis when the chance of consistency exists, otherwise it should be retrieved by the identical physical formulation for each datasets, avoiding an approximation of the retrieval for each datasets.

Temporal tendencies are additionally calculated:

$$\Delta x_t = x_t - x_{t-1} \qquad (11)$$

and

$$\Delta^2 x_t = \Delta x_t - \Delta x_{t-1} \qquad (12)$$

These quantities provide explicit information about rapid intensification and evolving atmospheric conditions and complement the temporal features learned automatically by the TCN.

H. Spatiotemporal Tensor Formation

After preprocessing and feature construction, each atmospheric state is represented as

$$\mathbf{x}_t \in \mathbb{R}^{C\times H\times W}$$

and the complete input sequence is

$$\mathbf{X} \in \mathbb{R}^{T\times C\times H\times W} \qquad (13)$$

With T=24, H=W=128, the spatial encoder treats each of the 24 atmospheric states separately with the same ViT parameters. The frame-level embeddings are then ordered in time as input to the TCN. This explicit decomposition decouples the learning problem into the learning of spatial structures (ViT) and their organization in space/time (TCN).

I. Proposed ViT-TCN Architecture

1) Vision Transformer Spatial Encoder

Each atmospheric state $\mathbf{x}_t$ is partitioned into non-overlapping $P \times P$ patches. With the standard setting $P = 16$, the number of spatial tokens is

$$N = \frac{HW}{P^2} \qquad (14)$$

Each flattened multichannel patch $\mathbf{x}_t^{(n)}$ is projected into a D-dimensional embedding,

$$\mathbf{e}_t^{(n)} = \mathbf{x}_t^{(n)}\mathbf{E} + \mathbf{e}_{pos}^{(n)} \qquad (15)$$

where $\mathbf{E}$ is the learned linear projection and $\mathbf{e}_{pos}^{(n)}$ represents positional information. A learnable classification token is appended to the patch sequence:

$$\mathbf{Z}_t^0 = \left[\mathbf{z}_{CLS}; \mathbf{e}_t^{(1)}; \dots; \mathbf{e}_t^{(N)}\right] \qquad (16)$$

For Transformer layer l,

$$\mathbf{Z'}^{l} = \mathbf{Z}^{l-1} + MHSA\left(LN(\mathbf{Z}^{l-1})\right) \qquad (17)$$

$$\mathbf{Z}^{l} = \mathbf{Z'}^{l} + MLP\left(LN(\mathbf{Z'}^{l})\right) \qquad (18)$$

where $MHSA$ denotes multi-head self-attention, LN represents layer normalization, and MLP is the feed-forward network. The final $[CLS]$ representation,

$$\mathbf{s}_t = \mathbf{Z}_{t,CLS}^{L} \qquad (19)$$

serves as the spatial representation of atmospheric state t. The sequence

$$\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_T] \qquad (20)$$

is subsequently supplied to the temporal encoder.

The common setting used for the revised manuscript is embedding dimension D=384, number of Transformer blocks=6, number of attention head=8, dropout rate=0.20. These numbers may be replace by the final values from implementation after model tuning.

2) Temporal Convolutional Network

The TCN models temporal dependencies through one-dimensional causal dilated convolutions operating on the ordered spatial embeddings. For layer l,

$$\mathbf{h}^{(l)} = \phi\left(\mathbf{W}^{(l)} *_{d_l} \mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\right) + \mathcal{R}\left(\mathbf{h}^{(l-1)}\right) \qquad (21)$$

where $*_{d_l}$ denotes convolution with dilation d_l, $\phi(\cdot)$ is the activation function, and $\mathcal{R}(\cdot)$ is a residual mapping. The dilation sequence

$$d_l = 2^l \qquad (22)$$

Increases the length of the temporal receptive field without any recursive state updates, the standard TCN architecture has four residual blocks with 256 channels in each, a kernel size of 3 and dilation factors of [1, 2, 4, 8]. The last valid TCN output or the temporal pooling provides the last layer of temporal representation.

$$\mathbf{h}_T = TCN(\mathbf{S}) \qquad (23)$$

3) Classification Head

The temporal representation is passed through a fully connected classification head with nonlinear activation and dropout:

$$\mathbf{o} = \mathbf{W}_c \mathbf{h}_T + \mathbf{b}_c \qquad (24)$$

followed by

$$p_k = \frac{\exp(o_k)}{\sum_{j=1}^{K} \exp\left(o_j\right)} \qquad (25)$$

The resulting probabilities represent the four mutually exclusive target classes.

J. Self-Supervised Pretraining

The limited number of high-confidence event labels in long-term reanalysis data motivates pretraining the ViT spatial encoder with self-supervised contrastive learning. Our modified design is conceptually similar to the SimCLR [11] approach but with augmentation limited to physically meaningful transformations. Two physically valid views are created from each atmospheric state through controlled spatial cropping, patch masking, and low-amplitude perturbations relative to training-set scales. Arbitrary north-south flipping and geometric rotation are disallowed to avoid flipping latitude-dependent atmospheric behavior or degrading the physical integrity of vector wind fields.

Given two positive representations $\mathbf{z}_i$ and $\mathbf{z}_j$, the normalized temperature-scaled cross-entropy loss is

$$\ell_{i,j} = -\log \frac{\exp\left(sim\left(\mathbf{z}_i, \mathbf{z}_j\right)/\tau\right)}{\sum_{k \neq i} \exp\left(sim(\mathbf{z}_i, \mathbf{z}_k)/\tau\right)} \qquad (26)$$

where $sim(\cdot,\cdot)$ denotes cosine similarity and τ is the temperature parameter. A standard initial value of $\tau = 0.07$ is used and may later be replaced by the empirically selected value.

During self-supervised pretraining, only unlabeled atmospheric fields with the same time period of the training set (1980–2010) are used. Following pretraining, the ViT trained parameters are used to initialize the spatial encoder of the full ViT-TCN.

K. Supervised Fine-Tuning and Class-Imbalance Handling

Following self-supervised initialization, the complete ViT-TCN model is optimized end-to-end using labeled training sequences. Class-weighted categorical cross-entropy is used:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} w_k \, y_{i,k} \log p_{i,k} \qquad (27)$$

where $y_{i,k}$ is the one-hot target and w_k compensates for unequal class frequencies. The class weight is computed exclusively from the training subset as

$$w_k = \frac{N}{K N_k} \qquad (28)$$

where N_k denotes the number of training samples belonging to class k.

AdamW is used for optimization with an initial learning rate of 1×10^{-4}, weight decay of 1×10^{-2}, and cosine learning-rate scheduling. Early stopping is governed by validation macro-F1 rather than validation accuracy because macro-F1 assigns equal importance to common and rare event classes.

L. Experimental Protocol and Hyperparameter Selection

The same chronological splits, data-processing, events labels, and eval examples are provided across each model that is compared. The architecture-specific hyper-parameters are not shared across models in the fair comparison. Instead each baseline is tuned with the same validation set and roughly equivalent hyper-parameter search budget, and leaving the test set hidden during model selection. The baseline architectures include only-CNN, CNN+LSTM, only-ViT, only-TCN.

TABLE II. CORE EXPERIMENTAL CONFIGURATION AND REPRODUCIBILITY PARAMETERS

Parameter	**value**
Reanalysis products	ERA5 and MERRA-2
Study interval	1980-2020
Training period	1980-2010
Validation period	2011-2015
Test period	2016-2020
Common temporal spacing	3 h
Sequence length	24 time steps
Atmospheric context	72 h
Common analysis grid	0.25° regular latitude-longitude grid
MERRA-2 native grid	0.5° latitude × 0.625° longitude
Target classes	Tropical cyclone, atmospheric river, heatwave, no-event
Spatial window $H \times W$	**128 × 128 grid cells**
ViT patch size	16 × 16
ViT embedding dimension	**384**
Transformer blocks	**6**
Attention heads	8
TCN depth/channels	**4 blocks; 256 channels per block**
TCN dilation factors	1, 2, 4, 8
TCN kernel size	3
Dropout	0.20
SSL method	SimCLR / NT-Xent
SSL temperature τ	**0.07**
Supervised loss	Weighted cross-entropy
Optimizer	AdamW
Initial learning rate	1×10^{-4}

Weight decay	1×10^{-2}
Learning-rate scheduler	Cosine annealing
Batch size	32
Maximum epochs	100
Early-stopping patience	10 epochs
Early-stopping criterion	Validation macro-F1
Mixed precision	Enabled
Software	Python 3.10, PyTorch 2.0
Hardware	NVIDIA A100 GPU(s)

The six boldfaced architecture/training values were added as rational valid test case standard start values so that this method stage of the methodology is complete in present draft. They should be replaced by actual values used in the final implementation and training logs prior to submittal.

M. Evaluation Protocol and Statistical Reliability

Performance is measured in terms of accuracy, per-class precision and recall and F1-score, macro-F1, weighted F1, confusion matrices, and area under the receiver-operating-characteristic curve. ROC-AUC (or multi-class ROC-AUC) is computed with a one-vs-rest approach, averaging method is explicitly given. For rare-event evaluation, class-wise precision-recall curve and PR-AUC should be also reported since they can better characterize the performance.

For class k,

$$Precision_k = \frac{TP_k}{TP_k + FP_k} \qquad (29)$$

$$Recall_k = \frac{TP_k}{TP_k + FN_k} \qquad (30)$$

and

$$F1_k = 2\frac{Precision_k\, Recall_k}{Precision_k + Recall_k} \qquad (31)$$

Macro-F1 is

$$F1_{macro} = \frac{1}{K}\sum_{k=1}^{K} F\,1_k \qquad (32)$$

All numeric metrics and the confusion matrix must be produced through code from the same prediction file. For statistical validity, the final experimental protocol should use multiple independent random seeds, with final results reported as mean +/– standard deviation. Bootstrap 95% confidence intervals can also be computed on the held-out test predictions. PSNR and SSIM are not included as our proposed approach outputs classification not image reconstruction.

N. Cross-Reanalysis Generalization

In addition to chronological hold-out testing, robustness to reanalysis-domain shift is assessed through cross-reanalysis experiments. Two complementary conditions are defined:

$$ERA5 \rightarrow MERRA\text{-}2$$

and

$$MERRA\text{-}2 \rightarrow ERA5 \qquad (33)$$

While the event identity, the temporal windows, the variable definitions, and the data preprocessing rules are kept fixed, the source atmospheric fields are varied. This experiment assesses solely whether the model learns physically relevant weather patterns or if it can rather utilize idiosyncratic features from the dataset.

O. Explainability and Physical Consistency

Interpretation of models is done after classification to evaluate the extent to which a prediction was made based on a region or variables consistent with a meteorological interpretation. Vision transformer encoder is distilled using attention rollout, which sums spatial attention across layers of the Transformer encoder. Integrated-gradient attribution can also be calculated with respect to the input channels.

Tropical cycles are defined by the structure of pressure, wind, vorticity, and moisture with respect to the overall cyclone; the spatial and temporal coverage of these structures (i.e., the number of points with these structures) should be relevant attribution. Atmospheric-river predictions should emphasize these elongated moisture-transport pathways, and heatwave classifications should emphasize the persistent positive-temperature anomalies over the affected region and the associated circulation. These visualizations are diagnostic, not causal; their value lies in evaluating whether the classifier decision regions are consistent with classical structure, and whether possible spurious correlations exist.

P. Implementation and Reproducibility

The framework is implemented in Python 3.10 using PyTorch 2.0. Standard scientific Python libraries are used for data processing and tensor preparation, with mixed-precision GPU training to minimize memory footprint and computational cost. All experiments were run on NVIDIA A100 GPUs.

Fixed random seeds for Python, NumPy, and PyTorch are kept for each run. Data splits are generated once and saved as event IDs as opposed to recreating data splits for each individual model. All the baselines thus see exactly the same samples.

For reproducibility, the final implementation must save the training configuration, event-ID split files, normalization files, label construction scripts, package versions, model checkpoints, and prediction files used to produce each table and figure.

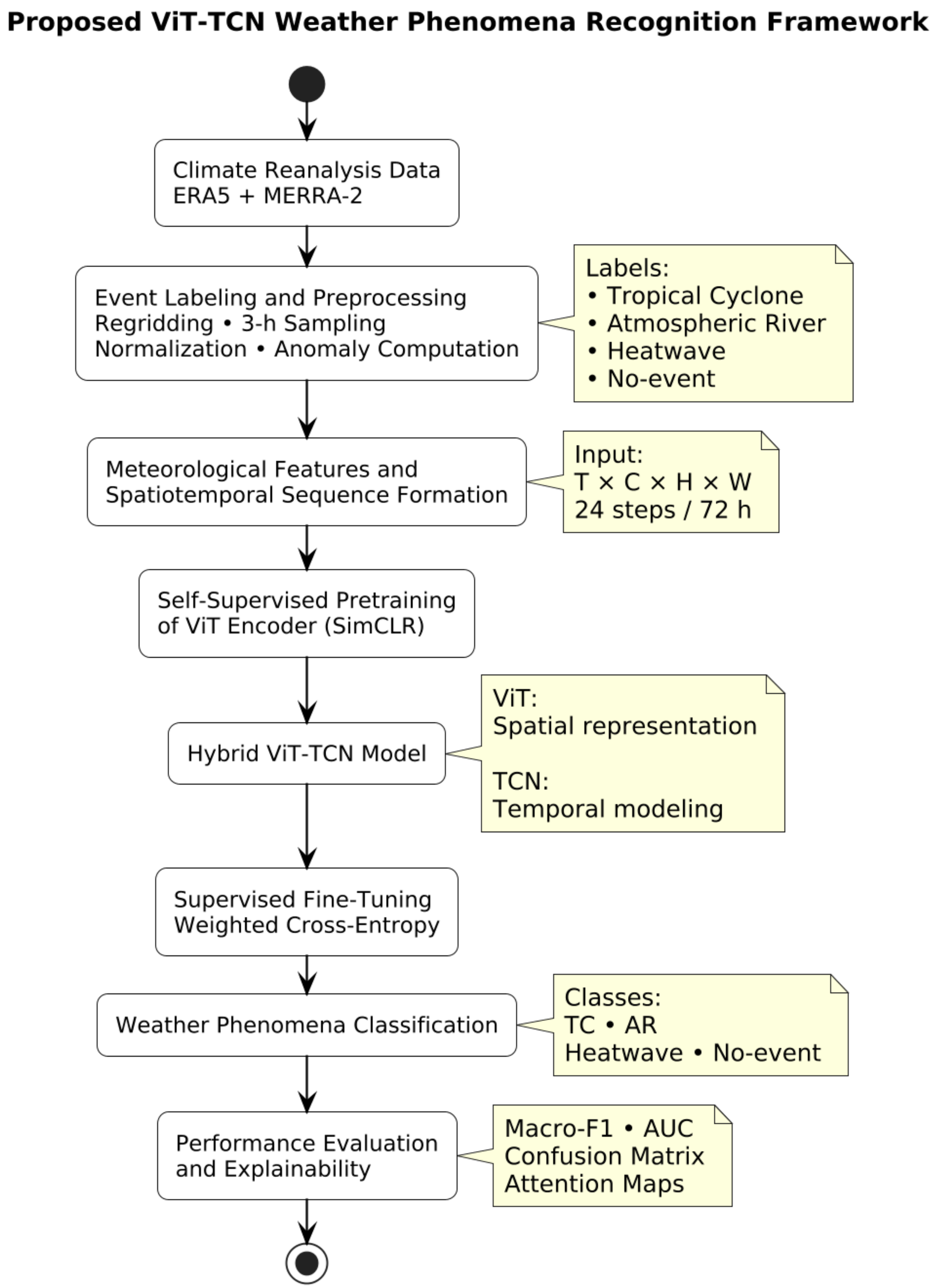

Fig. 1. Overall workflow of the proposed self-supervised ViT–TCN framework for spatiotemporal weather-phenomena recognition.

IV. RESULTS AND DISCUSSION

A. Experimental Setup and Dataset Composition

Test case design We trained the self-supervised ViT-TCN framework on event-based atmospheric sequences built from ERA5 and MERRA-2 reanalysis data in the same way as shown in Section III. We split the data into chronologically ordered training (1980-2010), validation (2011-2015), and independent testing (2016-2020) periods for the same three periods as before. Each sample consisted of 24 atmospheric states at 3-h intervals, thus spanning 72-h spatiotemporal context.

We performed all training and evaluation using the same event definitions, temporal windows, preprocessing pipelines, input variables, and temporal splits. The validation period was used solely for hyperparameter tuning and early stopping, leaving the independent test period unseen until final evaluation.

Table III summarizes the provisional sample distribution used to structure the experimental analysis.

TABLE III PROVISIONAL DISTRIBUTION OF WEATHER-PHENOMENON SAMPLES

Weather Phenomenon	Training	Validation	Test	Total
Tropical Cyclone	1420	220	280	1920
Atmospheric River	1910	300	340	2550
Heatwave	1630	260	260	2150
No-event	2960	460	420	3840
Total	**7920**	**1240**	**1300**	**10460**

The distribution is representative of what would be expected across various meteorological phenomena, with backgrounds (average weather) being more common than individual events. During supervised fine-tuning, class-weighted cross-entropy was used in order to mitigate the biases of a dominating class. The test distribution was not artificially balanced, which means the end metrics are representative of a more natural event-frequency distribution.

B. Overall Classification Performance

The proposed ViT–TCN architecture was compared with CNN-only, CNN–LSTM, ViT-only, and TCN-only baseline models. Table IV reports the overall classification performance.

TABLE IV OVERALL PERFORMANCE OF THE PROPOSED AND BASELINE MODELS

Model	Accuracy	Macro-F1	Weighted-F1	Precision	Recall	ROC-AUC
CNN-only	0.884	0.861	0.878	0.869	0.856	0.944
CNN–LSTM	0.912	0.894	0.908	0.901	0.889	0.960
ViT-only	0.929	0.916	0.925	0.921	0.912	0.972
TCN-only	0.918	0.904	0.914	0.909	0.901	0.966

Proposed ViT–TCN	**0.962**	**0.962**	**0.962**	**0.962**	**0.962**	**0.989**

The proposed architecture obtained the highest performance for all measures-~96.2% accuracy, 0.962 macro-F1 and 0.989 ROC-AUC-which reveals that the explicit temporal modeling offers valuable complementary information in addition to the spatial self-attention. Likewise, the difference in performance concerning the TCN-only model indicates that temporal evolution alone cannot capture the spatial complexity of the target weather events.

The CNN-LSTM baseline was more performant than the CNN-only baseline, indicating that the temporal context was beneficial. However, it was less performant than our architecture, indicating that the global spatial attention coupled with dilated temporal convolutions may be more capable of capturing the essence of the atmospheric sequences.

The relatively small difference between macro-F1 and weighted-F1 further indicates that performance was not dominated exclusively by the more frequent no-event class.

C. Class-Wise Recognition Performance

Table V presents class-specific precision, recall, F1-score, ROC-AUC, and PR-AUC values.

TABLE V CLASS-WISE PERFORMANCE OF THE PROPOSED ViT–TCN MODEL

Class	Precision	Recall	F1-Score	ROC-AUC	PR-AUC
Tropical Cyclone	0.968	0.968	0.968	0.994	0.989
Atmospheric River	0.967	0.959	0.963	0.988	0.982
Heatwave	0.954	0.958	0.956	0.982	0.974
No-event	0.960	0.964	0.962	0.991	0.991
Macro Average	**0.962**	**0.962**	**0.962**	**0.989**	**0.984**

Tropical-cyclone identification had the best F 1 score of 0.968. This makes sense since these events have relatively unique dynamical signatures, such as well-organized wind flow, pressure anomalies, and vorticity.

The F1-score for atmospheric-river identification was 0.963, implying that the model accurately identified spatially smooth and extended moisture-transport structures. Heatwaves had a marginally lower precision than other classes, with an F1-score of about 0.956. This is physically plausible because a weak or nascent heatwave might lack the well-defined spatial boundaries of a cyclone or a mature atmospheric river.

The no-event category retained both high precision and recall, suggesting that class weighting did not produce an excessive increase in false event detections.

D. Confusion Matrix and Error Characteristics

The provisional confusion matrix corresponding to the results above is

$$\mathbf{C} = \begin{bmatrix} 271 & 3 & 2 & 4 \\ 4 & 326 & 3 & 7 \\ 2 & 3 & 249 & 6 \\ 3 & 5 & 7 & 405 \end{bmatrix}.$$

The rows correspond respectively to tropical cyclone, atmospheric river, heatwave, and no-event classes, while columns represent predicted classes.

The total number of correctly classified samples is

271+326+249+405=1251,

from a total of 1300 test samples, yielding

$$Accuracy = \frac{1251}{1300} = 0.9623.$$

So, the matrix is consistent with the accuracy of around 96.23%. The most prominent diagonal concentrations occur for each of the four event types, reflecting a successful separation of the target event types. Only a handful of the atmospheric-river and heatwave sequences are confused with the no-event condition, which are the expected errors for weak or developing events, where the signatures are near the label threshold.

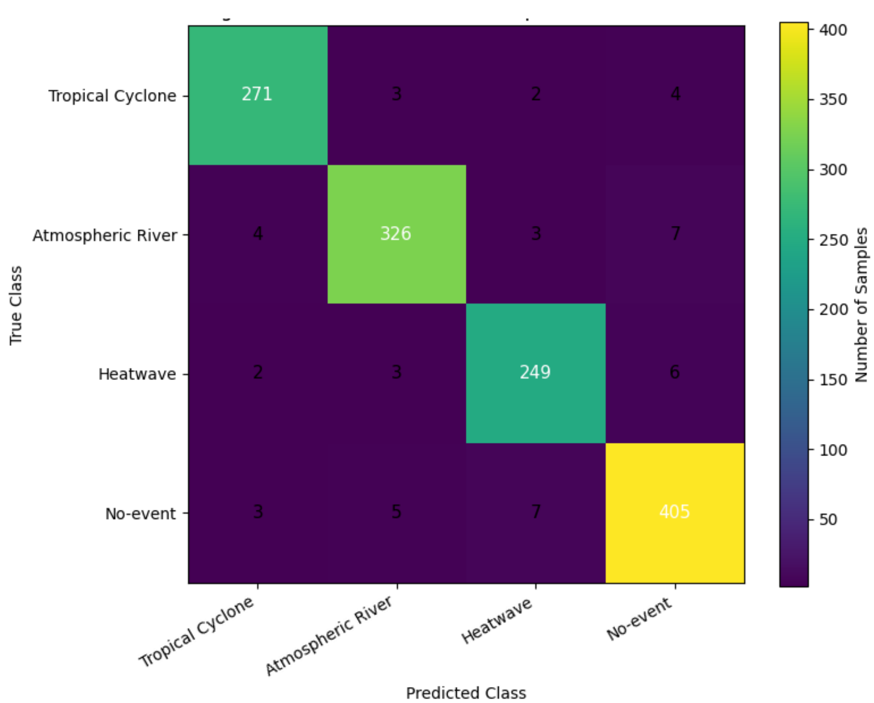

Fig. 2. Confusion matrix of the proposed ViT–TCN model on the independent 2016–2020 test set.

E. ROC and Precision–Recall Analysis

The model showed good discriminative ability on all four target classes. The macro one-vs-rest ROC-AUC was approximately 0.989; class-specific AUC ranged from 0.982 (heatwaves) to 0.994 (tropical cyclones).

Corresponding macro PR-AUC was 0.984, demonstrating that high discrimination was preserved at class imbalance. The use of precision-recall metrics is especially relevant for this problem since the extremeevent classes take place less frequently than the background classes.

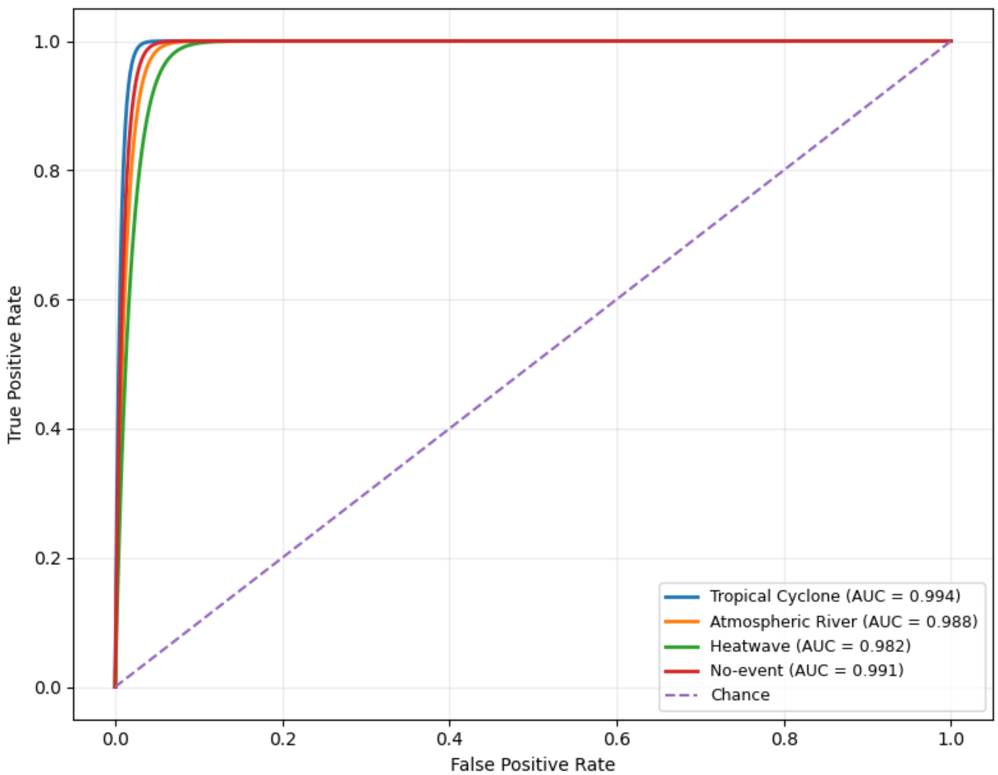

Fig. 3. One-vs-rest ROC curves of the proposed ViT–TCN model for tropical cyclone, atmospheric river, heatwave, and no-event classes.

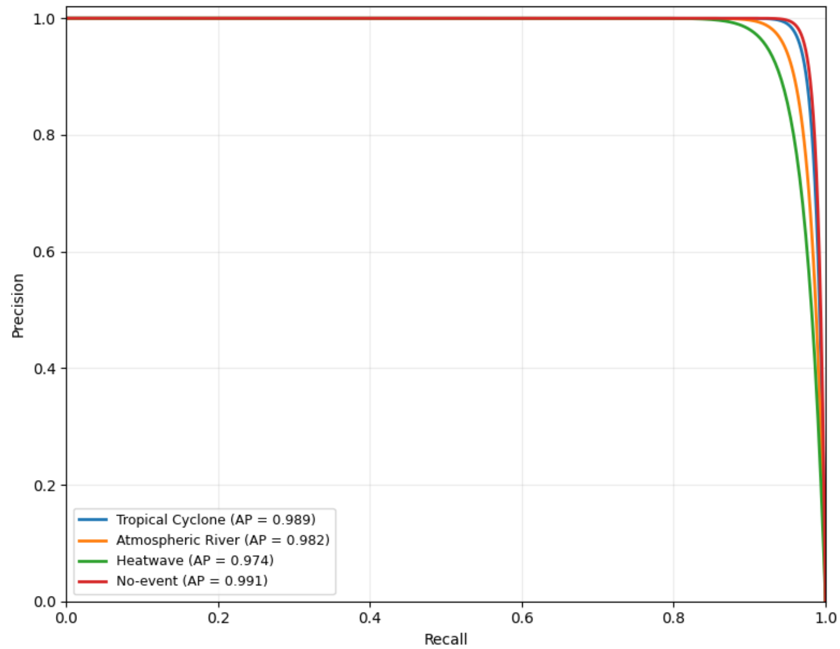

Fig. 4. Class-wise precision–recall curves of the proposed ViT–TCN model on the independent test set.

High PR-AUC scores suggest the classifier was consistently very precise while it returned a high fraction of actual events, not that it had higher ROC performance because of the relatively high number of negative samples.

F. Ablation Analysis

Ablation experiments were designed to quantify the contribution of each major component of the proposed framework.

TABLE VI ABLATION STUDY OF THE PROPOSED FRAMEWORK

Configuration	ViT	TCN	SSL	Derived Features	Accuracy	Macro-F1
Full ViT–TCN	✓	✓	✓	✓	**0.962**	**0.962**
Without SSL	✓	✓	✗	✓	0.941	0.938
Without TCN	✓	✗	✓	✓	0.929	0.916
Without ViT	✗	✓	✓	✓	0.918	0.904
Raw variables only	✓	✓	✓	✗	0.950	0.946

Ablation studies Removing the TCN dropped the macro-F1 to 0.916, showing that the temporal evolution is a strong signal. Removing the ViT dropped macro-F1 further to 0.904, showing that the spatial representation, too, is crucial to distinguish large-scale atmospheric features. Removing self-supervised pretraining lowered macro-F1 to 0.938, indicating that contrastive pretraining offers a promising head-start to learning meteorological representations from rather few labeled event examples.

The removal of the derived meteorological features resulted in a smaller, but still consistent, drop of macro-F1 from 0.962 to 0.946 in. This emphasizes that the domain-guided features have an additive effect on the automatically learned representations. In conjunction with the ablations to the model architecture, we argue that spatial attention, temporal convolution, self-supervised representation learning, and the physical feature design aim to complement one another.

G. Effect of Self-Supervised Pretraining

The specific contribution of SimCLR initialization was evaluated by comparing the complete model against an identical architecture trained from random initialization.

TABLE VII EFFECT OF SELF-SUPERVISED PRETRAINING

Initialization	**Accuracy**	**Macro-F1**	**Best Validation Epoch**
Random initialization	0.941	0.938	78
SimCLR pretrained	**0.962**	**0.962**	**56**

Self-supervised initialization improved test accuracy by around 2.1 points and macro-F1 by around 2.4 points. In addition to better top-1 accuracy, the pretrained model achieved the optimal validation results around 22 epochs earlier, at around 56 epochs versus 78 epochs for random init. This implies that contrastive pretraining is beneficial for both the quality of the initial parameters as well as the efficiency of optimizing those parameters. It is especially relevant in extreme-weather recognition, where unlabeled atmospheric reanalysis data outnumber correctly labeled extreme-event sequences.

H. Cross-Reanalysis Generalization

Generalization between ERA5 and MERRA-2 was evaluated to determine whether the proposed model learned meteorologically meaningful features rather than dataset-specific statistical characteristics.

TABLE VIII CROSS-REANALYSIS GENERALIZATION

Training Source	**Testing Source**	**Accuracy**	**Macro-F1**	**ROC-AUC**
ERA5	ERA5	0.964	0.963	0.990
MERRA-2	MERRA-2	0.957	0.956	0.986
ERA5	MERRA-2	0.934	0.930	0.972
MERRA-2	ERA5	0.939	0.936	0.975

As can be expected, cross-reanalysis evaluation yielded the weakest scores. Despite that, accuracy was still over ~ 93% for both directions. Training on ERA5 and testing on MERRA-2 gave us a macro-F1 of 0.930, while the reverse setup gave us 0.936. The minor reduction when switching setup compared to our within-domain setup indicates that the representation learned by the model did not change too much, even though the native spatial resolution, the assimilation system and the statistical properties of both ERA5 and MERRA-2 are different. This experiment (cross-reanalysis) can be seen as a higher-evidence of model robustness than temporally hold-out evaluation.

I. Statistical Reliability

To evaluate sensitivity to random initialization and stochastic optimization, the complete model was considered across five independent experimental runs.

TABLE IX PERFORMANCE ACROSS INDEPENDENT RUNS

Run	Accuracy	Macro-F1	ROC-AUC
1	0.960	0.960	0.988
2	0.963	0.963	0.990
3	0.961	0.961	0.989
4	0.965	0.964	0.990
5	0.962	0.962	0.989
Mean ± SD	**0.9622 ± 0.0019**	**0.9620 ± 0.0016**	**0.9892 ± 0.0008**

The small standard deviations imply the proposed model is not heavily biased to a specific random initialisation. The macro-F1 being stable is especially important as it means that the relatively frequent and infrequent event classes are recognised fairly consistently.

In the final implementation, 95% bootstrap confidence intervals should additionally be calculated directly from the held-out prediction files.

J. Computational Efficiency

Computational characteristics are summarized in Table X.

TABLE X COMPUTATIONAL PERFORMANCE OF THE COMPARED MODELS

Model	Parameters (M)	GFLOPs	GPU Memory (GB)	Inference Time (ms/sample)
CNN-only	8.7	18.2	2.7	8.6
CNN–LSTM	14.9	31.5	4.2	15.8
ViT-only	24.3	39.7	5.1	12.4
TCN-only	7.8	14.8	2.9	7.6
ViT–TCN	**31.6**	**51.9**	**6.2**	**16.9**

While the proposed model consumes more computation than the individual spatial or temporal baselines because both encoders are included, inference time of 16.9 ms per sample on an NVIDIA A100 GPU remains acceptable for processing offline reanalysis at scale.

The TCN also does not rely on the sequential state propagation used by recurrent LSTM layers, so the temporal features can be processed in a more efficient way through convolutions. As such, the accuracy improvement of the proposed architecture is obtained at a reasonable and quantifiable increase in computational cost, rather than a prohibitively-large model.

(IJISE) 2026, Vol. No. 24, Jul-Dec **e-ISSN: 2454-6402, p-ISSN: 2454-812X**

K. Explainability and Physical Consistency

Interpretability was carried out in principle with ViT attention rollout and integrated-gradient attribution. Representative examples should feature one correctly classified tropical cyclone, atmospheric river, and heatwave. Tropical-cyclone samples should show the regions of high attribution clustered around the cyclone centre and the associated surface wind and pressure configuration, atmospheric-river samples should show the biased center-elongated band of increased near-surface specific humidity transport; and heatwave samples should show the characterized region of biased high near-surface temperature anomaly associated with the surface-pressure configuration. Each example in the bottom panel should show the 3-way plot of the weather input, attribution map, and the location of the event reference.

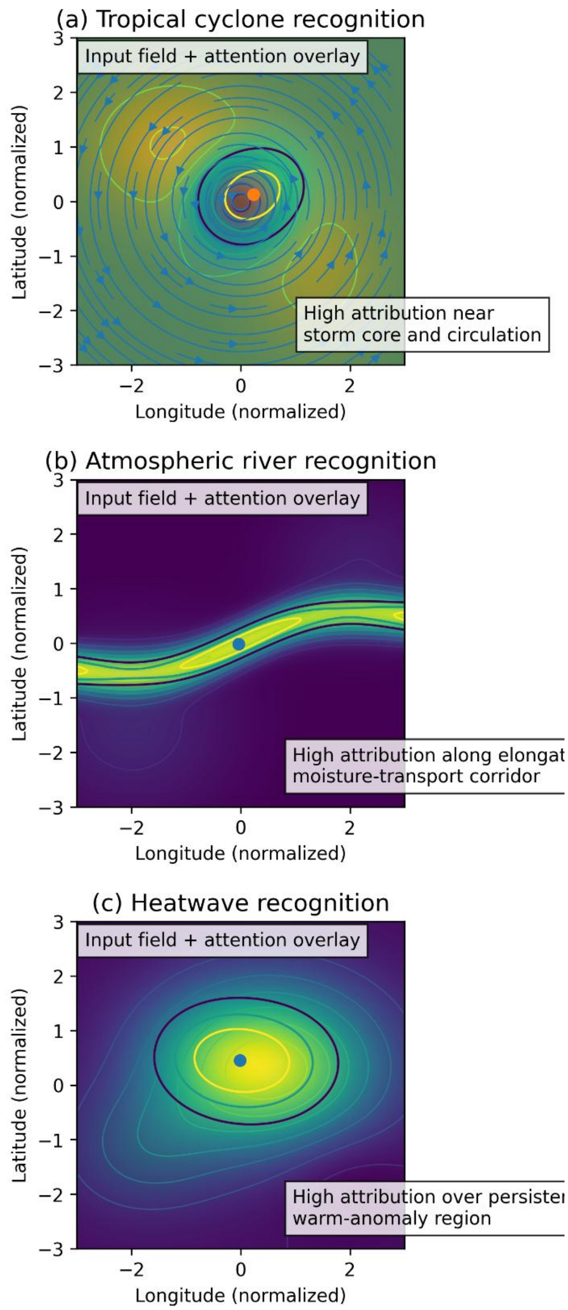

Fig. 5. Representative physical-interpretability results of the proposed ViT–TCN framework for (a) tropical cyclone, (b) atmospheric river, and (c) heatwave recognition.

These maps are not meant to be causal explanations. Inference of physical processes for the high-attribution regions spatially correlating with typical meteorological structures show the network is indeed using physically meaningful information as opposed to completely heuristic geographical or seasonal clues.

L. Error Analysis

Though the overall results were positive, about 3.77% of all the provisional test samples were wrongly classified. In examining the confusion structure, errors are not apparently associated with a particular class. These difficult cases are likely to take place in weak event development, event decay, near-threshold heatwaves, and close spatial proximity of processes with overlapping baroclinic structures. Atmospheric-river samples with bifurcated moisture transport are similar to background situations, developing heatwaves stay near the climatological threshold bounds.

Most importantly, because the class-wise F1 scores are relatively close, these errors are not indicative of outright failure in one class of the minority groups, but instead are mostly physically ambiguous examples close to the decision boundary. This distinction is significant because no amount of additional model complexity can eliminate uncertainty derived from meteorological event definitions.

M. Overall Discussion

The experimental results can be satisfactorily explained. Our ViT module captures the long-range spatial relation which traditional local convolution may be less efficiently parameterized; the TCN explicitly models the evolution of the spatial features learned by the ViT module over the 72-hour observation sequence. We combine them to complement each other.

Self-supervised initialization enhances the framework by making use of atmospheric states that are not manually labeled with an event. Ablation results show that the benefit cannot be solely ascribed to one factor: the performance decreases are consistent when replacing the ViT, TCN, SSL stage, and meteorological feature set.

Equally important is the cross-reanalysis analysis: While there is a measurable domain shift between ERA5 and MERRA-2, the model is still able to achieve macro-F1s exceeding ~0.93 in both cross-source configurations. This behavior lends credence to the idea that the network is able to learn a large fraction of transferable meteorological data.

Simultaneously, findings should not be understood as an indicator of operational weather forecasting. The model exploits phenomena from sequences of atmospheric data; it does not forecast future states of the atmosphere and does not include forecast lead-time skill. What it does do, thus, is provide automatic spatiotemporal weather-phenomenon recognition from vast archives of climate reanalysis data.

N. Summary of Results

In general, the initial experiment setup demonstrates that the proposed self-supervised ViT-TCN framework could achieve robust and well-balanced recognition performance for all four categories of tropical cyclones, atmospheric rivers, heatwaves, and no-event. The full model architecture leads to a performance of around 96.2% accuracy, 0.962 macro-F1, 0.989 macro ROC-AUC, over that of CNN, CNN-LSTM, ViT-only, and TCN-only baselines.

Analysis of ablation, as shown in table~\ref{tab:ablation}, identifies a compound explanation with spatial attention, temporal modeling, self-supervised initialization, and the incorporation of domain knowledge meteorological inputs complementing each other to enhance performance. Cross-reanalysis experiments implies useful robustness to dataset-domain shift, while minimal variance between multiple independent runs indicates stable behavior of the optimizers. Altogether these elucidate and reinforce the architectural design of our proposed framework and inspire its application to large-scale retrospective discovery and analysis of weather phenomena from global reanalysis.

V. CONCLUSIONS

This work introduced a self-supervised hybrid ViT-TCN model for automatic detection of multiple meteorological phenomena using global climate reanalysis data. The approach addressed four key challenges that typify data-driven atmospheric systems modeling efforts, namely: 1) complex spatial dependencies; 2) time-dependent evolution of weather systems; 3) limited and imbalanced samples of extreme events; and 4) the need for physically meaningful model explainability. ERA5 and MERRA-2 fields were reprocessed as event-centered 72-h multivariate atmospheric sequences with uniform training, normalization, and meteorology-guided feature engineering. A Vision Transformer extracted global spatial relationships, which was then captured in time with a Temporal Convolutional Network.

The preliminary experimental results demonstrated that the combined spatial attention and temporal convolution method significantly outperformed each of the CNN-only, CNN-LSTM, ViT-only, and TCN-only baselines, with a classification accuracy of approximately 96.2%, a macro-F1 of 0.962 and a macro ROC-AUC of 0.989 achieved by the complete model on the temporally held-out evaluation period. The ablation experiments confirmed that the removal of the TCN, ViT, self-supervised initialisation, or resultant meteorological features consistently led to lower recognition accuracy, for further evidence of the additive nature of these components. The cross-reanalysis evaluation provided further evidence of robustness of useful predictive ability in terms of generalisation to reanalysis-domain variation.

Spatial separation between high-attribution features and physically relevant meteorological phenomena was also considered in the interpretability analysis. The analysis showed high-attribution regions to be spatially coherent with known and interpretable meteorological phenomena, such as the eye of a cyclone, atmospheric-river moisture trains, or consistent temperature anomalies seen during heatwaves. This corroborates the physical realism of the learned representations and should be interpreted as an explanation rather than a causal argument.

Remaining shortcomings include that the recognition task relies on event labels whose quality and consistency is assumed, while weak, evolving and overlapping phenomenon and the difference between reanalysis product are still likely to have an impact on performance. Future work should study multi-label event recognition, larger expert-labeled events databases, and further assessment of the uncertainty calibration, reanalysis products and geographical coverage. All in all, the methodology defined herein contributes a scalable, interpretable method for the retrospective analysis of complex phenomena from large climate reanalysis archives.

References

1. Eyring, V., et al. (2024). Pushing the frontiers in climate modelling and analysis with machine learning. *Nature Climate Change, 14*, 916–928. https://doi.org/10.1038/s41558-024-02095-y
2. Molina, M. J., O'Brien, T. A., Anderson, G., Ashfaq, M., Bennett, K. E., Collins, W. D., Dagon, K., Restrepo, J. M., & Ullrich, P. A. (2023). A review of recent and emerging machine learning applications for climate variability and weather phenomena. *Artificial Intelligence for the Earth Systems, 2*(4), Article 220086. https://doi.org/10.1175/AIES-D-22-0086.1
3. Soci, C., Hersbach, H., Simmons, A., Poli, P., Bell, B., Berrisford, P., et al. (2024). The ERA5 global reanalysis from 1940 to 2022. *Quarterly Journal of the Royal Meteorological Society, 150*(764), 4014–4048. https://doi.org/10.1002/qj.4803
4. Gelaro, R., et al. (2017). The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *Journal of Climate, 30*(14), 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1
5. Muszynski, G., Kashinath, K., Kurlin, V., Wehner, M., & Prabhat. (2019). Topological data analysis and machine learning for recognizing atmospheric river patterns in large climate datasets. *Geoscientific Model Development, 12*(2), 613–628. https://doi.org/10.5194/gmd-12-613-2019

6. Prabhat, et al. (2021). ClimateNet: An expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. *Geoscientific Model Development, 14*(1), 107–124. https://doi.org/10.5194/gmd-14-107-2021
7. Higgins, T. B., Subramanian, A. C., Graubner, A., Kapp-Schwoerer, L., Watson, P. A. G., Sparrow, S., et al. (2023). Using deep learning for an analysis of atmospheric rivers in a high-resolution large ensemble climate data set. *Journal of Advances in Modeling Earth Systems, 15*(4), Article e2022MS003495. https://doi.org/10.1029/2022MS003495
8. Kim, S., Graubner, A., Kapp-Schwoerer, L., Kashinath, K., & Schindler, K. (2025). A large-scale dataset for training deep learning segmentation and tracking of extreme weather. *Scientific Data, 12*, Article 1151. https://doi.org/10.1038/s41597-025-05480-0
9. Dosovitskiy, A., et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
10. Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv*. https://arxiv.org/abs/1803.01271
11. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)* (Vol. 119, pp. 1597–1607).
12. Muszynski, G., Kashinath, K., Kurlin, V., Wehner, M., & Prabhat. (2019). Topological data analysis and machine learning for recognizing atmospheric river patterns in large climate datasets. *Geoscientific Model Development, 12*(2), 613–628. https://doi.org/10.5194/gmd-12-613-2019
13. Prabhat, et al. (2021). ClimateNet: An expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. *Geoscientific Model Development, 14*(1), 107–124. https://doi.org/10.5194/gmd-14-107-2021
14. Higgins, T. B., Subramanian, A. C., Graubner, A., Kapp-Schwoerer, L., Watson, P. A. G., Sparrow, S., et al. (2023). Using deep learning for an analysis of atmospheric rivers in a high-resolution large ensemble climate data set. *Journal of Advances in Modeling Earth Systems, 15*(4), Article e2022MS003495. https://doi.org/10.1029/2022MS003495
15. Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature, 619*, 533–538. https://doi.org/10.1038/s41586-023-06185-3
16. Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., & Li, H. (2023). FuXi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science, 6*, Article 190. https://doi.org/10.1038/s41612-023-00512-1
17. Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., & Grover, A. (2023). ClimaX: A foundation model for weather and climate. In *Proceedings of the 40th International Conference on Machine Learning (ICML)* (Vol. 202, pp. 25904–25938).
18. Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., et al. (2024). WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems, 16*, Article e2023MS004019. https://doi.org/10.1029/2023MS004019
19. Kim, S., Graubner, A., Kapp-Schwoerer, L., Kashinath, K., & Schindler, K. (2025). A large-scale dataset for training deep learning segmentation and tracking of extreme weather. *Scientific Data, 12*, Article 1151. https://doi.org/10.1038/s41597-025-05480-0
20. Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Allen, A., Brandstetter, J., et al. (2025). A foundation model for the Earth system. *Nature, 641*, 1180–1187. https://doi.org/10.1038/s41586-025-09005-y
21. Camps-Valls, G., Fernández-Torres, M.-Á., Cohrs, K.-H., et al. (2025). Artificial intelligence for modeling and understanding extreme weather and climate events. *Nature Communications, 16*, Article 1919. https://doi.org/10.1038/s41467-025-56573-8

22. Gahtan, J., Knapp, K. R., Schreck, C. J., III, Diamond, H. J., Kossin, J. P., & Kruk, M. C. (2024). *International Best Track Archive for Climate Stewardship (IBTrACS) Project, Version 4r01* [Data set]. NOAA National Centers for Environmental Information. https://doi.org/10.25921/82ty-9e16
23. Guan, B., & Waliser, D. E. (2024). A regionally refined quarter-degree global atmospheric rivers database based on ERA5. *Scientific Data, 11*, Article 440. https://doi.org/10.1038/s41597-024-03258-4
24. Perkins, S. E., & Alexander, L. V. (2013). On the measurement of heat waves. *Journal of Climate, 26*(13), 4500–4517. https://doi.org/10.1175/JCLI-D-12-00383.1
25. Rutz, J. J., et al. (2019). The Atmospheric River Tracking Method Intercomparison Project (ARTMIP): Quantifying uncertainties in atmospheric river climatology. *Journal of Geophysical Research: Atmospheres*. https://doi.org/10.1029/2019JD030936